LEAN SIGMA
CORPORATION

Lean Six Sigma Black Belt Training

Featuring Examples from SigmaXL

1.0 Define Phase

1.1 Overview of Six Sigma

Black Belt Training: Define Phase

1.1 Six Sigma Overview

- 1.1.1 What is Six Sigma
- 1.1.2 Six Sigma History
- 1.1.3 Six Sigma Approach Y = f(x)
- 1.1.4 Six Sigma Methodology
- 1.1.5 Roles and Responsibilities

1.2 Six Sigma Fundamentals

- 1.2.1 Defining a Process
- 1.2.2 VOC and CTQs
- 1.2.3 QFD
- 1.2.4 Cost of Poor Quality (COPQ)
- 1.2.5 Pareto Analysis (80 : 20 rule)

1.3 Lean Six Sigma Projects

- 1.3.1 Six Sigma Metrics
- 1.3.2 Business Case and Charter
- 1.3.3 Project Team Selection
- 1.3.4 Project Risk Management
- 1.3.5 Project Planning

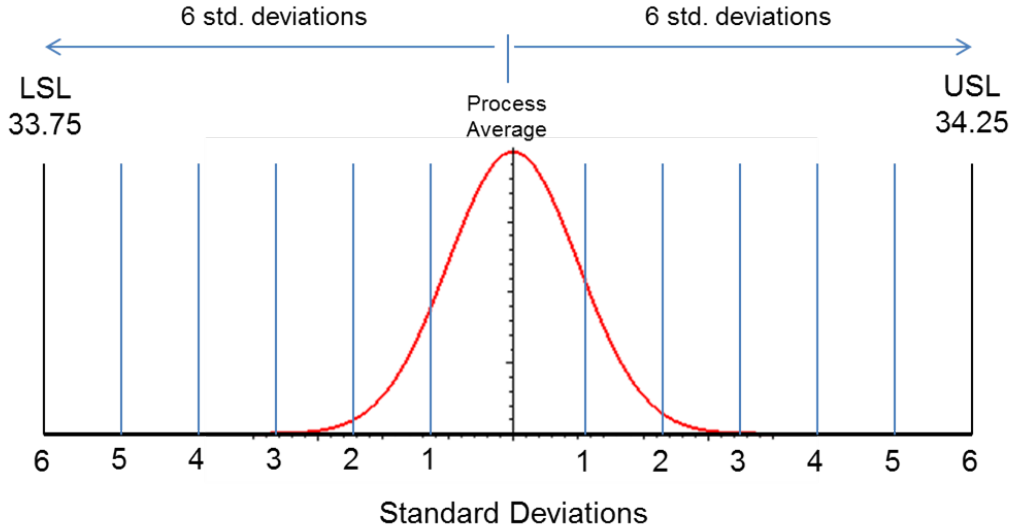
1.4 Lean Fundamentals

- 1.4.1 Lean and Six Sigma
- 1.4.2 History of Lean
- 1.4.3 The Seven Deadly Muda
- 1.4.4 Five-S (5S)

1.1.1 What is Six Sigma

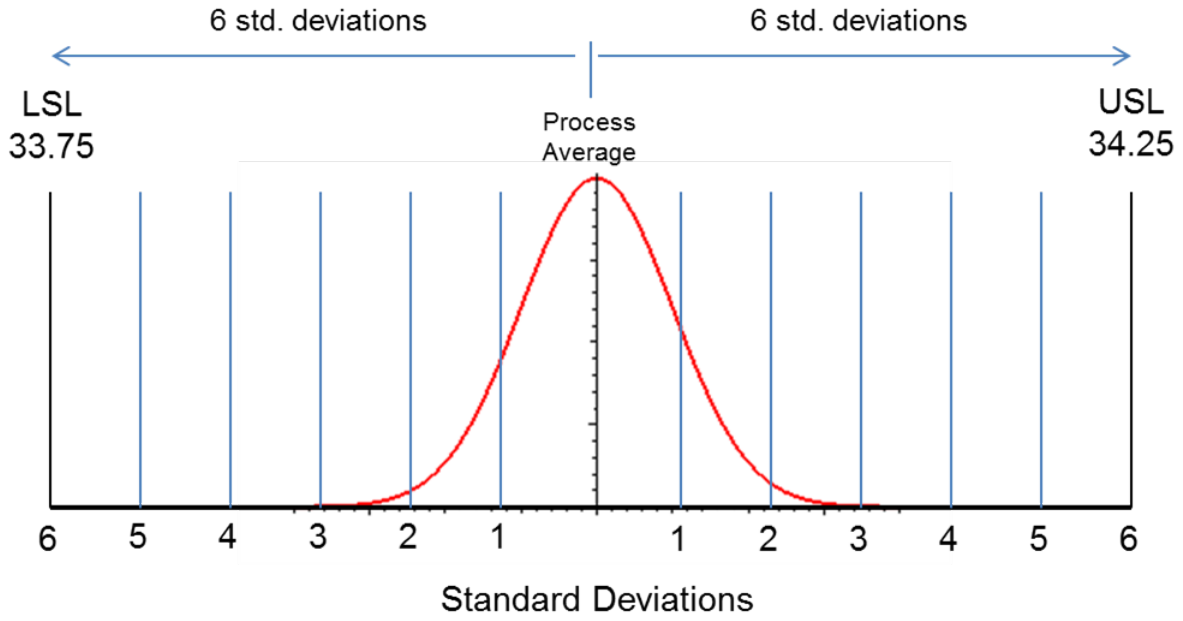
What is Six Sigma?

- What is “sigma”?
  - In statistics, **sigma** (σ) refers to “standard deviation,” which is a measure of variation.
  - You will come to learn that variation is the enemy of any quality process. We need to understand, manage, and minimize process variation.
- What is “Six Sigma”?
  - **Six Sigma** is an aspiration or goal of process performance.
  - A Six Sigma “goal” is for a process average to operate approximately 6σ away from customer’s high and low specification limits.



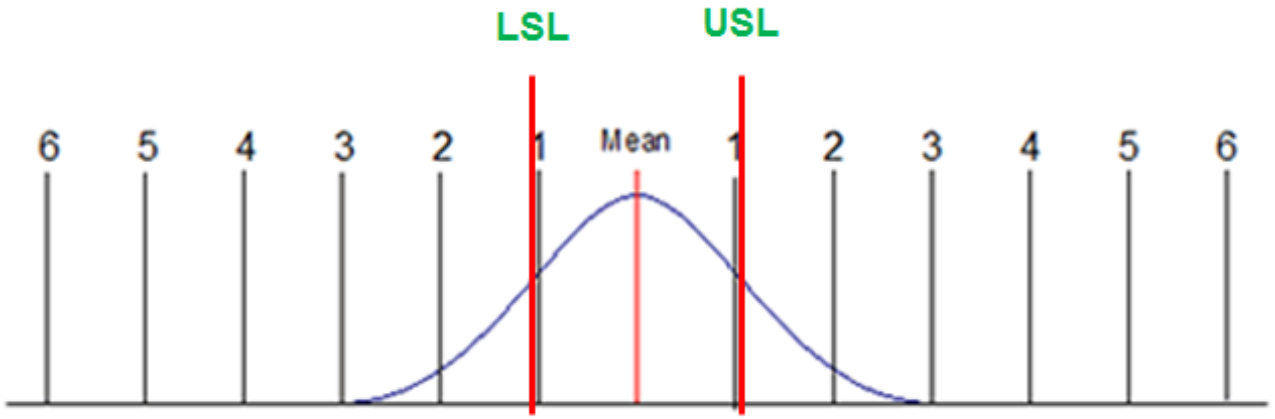
What is Six Sigma?

- A process whose average is about 6σ away from the customer's high and low specification limits has abundant room to "float" before approaching the customer's specification limits.
- A Six Sigma process only yields 3.4 defects for every million opportunities! In other words, 99.9997% of the products are defect-free!



What is Six Sigma: Sigma Level

- **Sigma level** measures how many “sigma” there are between your process average and the nearest customer specification.
- Let us assume that your customers upper and lower specifications limits (USL & LSL) were narrower than the width of your process spread.
- The USL & LSL below stay about 1 standard deviation away from the process average. Therefore, this process operates at **1 sigma**.



What is Six Sigma: Sigma Level

- A process operating at 1 sigma has a defect rate of approximately 70%.

  - This means that the process will generate defect-free products only 30% of the time.

- What about processes with more than 1 sigma level?

- A higher sigma level means a lower defect rate.

- Let us take a look at the defect rates of processes at different sigma levels.

What is Six Sigma: Sigma Level

- This table shows each sigma level's corresponding defect rate and DPMO (defects per million opportunities).
- The higher the sigma level, the lower the defective rate and DPMO.

Sigma Level	Defect Rate	DPMO
1	69.76%	697612
2	30.87%	308770
3	6.68%	66810
4	0.62%	6209
5	0.023%	232
6	0.00034%	3.4

These Defect Rates Assume a 1.5 sigma shift

- How does this translate into things you might easily relate to?

What is Six Sigma: Sigma Level

- Let us take a look at processes operating at 3 sigma.
- 3 sigma processes have a defect rate of approximately 7%. What would happen if processes operated at 3 sigma?
  - Virtually no modern computer would function*.
  - 10,800,000 health care claims would be mishandled each year.
  - 18,900 US savings bonds would be lost every month.
  - 54,000 checks would be lost each night by a single large bank.
  - 4,050 invoices would be sent out incorrectly each month by a modest-sized telecommunications company.
  - 540,000 erroneous call details would be recorded each day from a regional telecommunications company.
  - 270 million erroneous credit card transactions would be recorded each year in the United States.

*(*http://www.qualityamerica.com)*

What is Six Sigma: Sigma Level

- What if processes operated with 1% defect rate?
  - 20,000 lost articles of mail per hour*.
  - Unsafe drinking water almost 15 minutes per day.
  - 5,000 incorrect surgical operations per week.
  - Short or long landings at most major airports each day.
  - 200,000 wrong drug prescriptions each year.
  - No electricity for almost 7 hours per month.

- Even at 1% defect rate, some processes would be unacceptable to you and many others.

- **So what is Six Sigma?**
  - Sigma level is the measure!
  - Six is the goal!

(Implementing Six Sigma – Forest W. Breyfogle III)*

What is Six Sigma: The Methodology

- Six Sigma itself is the **goal**, not the method.

- In order to achieve Six Sigma, you need to improve your process performance by:
  - Minimizing the process variation so that your process has enough room to fluctuate within customer's spec limits
  - Shifting your process average so that it is centered between your customer's spec limits.

- Accomplishing these two process improvements *(along with stabilization and control),* you can achieve Six Sigma.

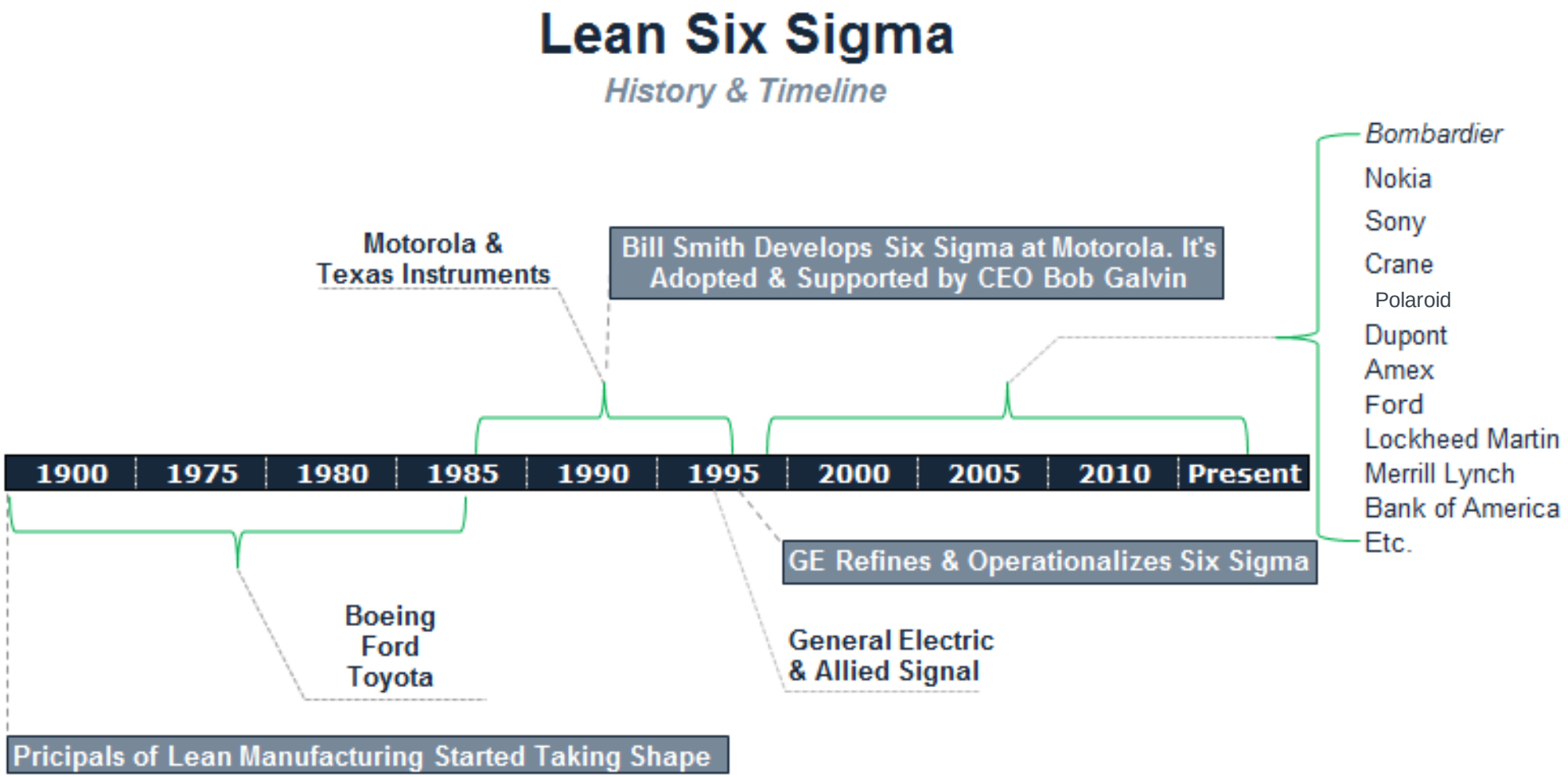
- DMAIC is the systematic methodology prescribed to achieve Six Sigma.

What is Six Sigma: The Methodology

- DMAIC is a systematic and rigorous methodology that can be applied to any process in order to achieve Six Sigma.
- It consists of 5 phases of a project:
  - **D**efine
  - **M**easure
  - **A**nalyze
  - **I**mprove
  - **C**ontrol.
- You will be heavily exposed to many concepts, tools, and examples of the DMAIC methodology through this training.
- You will be capable of applying the DMAIC methodology to improve the performance of any process at the completion of the curriculum.

1.1.2 Six Sigma History

Six Sigma History



Six Sigma History

- The “Six Sigma” terminology was originally adopted by Bill Smith at Motorola in the late 1980s as a quality management methodology.
- As the “Father of Six Sigma,” Bill forged the path for Six Sigma through Motorola’s CEO Bob Galvin who strongly supported Bill’s passion and efforts.
- Starting from the late 1980s, Motorola extensively applied Six Sigma as a process management discipline throughout the company, leveraging Motorola University.
- In 1988, Motorola was recognized with the prestigious Malcolm Baldrige National Quality Award for its achievements in quality improvement.

Six Sigma History

- Six Sigma has been widely adopted by companies as an efficient way of improving the business performance since General Electric implemented the methodology under the leadership of Jack Welch in the 1990s.

- As GE connected Six Sigma results to its executive compensation and published the financial benefits of Six Sigma implementation in their annual report, Six Sigma became a highly sought-after discipline of quality.

Six Sigma History

- Most Six Sigma programs cover the aspects, tools, and topics of Lean or Lean Manufacturing.

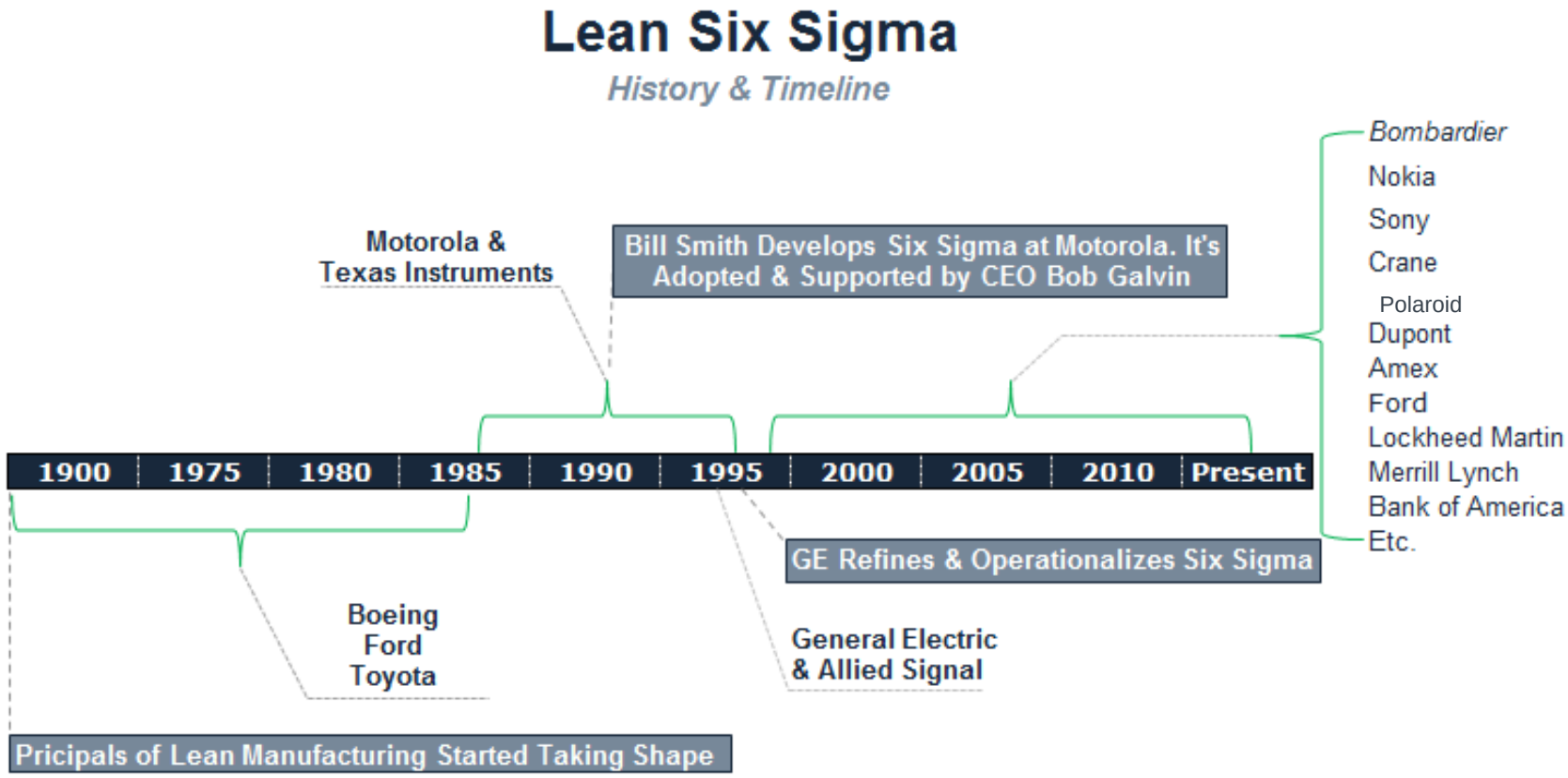
- The two work hand in hand, benefitting each other.
  - Six Sigma focuses on minimizing process variability, shifting the process average, and delivering within customer's specification limits.
  - Lean focuses on eliminating waste and increasing efficiency.

- Lean and its popularity began to form and gain significant traction in the mid 1960s with the Toyota initiative "TPS" or Toyota Production System.

- The concepts and methodology of Lean, however, were fundamentally applied much earlier by both Ford and Boeing in the early 1900s.

Six Sigma History

- Despite the criticism and immaturity of Six Sigma in many aspects, its history continues to be written with every company and organization striving to improve its business performance.



1.1.3 Six Sigma Approach

Six Sigma Approach: Y = f(x)

- The Six Sigma approach to problem solving uses a transfer function.
- A **transfer function** is a mathematical expression of the relationship between the inputs and outputs of a system.
- **Y = f(x)** is the relational transfer function that is used by all Six Sigma practitioners.
- It is absolutely critical that you understand and embrace this concept.

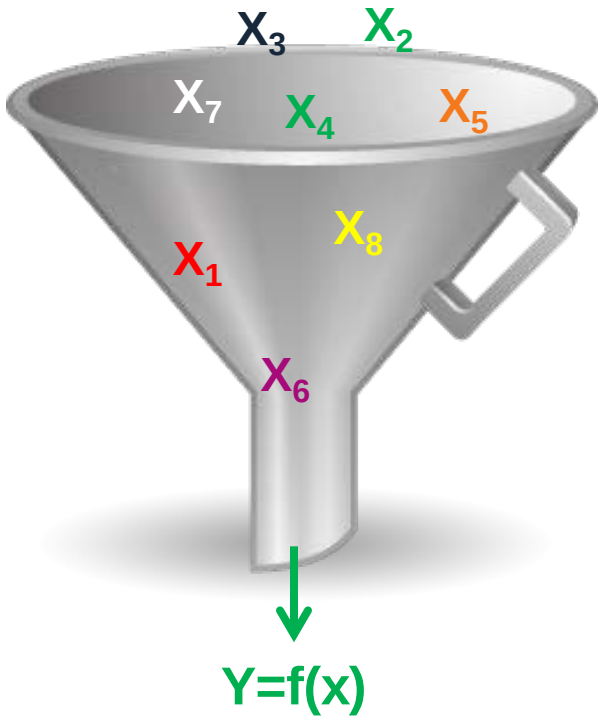
Six Sigma Approach: Y = f(x)

- “Y” refers to the measure or output of a process.
  - Y is usually your primary metric
  - Y is the measure of process performance that you are trying to improve.
- f(x) means “function of x.”
  - x’s are factors or inputs that affect the Y
- Combined, the Y = f(x) statement reads “Y is a function of x.”
- In simple terms: “My process performance is dependent on certain x’s.”
- The objective in a Six Sigma project is to identify the critical x’s that have the most influence on the output (Y) and adjust them so that the Y improves.

Six Sigma Approach: Y = f(x)

- Let us look at a simple example of a pizza delivery company that desires to meet customer expectations of on-time delivery.
  - Measure = on-time pizza deliveries
    - Y = percent of on-time deliveries
  - f(x) would be the x's or factors that heavily influence timely deliveries
    - x1: might be traffic
    - x2: might be the number of deliveries per driver dispatch
    - x3: might be the accuracy of directions provided to the driver
    - x4: might be the reliability of the delivery vehicle
    - etc.
- The statement Y = f(x) in this example will refer to the proven x's determined through the steps of a Six Sigma project.

Six Sigma Approach: Y = f(x)



- With this approach, all potential x's are evaluated throughout the DMAIC methodology.
- The x's should be narrowed down until the vital few x's that significantly influence "on-time pizza deliveries" are identified!

Six Sigma Approach: Y = f(x)

- This approach to problem solving will take you through the process of determining all potential x's that **might** influence on-time deliveries and then determining through measurements and analysis which x's **do** influence on-time deliveries.
- Those significant x's become the ones used in the Y = f(x) equation.

- The Y = f(x) equation is a very powerful concept and requires the ability to measure your output and quantify your inputs.
- Measuring process inputs and outputs is crucial to effectively determining the significant influences to any process.

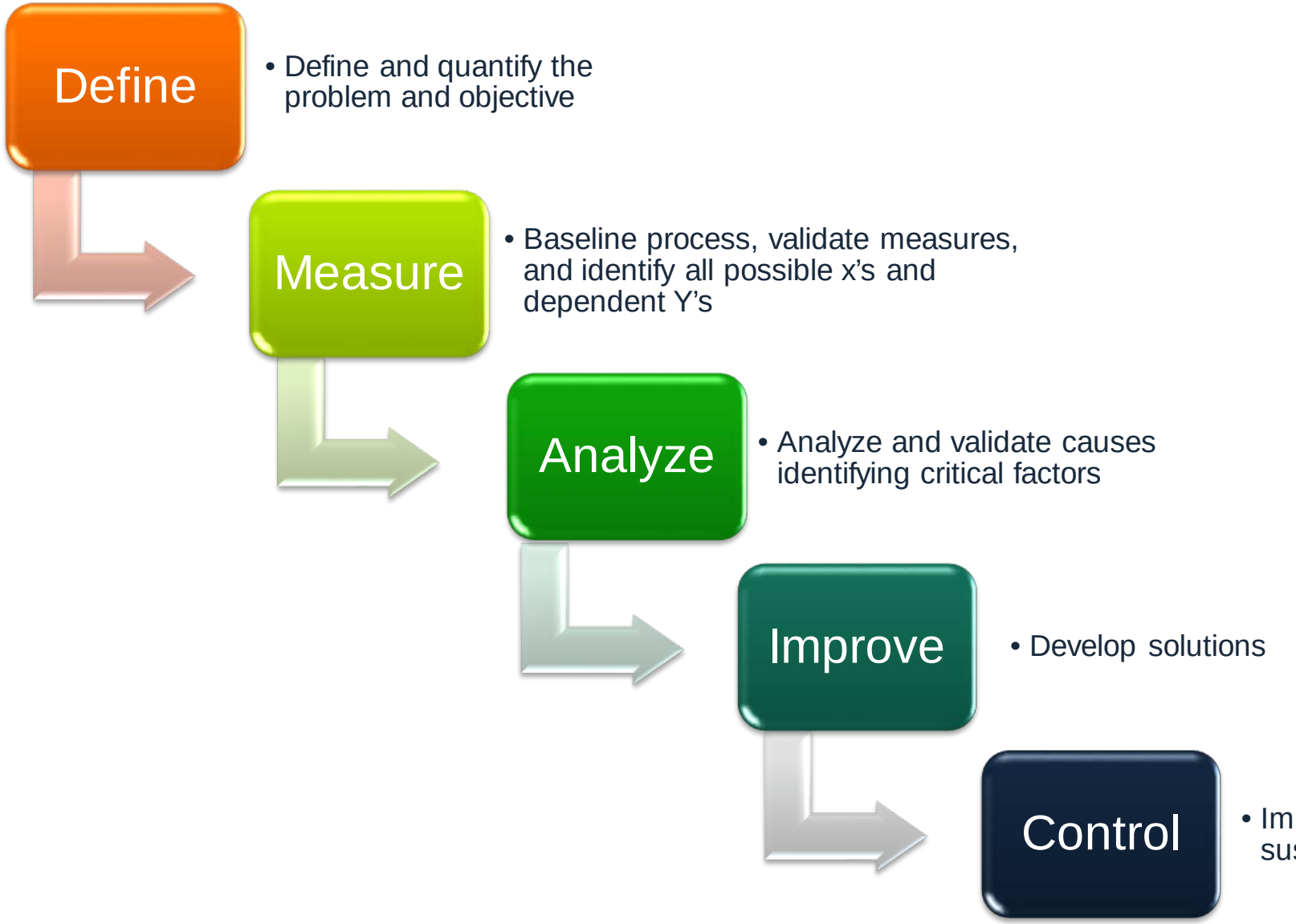
1.1.4 Six Sigma Methodology

Six Sigma Methodology

- Six Sigma follows a methodology that is conceptually rooted in the principles of a five-phase project.
- Each phase has a specific purpose and specific tools and techniques that aid in achieving the phase objectives.

- The 5 phases of DMAIC:
  1. **D**efine
  2. **M**easure
  3. **A**nalyze
  4. **I**mprove
  5. **C**ontrol

Six Sigma Methodology



Six Sigma Methodology: Define Phase

- The goal of the **Define** phase is to establish a solid foundation and business case for a Six Sigma project.
- Define is arguably the most important aspect of any Six Sigma project.
- All successful projects start with a current state challenge or problem that can be articulated in a quantifiable manner.
  - It is not enough to just know the problem, you must quantify it and also determine the goal.
- Once problems and goals are identified and quantified, the rest of the define phase will be about valuation, team, scope, project planning, timeline, stakeholders, Voice Of the Customer (VOC), and Voice Of the Business (VOB).

Six Sigma Methodology: Define Phase

- **Define Phase Tools and Deliverables**
  - Project Charter – Establish the:
    - Business Case
    - Problem Statement
    - Project Objective
    - Project Scope
    - Project Timeline
    - Project Team.
  - Stakeholder Assessment
  - High-Level Pareto Chart Analysis
  - High-Level Process Map
  - VOC/VOB and CTQs Identified and Defined
  - Financial Assessment

Six Sigma Methodology: Measure Phase

- The goal of the **Measure** phase is to gather baseline information about the process (process performance, inputs, measurements, customer expectations etc.).
- Throughout the Measure phase you will seek to achieve a few important objectives:
  - Gather All Possible x's
  - Assess Measurement System and Data Collection Requirements
  - Validate Assumptions
  - Validate Improvement Goals
  - Determine COPQ (Cost of Poor Quality)
  - Refine Process Understanding
  - Determine Process Stability
  - Determine Process Capability.

Six Sigma Methodology: Measure Phase

- **Measure Phase Tools and Deliverables**
  - Process Maps, SIPOC, Value Stream Maps
  - Failure Modes and Effects Analysis (FMEA)
  - Cause-and-Effect Diagram
  - XY Matrix
  - Six Sigma Statistics
    - Basic Statistics
    - Descriptive Statistics
  - Measurement Systems Analysis
    - Variable and/or Attribute Gage R&R
    - Gage Linearity and Accuracy or Stability
  - Basic Control Charts
  - Process Capability (Cpk, Ppk) and Sigma Levels
  - Data Collection Plan

Six Sigma Methodology: Analyze Phase

- The **Analyze** phase is all about establishing verified drivers.

- In the DMAIC methodology, the Analyze phase uses statistics and higher-order analytics to discover relationships between process performance and process inputs (in other words, what are the root causes or drivers of the improvement effort).

- Ultimately, the Analyze phase establishes a reliable hypothesis for improvement solutions.
  - Establish the Transfer Function Y = f(x)
  - Validate the List of Critical x's and Impacts
  - Create a Beta Improvement Plan (e.g., pilot plan).

Six Sigma Methodology: Analyze Phase

- **Analyze Phase Tools and Deliverables**
  - The Analyze phase is about proving and validating critical x's using the appropriate and necessary analysis techniques. Examples include:
    - Hypothesis Testing
      - Parametric and Non-Parametric
    - Regression
      - Simple Linear Regression
      - Multiple Linear Regression
  - The Analyze phase is also about establishing a set of solution hypotheses to be tested and further validated in the Improve phase.

Six Sigma Methodology: Improve Phase

- The goal of the **Improve** phase is. . .you guessed it! "make the improvement." Improve is about designing, testing, and implementing your solution.
- To this point you have defined the problem and objective of the project, brainstormed possible x's, analyzed and verified critical x's. Now it's time to make it real!
  - Statistically Proven Results from Active Study/Pilot
  - Improvement/Implementation Plan
  - Updated Stakeholder Assessment
  - Revised Business Case with Return on Investment (ROI)
  - Risk Assessment/Updated FMEA
  - New Process Capability and Sigma.

Six Sigma Methodology: Improve Phase

- **Improve Phase Tools and Deliverables**
  - Any Appropriate Tool from Previous Phases
  - Design of Experiment (DOE)
    - Full Factorial
    - Fractional Factorial
  - Pilot or Planned Study Using:
    - Hypothesis Testing
    - Valid Measurement Systems
  - Implementation Plan

Six Sigma Methodology: Control Phase

- The last of the 5 core phases of the DMAIC methodology is the **Control** phase.
- The goal of the Control phase is to establish automated and managed mechanisms to maintain and sustain your improvement.
- A successful control plan also establishes a reaction and mitigation plan as well as an accountability structure.

Six Sigma Methodology: Control Phase

- **Control Phase Tools and Deliverables**
  - Statistical Process Control (SPC/Control Charts)
    - IMR, XbarS, XbarR, P, NP, U, C etc.
  - Control Plan Documents
    - Control Plan
    - Training Plan
    - Communication Plan
    - Audit Checklist
  - Lean Control Methods
    - Poka-Yoke
    - Five-S
    - Kanban

Six Sigma Methodology

Six Sigma DMAIC Roadmap

Define
- **Goal**: Problem Statement, Objective, Business Case, Project Scope, Team
- **Main Tools**: Project Charter, Pareto, Process Maps

Measure
- **Goal**: Brainstorm/Prioritize Possible x's, Validate measurement, Capability
- **Tools**: Basic Statistics, C & E, XY Matrix, Capability Analysis, MSA, Process Maps, Control Charts

Analyze
- **Goal**: Identify critical x's
- **Tools**: Hypotheses Tests (Normal/Non Normal), Regression and Correlation

Improve
- **Goal**: Design, Test, and Implement Improvement
- **Tools**: DOE, Implementation/Change/Communication Plan

Control
- **Goal**: Lock-in the Improvement
- **Tools**: Control Plan, Poka-Yoke, SPC, SOPs, Training Plans etc.

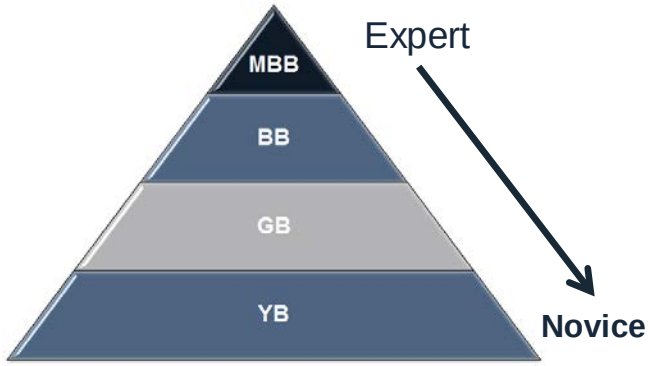
1.1.5 Roles and Responsibilities

Roles and Responsibilities

- The various roles in a Six Sigma program are commonly referred to as “Belts.”
- In addition to Belts, there are also other key roles with specific responsibilities.
- Let us explore the different roles and their corresponding responsibilities in a Six Sigma program.

Roles and Responsibilities

- Each of the four Six Sigma belts represents a different level of expertise in the field of Six Sigma.
  - Six Sigma Master Black Belt (MBB)
  - Six Sigma Black Belt (BB)
  - Six Sigma Green Belt (GB)
  - Six Sigma Yellow Belt (YB)



- In addition to Belts, there are other critical and complementary roles:
  - Champions
  - Sponsors
  - Stakeholders
  - Subject Matter Experts (SMEs).

Roles and Responsibilities: MBB

- The **Master Black Belt** (MBB) is the most experienced, educated, and capable Six Sigma expert.
- A typical MBB has managed dozens of Black Belt level projects.
- The MBB can simultaneously lead multiple Six Sigma Belt projects while mentoring and certifying Black Belt and Green Belt candidates.
- The MBB typically works with high-level operations directors, senior executives, and business managers to help with assessing and planning business strategies and tactics.

Roles and Responsibilities: MBB

- MBB commonly advises management team on the cost of poor quality of an operation and consults on methods to improve business performance.

- **Typical Responsibilities of a MBB**
  - Identifies and defines the portfolio of projects required to support a business strategy
  - Establishes scope, goals, timelines, and milestones
  - Assigns and marshals resources
  - Trains and mentors Green Belts and Black Belts
  - Facilitates tollgates or checkpoints for Belt candidates
  - Reports-out/updates stakeholders and executives
  - Establishes organization's Six Sigma strategy/roadmap
  - Leads the implementation of Six Sigma.

Roles and Responsibilities: BB

- The **Black Belt** (BB) is the most active and valuable experienced Six Sigma professional among all the Six Sigma Belts.

- A typical BB has
  - led multiple projects
  - trained and mentored various Green Belts candidates
  - understood how to define a problem and drive effective solution.

- The BB is well rounded in terms of project management, statistical analysis, financial analysis, meeting facilitation, prioritization, and a range of other value-added capabilities, which makes a BB highly valuable asset in the business world.

Roles and Responsibilities: BB

- BBs commonly serves as the dedicated resource continuing their line management role while simultaneously achieving a BB certification.

- **Typical Responsibilities of a BB**
  - Project Management
    - Defines projects, scope, teams etc.
    - Marshals resources
    - Establishes goals, timelines, and milestones
    - Provides reports and/or updates to stakeholders and executives.

Roles and Responsibilities: BB

- **Typical Responsibilities of a BB** *(continued)*
  - Task Management
    - Establishes the team's Lean Sigma roadmap
    - Plans and implements the use of Lean Sigma tools
    - Facilitates project meetings
    - Does project management of the team's work
    - Manages progress toward objectives.
  - Team Management
    - Chooses or recommend team members
    - Defines ground rules for the project team
    - Coaches, mentors, and directs project team
    - Coaches other Six Sigma Belts
    - Manages the team's organizational interfaces.

Roles and Responsibilities: GB

- The **Green Belt** (GB) is considered as a less intense version of Six Sigma professional than the Black Belt (BB).
- A GB is exposed to all the comprehensive aspects of Six Sigma with less focus on the statistical theories and some other advanced analytical methodologies such as Design of Experiment (DOE).
- When it comes to project management, a GB has almost the same responsibilities as a BB.
- In general, the GB works on less complicated and challenging business problems than a BB.

Roles and Responsibilities: GB

- **Typical Responsibilities of a Green Belt**
  - Project Management
    - Defines the project, scope, team etc.
    - Marshals resources
    - Sets goals, timelines, and milestones
    - Reports-out/updates stakeholders and executives.
  - Task Management
    - Establishes the team's Lean Sigma Roadmap
    - Plans and implements the use of Lean Sigma tools
    - Facilitates project meetings
    - Does Project Management of the team's work
    - Manages progress toward objectives.
  - Team Management
    - Chooses or recommends team members
    - Defines ground rules for the project team
    - Coaches, mentors, and directs project team
    - Coaches other Six Sigma Belts
    - Manages the team's organizational interfaces.

Roles and Responsibilities: YB

- The **Yellow Belt** (YB) understands the basic objectives and methods of a Six Sigma project.
- YB has an elementary understanding about what other Six Sigma Belts (GB, BB, MBB) are doing to help them succeed.
- In a Six Sigma project, YB usually serves as a subject matter expert regarding some aspects of the process or project.
- Supervisors, managers, directors, and sometimes executives are usually trained at the YB level.

Roles and Responsibilities: YB

- **Typical Responsibilities of a Yellow Belt**
  - Helps define process scope and parameters
  - Contributes to team selection process
  - Assists in information and data collection
  - Participates in experiential analysis sessions (FMEA, Process Mapping, Cause and Effect etc.)
  - Assists in assessing and developing solutions
  - Delivers solution implementations.

Roles and Responsibilities: Champions & Sponsors

- **Champions and sponsors** are those individuals (directors, executives, managers etc.) chartering, funding, or driving the Six Sigma projects that BBs and GBs are conducting.

- Champions and sponsors need to have a basic understanding of the concepts, tools, and techniques involved in the DMAIC methodology so that they can provide proper support and direction.

Roles and Responsibilities: Champions & Sponsors

- Champions and sponsors play critical roles in the successful deployment of Six Sigma.
- Strong endorsement of Six Sigma from the leadership team is critical for success.

- **Typical Responsibilities of a Champion/Sponsor**
  - Maintains a strategic oversight
  - Establishes strategy and direction for a portfolio of projects
  - Clearly defines success
  - Provides resolution for issues such as resources or politics
  - Establishes routine tollgates or project reviews
  - Clears the path for solution implementation
  - Assists in project team formation.

Roles and Responsibilities: Stakeholders

- **Stakeholders** are usually the recipients or beneficiaries of the success of a Six Sigma project.

- Stakeholders are individuals owning the process, function, or production/service line that a Six Sigma Belt focuses on improving the performance of.

- BBs and GBs need to keep strong working relationships with stakeholders because without their support, it would be extremely difficult to make the Six Sigma project a success.

Roles and Responsibilities: SMEs

- **Subject Matter Experts** (SMEs) are commonly known as the experts of the process or subject matter.

- Six Sigma Belts should proactively look to key SMEs to round out their working project team.

- SMEs play critical roles to the success of a project.
  - Based on SMEs' extensive knowledge about the process, they have the experience to identify which solutions can work and which cannot work.
  - SMEs who simply do not speak up can hurt the chances of the process' success.
  - SMEs are also the same people who prefer to keep the status quo. Six Sigma Belts may find many of them unwilling to help implement the changes.

Roles and Responsibilities

- Throughout this module we have reviewed the various common roles and corresponding responsibilities in any Six Sigma program:
  - Six Sigma Master Black Belt
  - Six Sigma Black Belt
  - Six Sigma Green Belt
  - Six Sigma Yellow Belt
  - Champion and Sponsors
  - Stakeholders
  - Subject Matter Experts (SMEs)

- These Six Sigma belts and other roles are designed to deliver value to the business effectively and successfully.

1.2. Six Sigma Fundamentals

Black Belt Training: Define Phase

1.1 Six Sigma Overview

- 1.1.1 What is Six Sigma
- 1.1.2 Six Sigma History
- 1.1.3 Six Sigma Approach Y = f(x)
- 1.1.4 Six Sigma Methodology
- 1.1.5 Roles and Responsibilities

1.2 Six Sigma Fundamentals

- 1.2.1 Defining a Process
- 1.2.2 VOC and CTQs
- 1.2.3 QFD
- 1.2.4 Cost of Poor Quality (COPQ)
- 1.2.5 Pareto Analysis (80 : 20 rule)

1.3 Lean Six Sigma Projects

- 1.3.1 Six Sigma Metrics
- 1.3.2 Business Case and Charter
- 1.3.3 Project Team Selection
- 1.3.4 Project Risk Management
- 1.3.5 Project Planning

1.4 Lean Fundamentals

- 1.4.1 Lean and Six Sigma
- 1.4.2 History of Lean
- 1.4.3 The Seven Deadly Muda
- 1.4.4 Five-S (5S)

1.2.1 Defining a Process

Defining a Process

- The basic method of defining and understanding a process is the **process map**.
- Process maps help determine where and how a process begins as well as all the steps and decisions in between.
- By learning the various types and methods of process maps, you can become adept at setting project scopes, identifying value-added and non-value-added steps, identifying problems in a process, etc.
- This module covers:
  - High-level process maps
  - Detailed process maps
  - Functional maps.
- In the Measure section we will touch on several other types and methods of process mapping.

What is a Process Map?

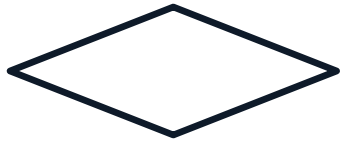
- A **process map** is a graphical representation of a process flow.
- It illustrates how the business process is accomplished step by step.
- It describes how the materials or information sequentially flow from one business entity to the next.
- It illustrates who is responsible for what between the process boundaries.
- It depicts the inputs and outputs of each individual process step.
- Always encourage your project team to map the current state of the process instead of the ideal state. Be honest with each other!

Process Map Basic Symbols

- The following four symbols are the most commonly used symbols in a process map.

Terminator (Oval):
Shows the start and end points in the process.

Process (Rectangle):
Indicates a single process step.



Decision (Diamond):
Indicates a question with two choices (e.g. Yes/No)

Flow Line (Arrow):
Shows the direction of the process flow.

Additional Process Symbols

- Additional Process Symbols:

Alternative Process:
Indicates a process step as an alternate of a normal step.

Predefined Process:
Indicates a formally-defined process step. Other documentation or instruction is needed to support further details of the step.

Manual Operation:
Indicates a process step conducted manually.

Preparation:
Indicates a preparation step.

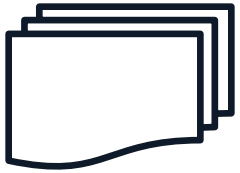
Delay:
Indicates a waiting period in the process.

Additional Process Symbols

- Additional File and Information Related Symbols:

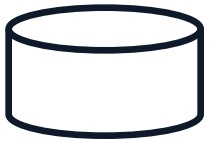
Data (I/O):
Shows the inputs and outputs of a process.

Document:
Indicates a process step that results in a document.



Multi-Document:
Indicates a process step that results in multiple documents.

Stored Data:
Indicates a process step that stores data.



Magnetic Disk:
Indicates a database.

Additional Process Symbols

- Additional Control of Flow Symbols:

Off-Page Connector:
Indicates the process flow continues onto another page.

Merge:
Indicates multiple processes merge into one.

Extract:
Indicates a process splits into multiple parallel processes.

Or:
Indicates a single data processing flow diverges to multiple branches with different criteria requirements.

Summing Junction:
Indicates multiple data processing flows converge into one.

How to Plot a Process Map

- Step 1: Define the boundaries of the process you want to map.
  - A process map can depict the flow of an entire process or a segment of it.
  - You need to identify and define the beginning and ending points of the process before starting to plot.
  - Use operational definitions where possible.

How to Plot a Process Map

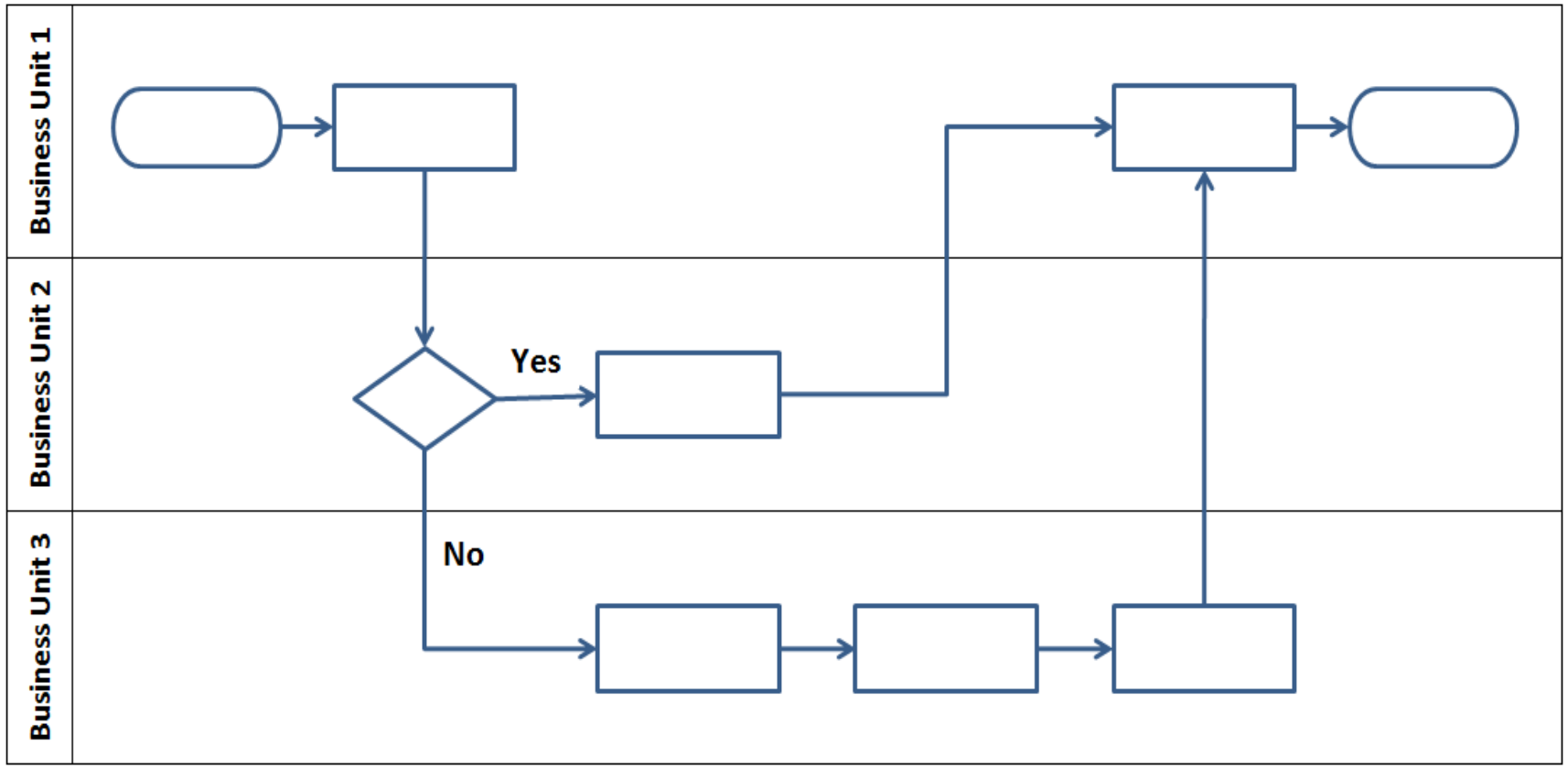
- Step 2: Define and sort the process steps with the flow.
  - Consult with process owners and SMEs or observe the process in action to understand how the process is actually performed.
  - Record the process steps and sort them according to the order of their occurrence.

How to Plot a Process Map

- Step 3: Fill the step information into the appropriate process symbols and plot the diagram.
  - In the team meeting of process mapping, place the sticky notes with different colors on a white board so you can move them around while the map is under-construction.
  - The flow lines can be plotted directly on the white board.
  - Decision steps. Rotate the sticky note 45 degrees.
  - When the map is completed on the white board, record the map using Excel, PowerPoint, Visio, Quality Companion, or other preferred software.

How to Plot a Process Map

- Step 3:
  - To illustrate the responsibility of different organizations involved in the process, use a Swim Lane Process Map.



How to Plot a Process Map

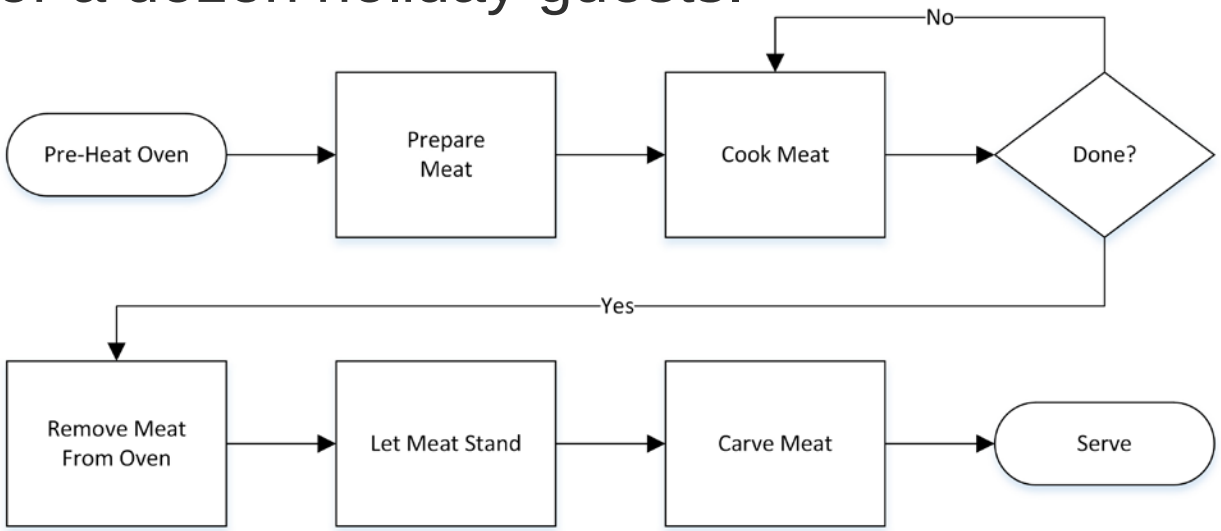
- Step 4: Identify and record inputs/outputs and their corresponding specifications for each process step.
  - The process map helps in understanding and documenting Y = f(x) of a process, where Y represents the outputs and x represents the inputs.
  - The inputs of each process step can be controllable or non-controllable, standardized operational procedures, or noise.
  - Inputs are the source of variation in the process and need to be analyzed qualitatively and quantitatively.
  - The outputs of each process step can be products, information, services, etc. They are the little Y's within the process.

How to Plot a Process Map

- Step 5: Evaluate the process map and adjust if needed.
  - If the process is too complicated to be covered in one single process map, you may create additional detailed sub-process maps for further information.
  - Number the process steps in the order of their occurrence for clarity.

High Level Process Map

- Most high-level process maps are also referred to as **flow charts**.
- The key to a high-level process map is to over-simplify the process being depicted so that it can be understood in its most generic form.
- As a general rule, high-level process maps should be no more than 4–6 steps.
- Below is an oversimplified version of a high-level process map for cooking a 10lb prime rib for a dozen holiday guests.

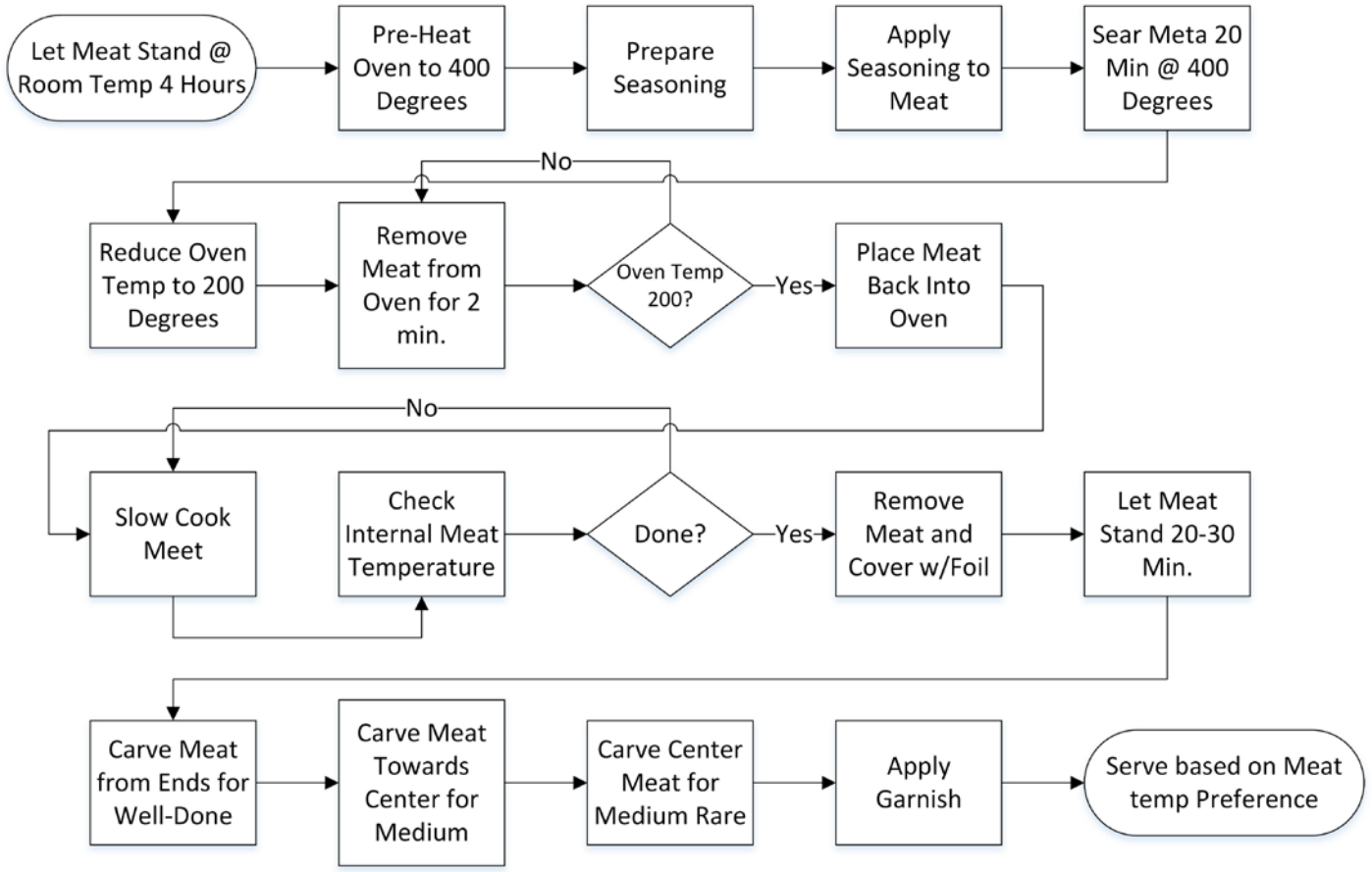


Detailed Process Map

- Detailed process maps or multi-level maps take the high-level map much further.
- Detailed maps can be 2–4 levels deeper than your high-level process map.
- A good guideline used to help create the second level is to take each step in the high level map and break it down into 2–4 steps (no more).
- Repeat this process (level 3, level 4 etc.) until reaching the desired level of detail.
- Some detailed maps are 2 or 3 levels deep, others can be 5–6 levels deep. Obviously, the deeper the levels, the more complex and the more burdensome.

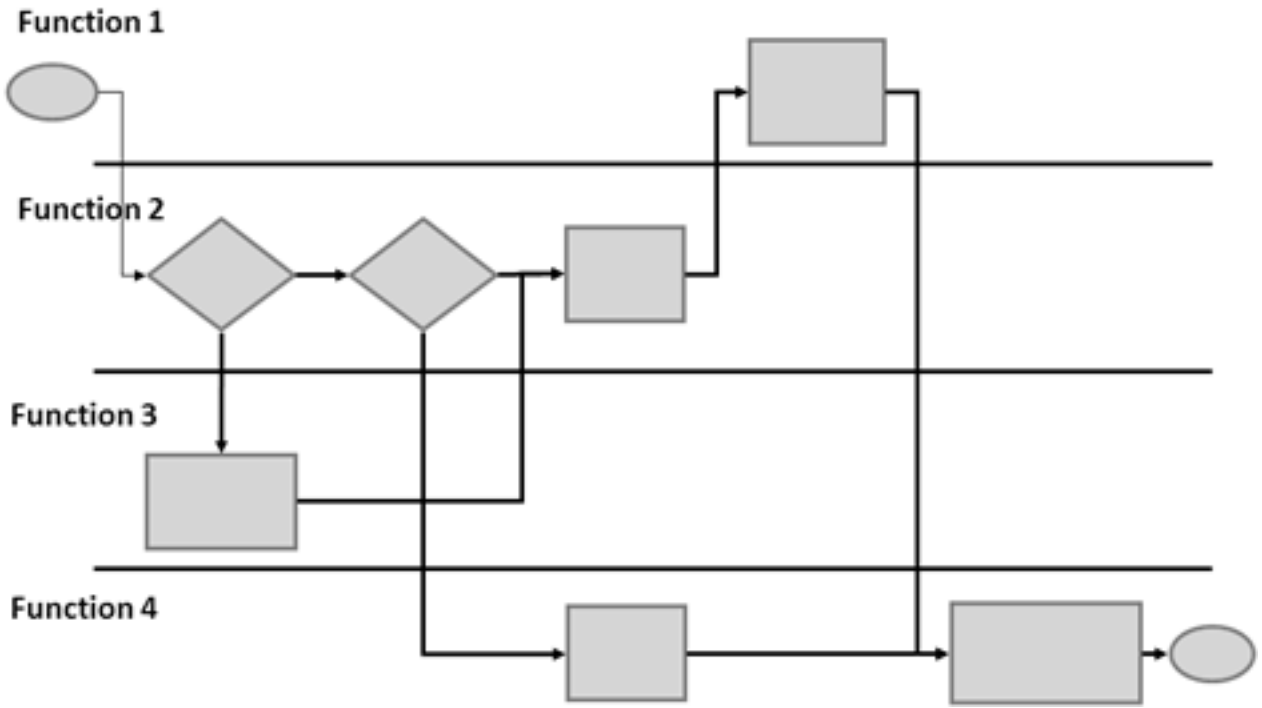
Detailed Process Map

- At right is our prime rib cooking example at **level 2 detail**.
- This process map has a few more decision points and process steps.
- You can see that going only one more level deep adds a fair amount of information to the process map.



Functional Process Map

- The functional map adds dimension to the high-level or detailed map.
- The dimension added is identifying which function or job performs the step or makes the decision.
- At right is a generic example of a functional map. Note that functions are identified in horizontal lanes and each process step is placed in the appropriate lane based on which function performs the step.



1.2.2 VOC and CTQs

Voice of the Customer

- **VOC** stands for “Voice of the Customer.”
- Voice of the customer is a term used for a data-driven plan to discover customer wants and needs.
- VOC is an important component to a successful Six Sigma project.
- There are also other “Voices” that need to be heard when conducting projects. The 3 primary forms are:
  - VOC: Voice of the Customer
  - VOB: Voice of the Business
  - VOA: Voice of the Associate.

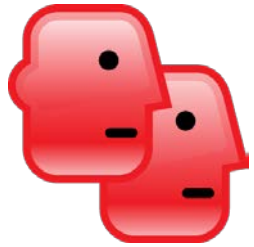
Gathering VOC

- Gathering VOC should be performed methodically.
- The two most popular methods of collecting VOC are
  1. Indirect
  2. Direct.
- 1. Indirect data collection for VOC involves passive information exchange:
  - Warranty claims
  - Customer complaints/compliments
  - Service calls
  - Sales reports.

- 2. Indirect methods are less effective, sometimes dated, require heavy interpretation, and are also more difficult to confirm.

Gathering VOC

- Direct data collection methods for VOC are active and planned customer engagements:
  - Conducting interviews
  - Conducting customer surveys
  - Conducting market research
  - Hosting focus groups.



- Direct methods are typically more effective for several reasons:
  - Less need to interpret meaning
  - Researchers can go a little deeper when interacting with customers
  - Customers are aware of their participation and will respond better upon follow-up
  - Researchers can properly plan engagements (questions, sample size, information collection techniques etc.).

Gathering VOC

- Gathering VOC requires consideration of many factors such as product or services types, customer segments, manufacturing methods or facilities etc.
- All this information will influence the sampling strategy.
- Consider which factors are important and build a sample size plan around them.
- Also, consider response rates and adjust the initial sample strategy to ensure adequate input is received.
- Once a sampling plan is in place, collect data via the direct and indirect methods discussed earlier.
- After gathering VOC it will be necessary to translate it into something meaningful: CTQs.

Critical to Quality: CTQ

- **CTQ** stands for Critical to Quality.
- CTQs are translated from VOC or "voice of the customer" feedback.
- VOC is often vague, emotional, or simply a generalization about products or services.

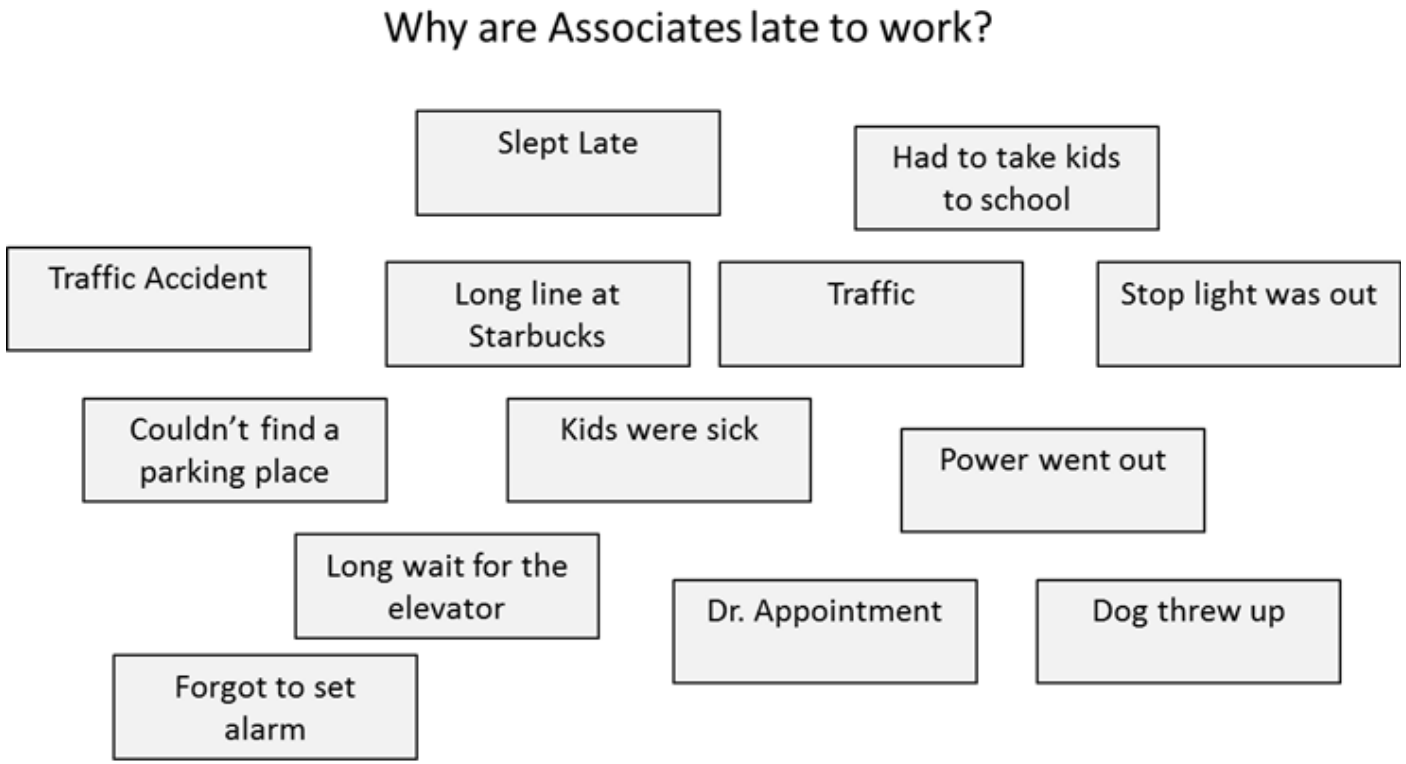
- CTQs are the quantifiable, measureable, and meaningful translations of VOC.
- Organizing VOC helps to identify CTQs.
- One effective way to organize VOC is to group or bucket it using an **affinity diagram**.
- Affinity diagrams are ideal for large amounts of soft data resulting from brainstorming sessions or surveys.

Affinity Diagram: Building a CTQ Tree

- Steps for conducting an Affinity Diagram exercise:
  - Step 1: Clearly define the question or focus of the exercise ("Why are associates late for work?").
  - Step 2: Record all participant responses on note cards or sticky notes (this is the sloppy part, record everything!).
  - Step 3: Lay out all note cards or post the sticky notes onto a wall.
  - Step 4: Look for and identify common themes.
  - Step 5: Begin moving the note cards or sticky notes into the themes until all responses are allocated.
  - Step 6: Re-evaluate and make adjustments.

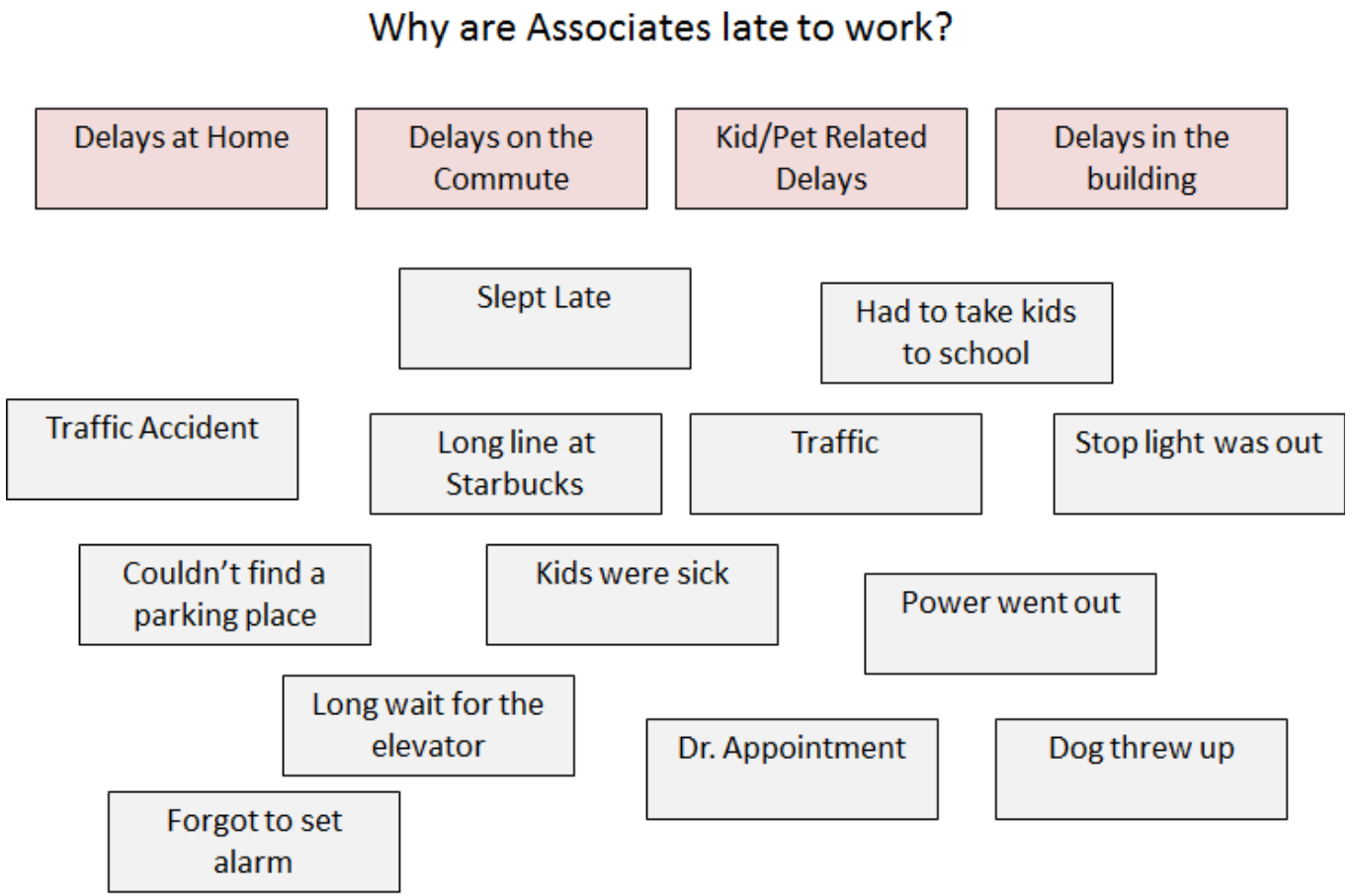
Affinity Diagram: Building a CTQ Tree

- Define the question or focus
- Record responses on note cards or sticky notes
- Display all note cards or sticky notes on a wall if necessary.



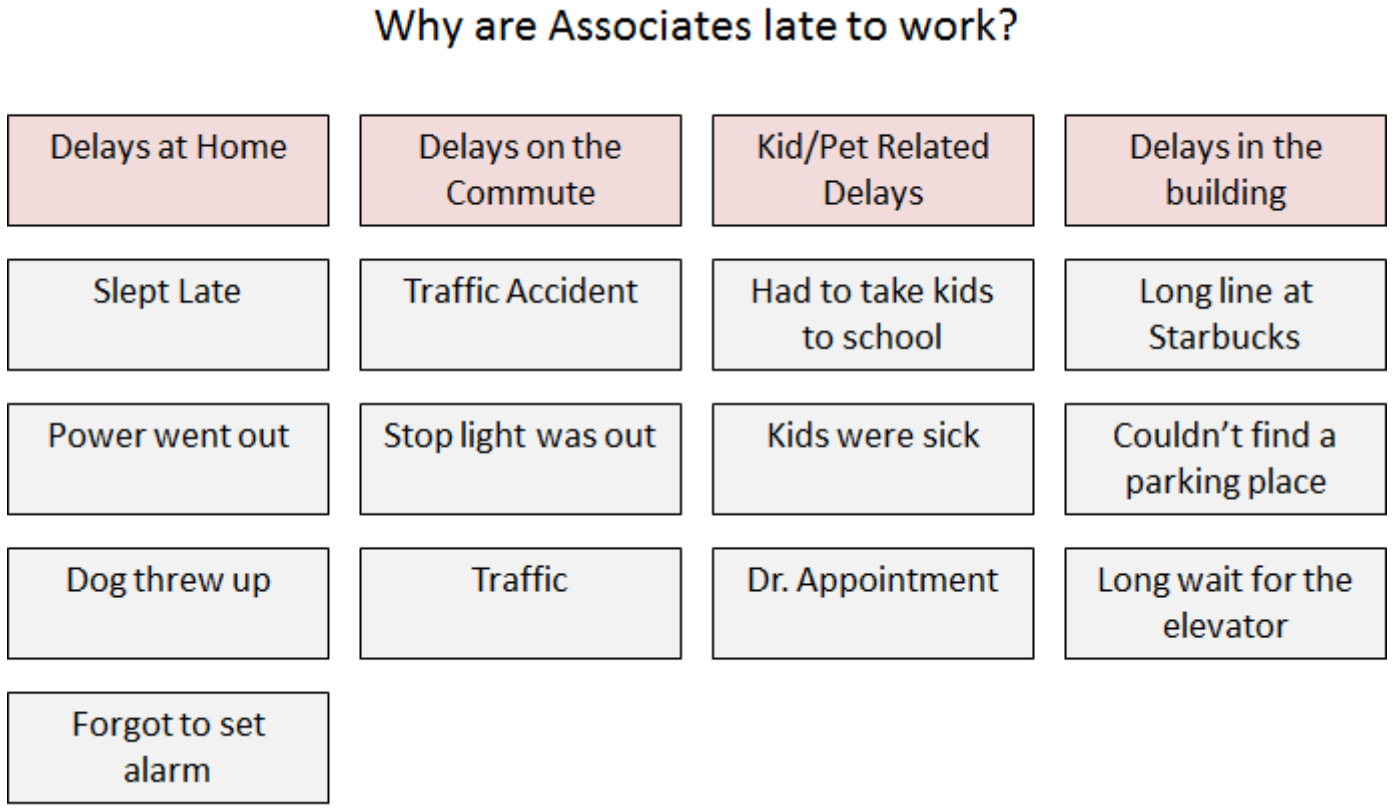
Affinity Diagram: Building a CTQ Tree

- Look for and identify common themes within the responses.



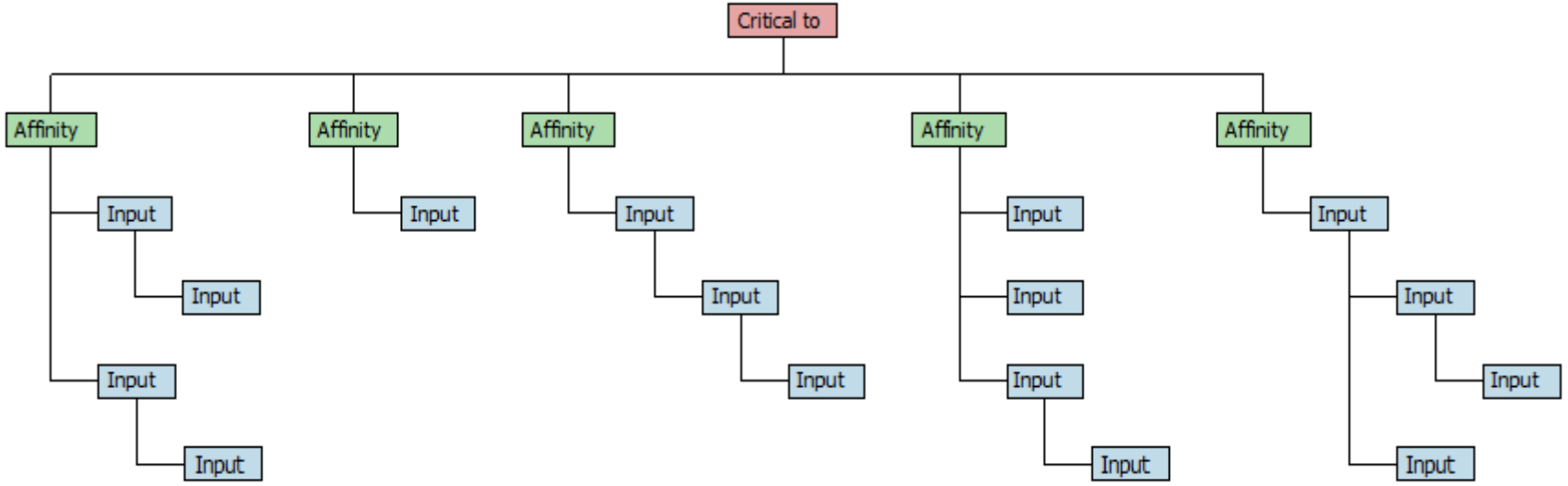
Affinity Diagram: Building a CTQ Tree

- Group note cards or sticky notes into themes until all responses are allocated.
- Re-evaluate and make final adjustments.



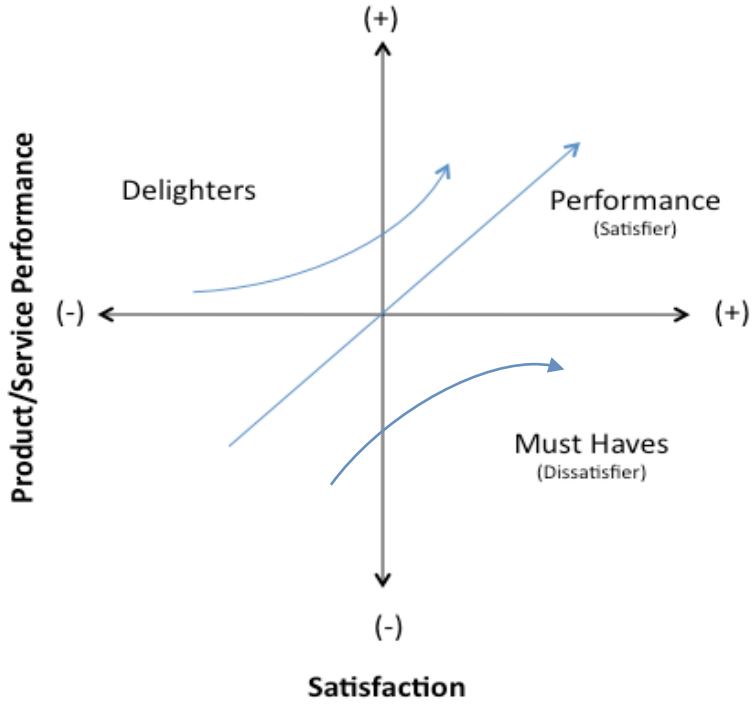
CTQ Tree

- Example of a generic CTQ tree transposed from a white board to a software package.



Kano

- Another VOC categorization technique is the Kano.
- The Kano model was developed by Noriaki Kano in the 1980s.
- The Kano model is a graphic tool that further categorizes VOC and CTQs into 3
  distinct groups:
  - Must Haves
  - Performance Attributes
  - Delighters.



- The Kano helps to identify CTQs that add incrementalvalue vs. those that are simply requirements and having more is not necessarily better.

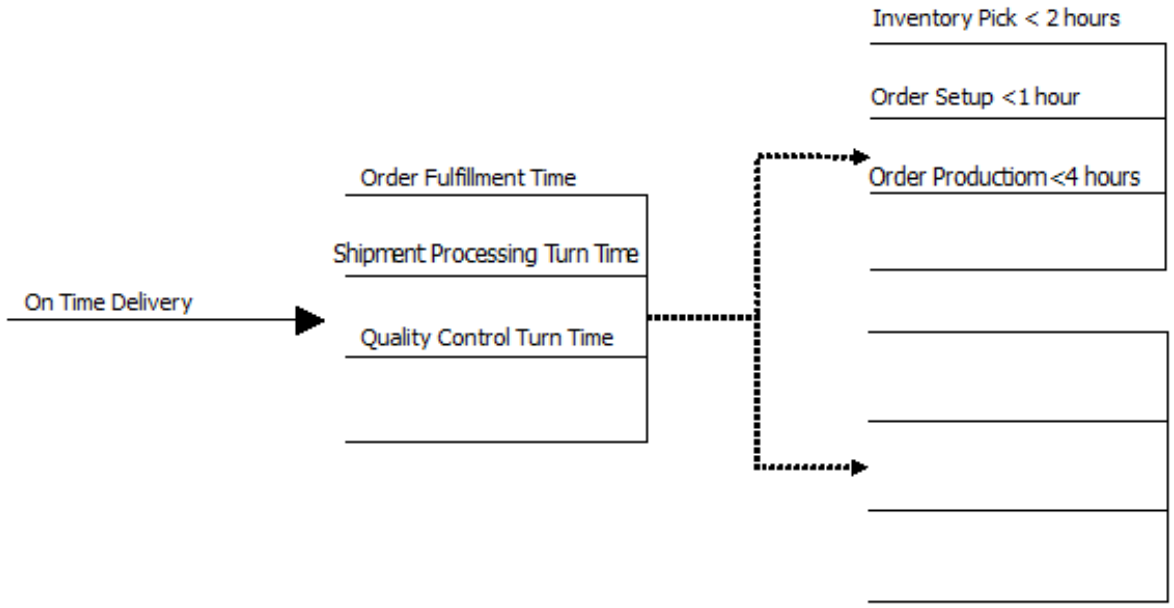
Validating VOC and CTQs

- After determining all CTQs, confirm them with the customer.
- Confirming can be accomplished by conducting surveys through one or more of the following methods:
  - Group sessions
  - One-on-one meetings
  - Phone interviews
  - Electronic means (chat, email, social media etc.)
  - Physical mail.
- Consider your confirming audience and try to avoid factors that may influence or bias responses such as inconvenience or overly burdensome time commitments.

Translating CTQs to Requirements

- Lastly, CTQs must be transformed into **specifics** that can be built upon in a process.

- A **requirements tree** translates CTQs to meaningful and measureable requirements for production processes and products.



1.2.3 Quality Function Deployment

History of QFD

- Developed by Shigeru Mizuno (1910–1989) and Yoji Akao (b. 1928) in Japan. Quality Function Deployment (QFD) aims to design products that assure customer satisfaction and value – the first time and every time.

- The QFD framework can be used for translating actual customer statements and needs (“The voice of the customer”) into actions and designs to build and deliver a quality product.

What is QFD?

- **Quality Function Deployment (QFD)** is a construction methodology and quantification tool used to identify and measure customer's requirements and transform them into meaningful and measureable parameters.

- QFD helps to prioritize actions to advance process or product to meet customer's anticipations.

- QFD is an excellent tool for contact between cross-functional groups.

Purpose of QFD

The quality function deployment has many purposes. Among the most important are:

- Market analysis to establish needs and expectations
- Examination of competitors' abilities
- Identification of key factors for success
- Translation of key factors into product and process characteristics.

Phases of QFD

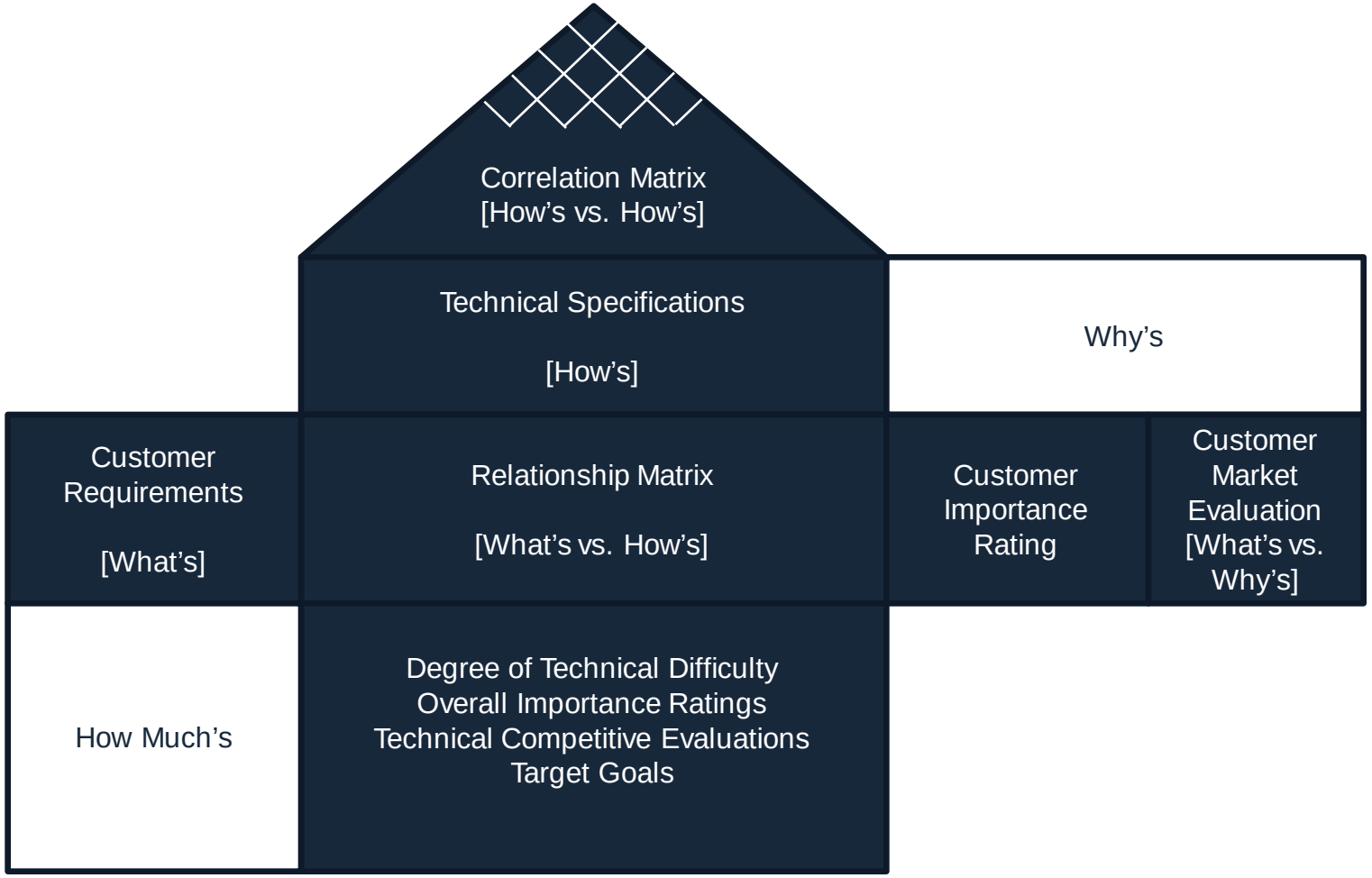
Four Key Phases of QFD

- **Phase I:** Product Planning Including the “House of Quality” (Requirements Engineering Life Cycle)
- **Phase II:** Product Design (Design Life Cycles)
- **Phase III:** Process Planning (Implementation Life Cycle)
- **Phase IV:** Process Control (Testing Life Cycle)

How to build a House of Quality

- Determine Customer Requirements (*What's* from VOC/CTQ)
- Technical Specifications/Design Requirements (*How's*)
- Develop Relationship Matrix (*What's* and *How's*)
- Prioritize Customer Requirements
- Conduct Competitive Assessments
- Develop Interrelationship (*How's*)
- Prioritize Design Requirements

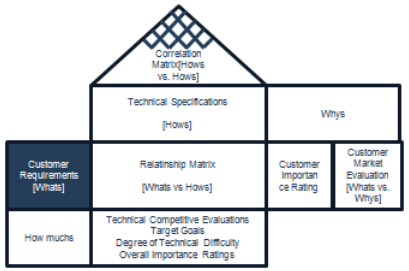
House of Quality



Step 1: Determine Customer Requirements

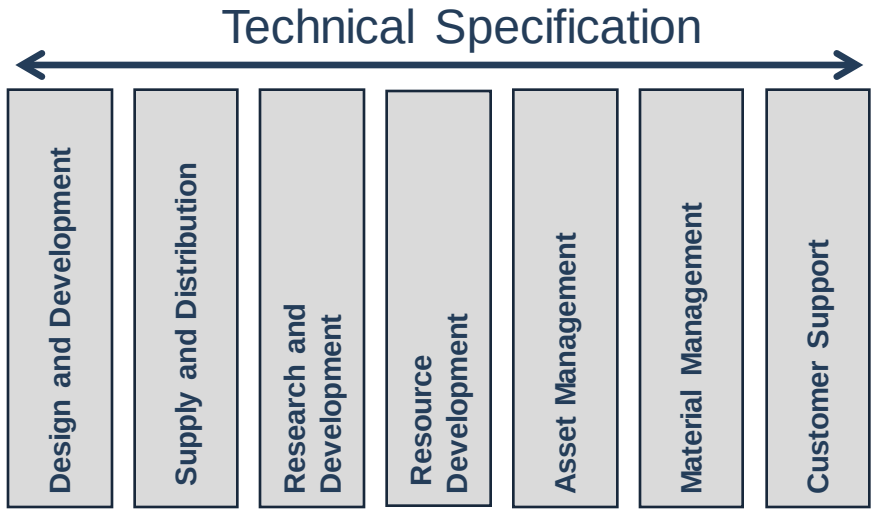
- Identify the important customer requirements. These are the “What’s” and are typically determined through the VOC/CTQ process.

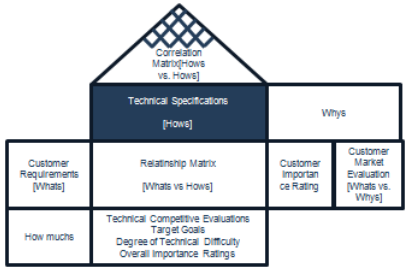
- Use the results from your requirements tree diagram as inputs for the customer requirements in your HOQ.



Step 2: Technical Specification

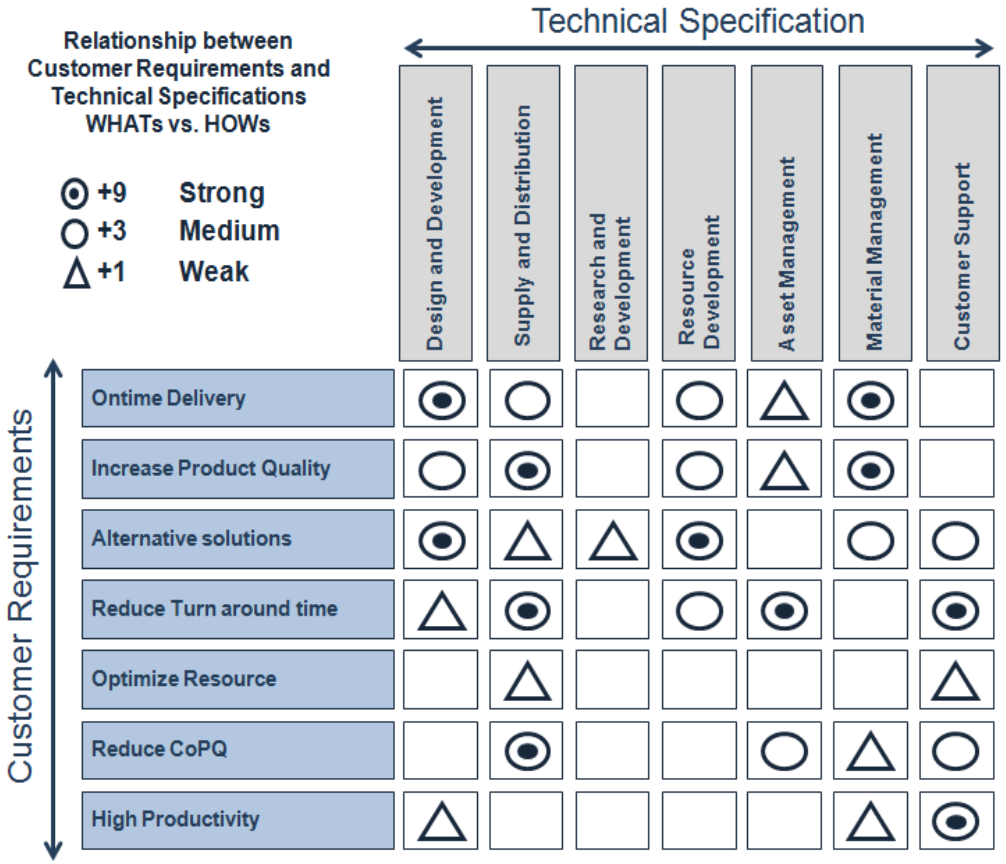
- Potential choices for product features
- Voice of Designers or Engineers
- Each “What” item must be refined to “How’s”

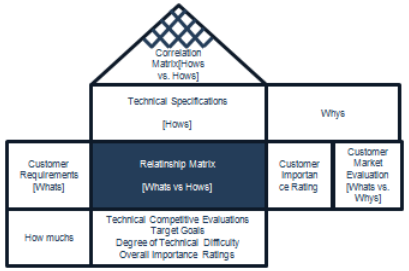




Step 3: Develop Relationship Matrix (What's & How's)

- This is the center portion of the house. Each cell represents how each technical specification relates to each customer requirement.

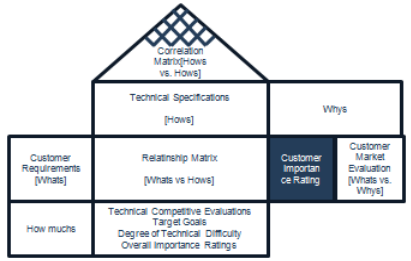




Step 4: Prioritize Customer Requirements

- This is the right portion of the house. Each cell represents customer requirements based on relative importance to customers and perceptions of competitive performance.

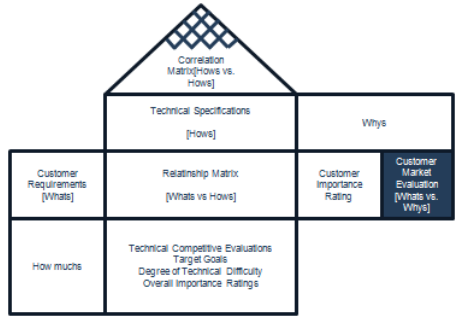
	Design and Development	Supply and Distribution	Research and Development	Resource Development	Asset Management	Material Management	Customer Support	Customer Preferences
Ontime Delivery	◉	○		○	△	◉		3
Increase Product Quality	○	◉		○	△	◉		5
Alternative solutions	◉	△	△	◉		○	○	4
Reduce Turn around time	△	◉		○	◉		◉	2
Optimize Resource		△					△	3
Reduce CoPQ		◉			○	△	○	4
High Productivity	△					△	◉	1



Step 5: Competitive Assessments

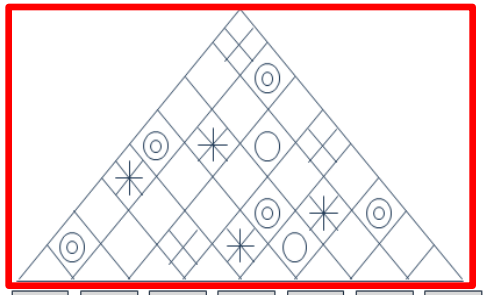
- This is the extreme right portion of the house. Comparison of the organization's product to competitors' products.

	Design and Development	Supply and Distribution	Research and Development	Resource Development	Asset Management	Material Management	Customer Support	Customer Preferences	Our Company	Competitor 1	Competitor 2
Ontime Delivery	◉	○		○	△	◉		3	4	4	3
Increase Product Quality	○	◉		○	△	◉		5	3	3	1
Alternative solutions	◉	△	△	◉		○	○	4	1	2	5
Reduce Turn around time	△	◉		○	◉		◉	2	2	1	2
Optimize Resource		△					△	3	3	4	3
Reduce CoPQ		◉			○	△	○	4	2	5	1
High Productivity	△					△	◉	1	5	3	3



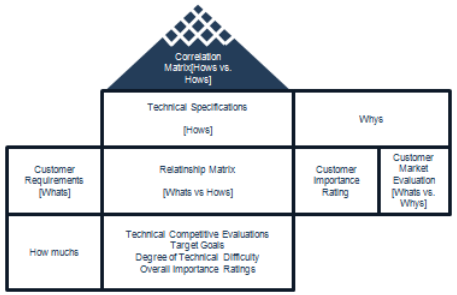
Step 6: Correlation Matrix

- This is the top portion of the house. It identifies the way “how” items either support (positive) or conflict (negative) with one another.



+9 ◎ **Strong Positive**
+3 ○ **Positive**
-3 ╳ **Negative**
-9 ✳ **Strong Negative**

	Design and Development	Supply and Distribution	Research and Development	Resource Development	Asset Management	Material Management	Customer Support	Customer Preferences	Our Company	Competitor 1	Competitor 2
Ontime Delivery	◎	○		○	△	◎		3	4	4	3
Increase Product Quality	○	◎		○	△	◎		5	3	3	1
Alternative solutions	◎	△	△	◎		○	○	4	1	2	5
Reduce Turn around time	△	◎		○	◎		◎	2	2	1	2
Optimize Resource		△					△	3	3	4	3
Reduce CoPQ		◎			○	△	○	4	2	5	1
High Productivity	△					△	◎	1	5	3	3



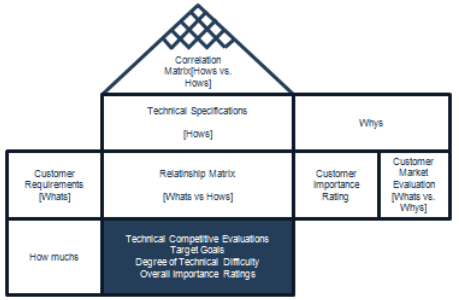
Step 7: Prioritize Design Requirements

- **Overall Importance Ratings**

  Function of relationship ratings and customer prioritization ratings

- **Technical Difficulty Assessment**

  Similar to customer market competitive evaluations but conducted by the technical team

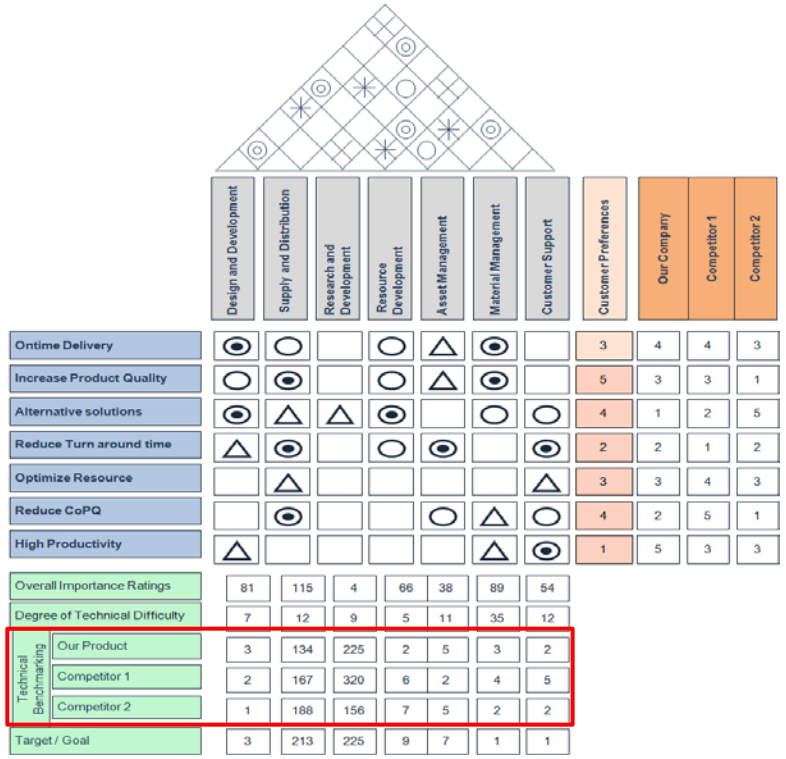


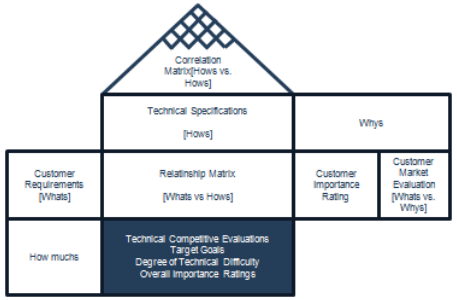
Step 7: Prioritize Design Requirements

- **Technical Specification Competitive Evaluation**

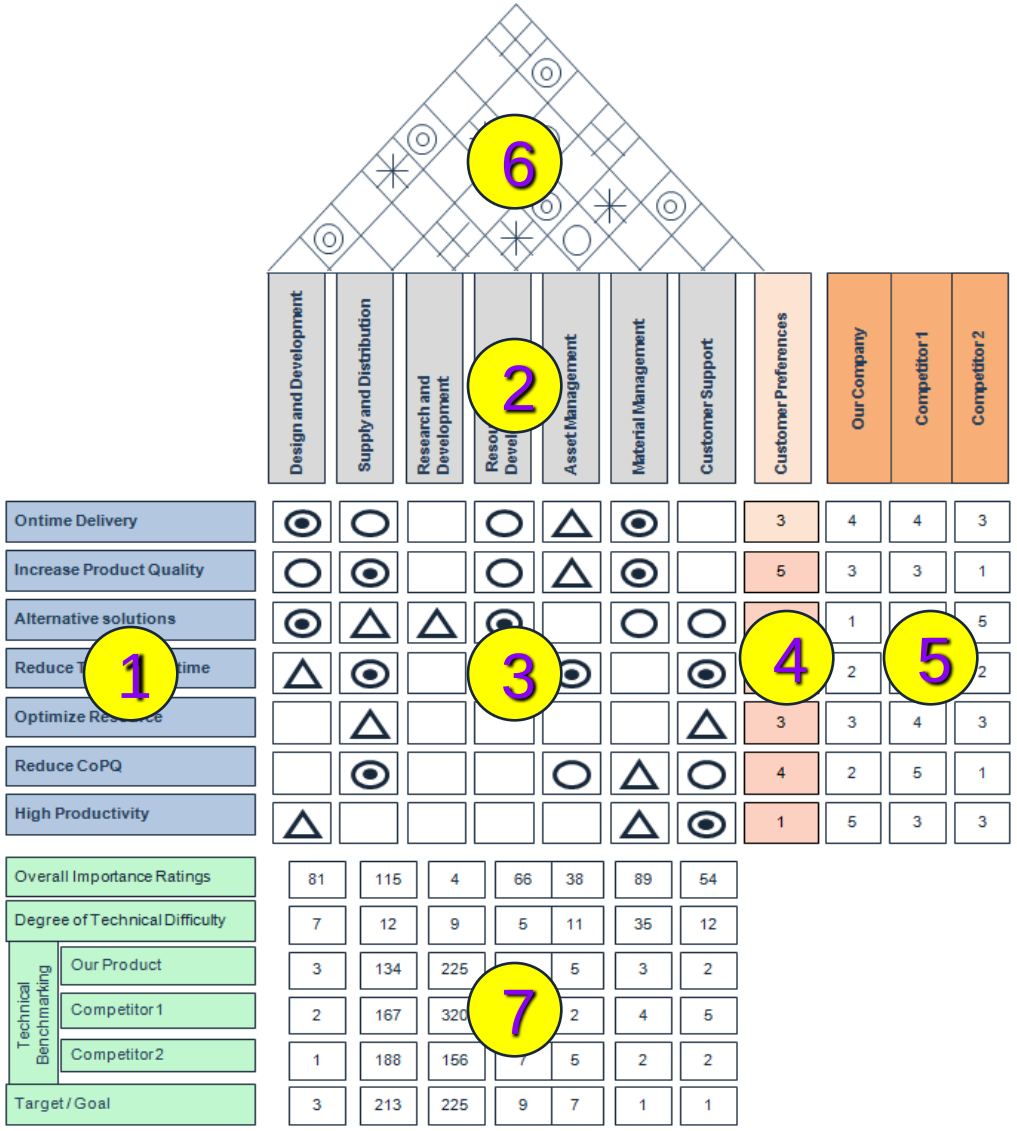
  Helps to establish the feasibility and realization of each “how” item
- **Target Goals**

  How much is good enough to satisfy the customer





House of Quality



Pros of QFD

- Focuses the design of the product or process on satisfying customer’s needs and wants.
- Improves the contact channels between customers, advertising, research and improvement, quality and production departments, which sustains better decision making.
- Reduces the new product development project period and cost.

Cons of QFD

- The relationship matrix can be too obscure with many process inputs and/or many customer constraints.
- It can be very complicated and difficult to implement without experience.
- If throughout the process new ideas, specifications, or requirements are not discovered, you run the risk of losing team members' trust in the process.

QFD Summary

When used properly, the quality function deployment is an extremely valuable approach to product/process design.

There are many benefits of QFD that can only be realized when each step of the process is completed thoroughly:

- Logical way of obtaining information and presenting it
- Smallest product development cycle
- Considerably condensed start-up costs
- Fewer engineering alterations
- Reduced chance of supervision during design process
- Collaborating environment
- Preserving everything in characters.

1.2.4 Cost of Poor Quality

Cost of Poor Quality

- Cost of Poor Quality (COPQ) is the expense incurred due to waste, inefficiencies, and defects.

- The COPQ has been proven to range from 5% to 30% of gross sales for most companies.
- The COPQ can be staggering when considering process inefficiencies, hidden factories, defective products, rework, scrap, etc.
- Understanding COPQ and where to look for it will help uncover process inefficiencies, defects, and hidden factories within your business.

Cost of Poor Quality

- There are 7 common forms of waste that are often referred to as the "**7 deadly muda.**"
- Technically, there are more than 7 forms of waste but if you can remember these you will capture over 90% of your waste.

1. Defects
2. Overproduction
3. Over-Processing
4. Inventory
5. Motion
6. Transportation
7. Waiting

Cost of Poor Quality

- The “7 deadly muda” are very important to understand. They are the best way to identify the COPQ.
- The presence of any muda causes many other forms of inefficiencies and hidden factories to manifest themselves.
    - There are four key categories of costs related to muda:

1. Costs Related to Production
2. Costs Related to Prevention
3. Costs Related to Detection
4. Costs Related to Obligation

COPQ: Costs Related to Production

- Costs related to **production** are the direct costs of the presence of muda. These forms of COPQ are usually understood and easily observable. They are in fact the “7 deadly muda” themselves.

1. Defects
2. Overproduction
3. Over-Processing
4. Inventory
5. Motion
6. Transportation
7. Waiting

COPQ: Costs Related to Prevention

- Costs related to the **prevention** of muda are those associated with trying to reduce or eliminate any of the “7 deadly muda.”
  - Costs for error proofing methods or devices
  - Costs for process improvement and quality programs
  - Costs for training and certifications
  - etc.

- *Any costs* directly associated with the prevention of waste and defects should be included in the COPQ calculation.

COPQ: Costs Related to Detection

- Costs related to the detection of muda are those associated with trying to find or observe any of the “7 deadly muda.”
  - Costs for sampling
  - Costs for quality control check points
  - Costs for inspection costs
  - Costs for cycle counts or inventory accuracy inspections
  - etc.
- Any costs directly associated with the detection of waste and defects should be included in the COPQ calculation.

COPQ: Costs Related to Obligation

- Costs related to **obligation** are those associated with addressing the muda that reaches a customer.
  - Repair costs
  - Warranty costs
  - Replacement costs
  - Customer returns and customer service overhead
  - etc.
- Any costs directly associated with customer obligations should be included in the COPQ calculation.

COPQ: Types of Cost

- There are two types of costs to be considered when determining COPQ
  1. **Hard Costs**
     - Tangible costs that can be traced to the income statement
  2. **Soft Costs**
     - Intangible costs: avoidance, opportunity costs, lost revenue etc.

- Calculating the COPQ
  1. Determine the types of waste that are present in your process
  2. Estimate the frequency of waste that occurs
  3. Estimate the cost per event, item, or time frame
  4. Do the math.

1.2.5 Pareto Charts and Analysis

Pareto Principle

- The **Pareto principle** is commonly known as the “law of the vital few” or “80:20 rule.”
- It means that the majority (approximately 80%) of effects come from a few (approximately 20%) of the causes.
- This principle was first introduced in early 1900s and has been applied as a rule of thumb in various areas.
- Example of applying the Pareto principle:
  - 80% of the defects of a process come from 20% of the causes.
  - 80% of sales come from 20% of customers.

Pareto Principle

- The Pareto principle helps us to focus on the vital few items that have the most significant impact.
- In concept, it also helps us to prioritize potential improvement efforts.
- Since this 80:20 rule was originally based upon the works of Wilfried Fritz Pareto (or Vilfredo Pareto), the Pareto principle and references to it should be capitalized because Pareto refers to a person (proper noun).
  - Mr. Pareto is also credited for many works associated with the 80:20, some more loosely than others:
    - Pareto's Law
    - Pareto efficiency
    - Pareto distribution etc.

Pareto Charts

- A **Pareto chart** is a chart of descending bars with an ascending cumulative line on the top.
  - **Sum or Count**:

The descending bars on a Pareto chart may be set on a scale that represents the total of all bars or relative to the biggest bucket, depending on the software you are using.

- **Percent to Total**: A Pareto chart shows the percentage to the total for individual bars.

- **Cumulative Percentage**: A Pareto chart also shows the cumulative percentage of each additional bar. The data points of all cumulative percentages are connected into an ascending line on the top of all bars.

Pareto Charts

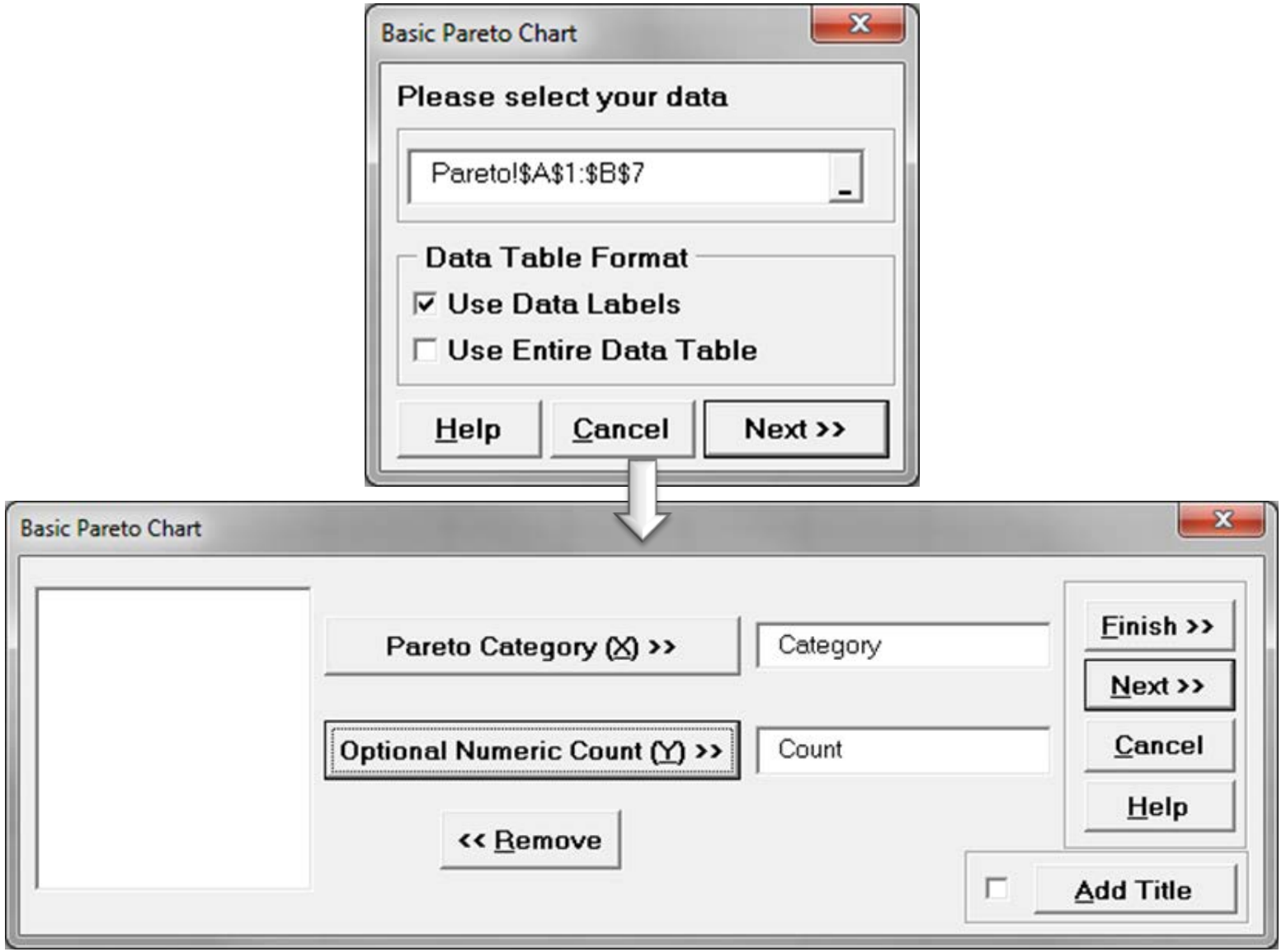
- Case study time!
  - Next we will use SigmaXL to run Pareto charts on exactly the same data set.
  - The following table shows the count of defective products by team.
  - Input the tabled data below into your software program and follow the instructions over the next few pages to run Pareto charts in the appropriate software.

Count	Category
2	team1
12	team2
4	team3
22	team4
2	team5
2	others

Create Pareto Chart in SigmaXL

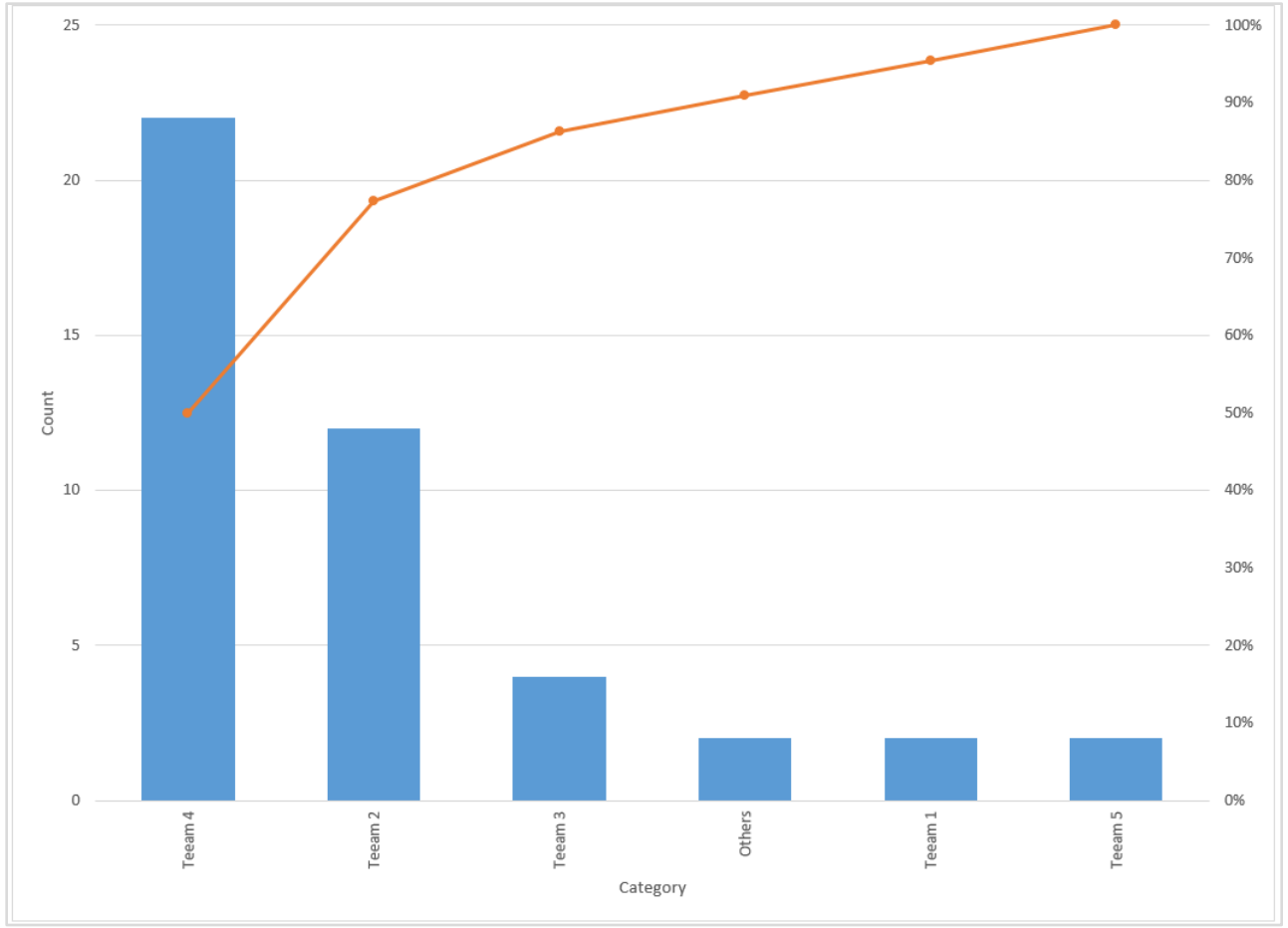
- Steps to generate a Pareto chart using SigmaXL:
  1. Open the Pareto Chart spreadsheet.
  2. Highlight both columns of “Count” and “Category.”
  3. Click SigmaXL → Graphical Tools → Basic Pareto Chart.
  4. A new window named “Pareto Chart” pops up.
  5. Click “Next>>.”
  6. A new window named “Basic Pareto Chart” pops up.
  7. Select “Category” as the “Pareto Category (X)” and “Count” as the “Optional Numeric Count (Y).
  8. Click “Finish.”
  9. The Pareto chart is created in a new tab.

Create Pareto Chart in SigmaXL



Create Pareto Chart in SigmaXL

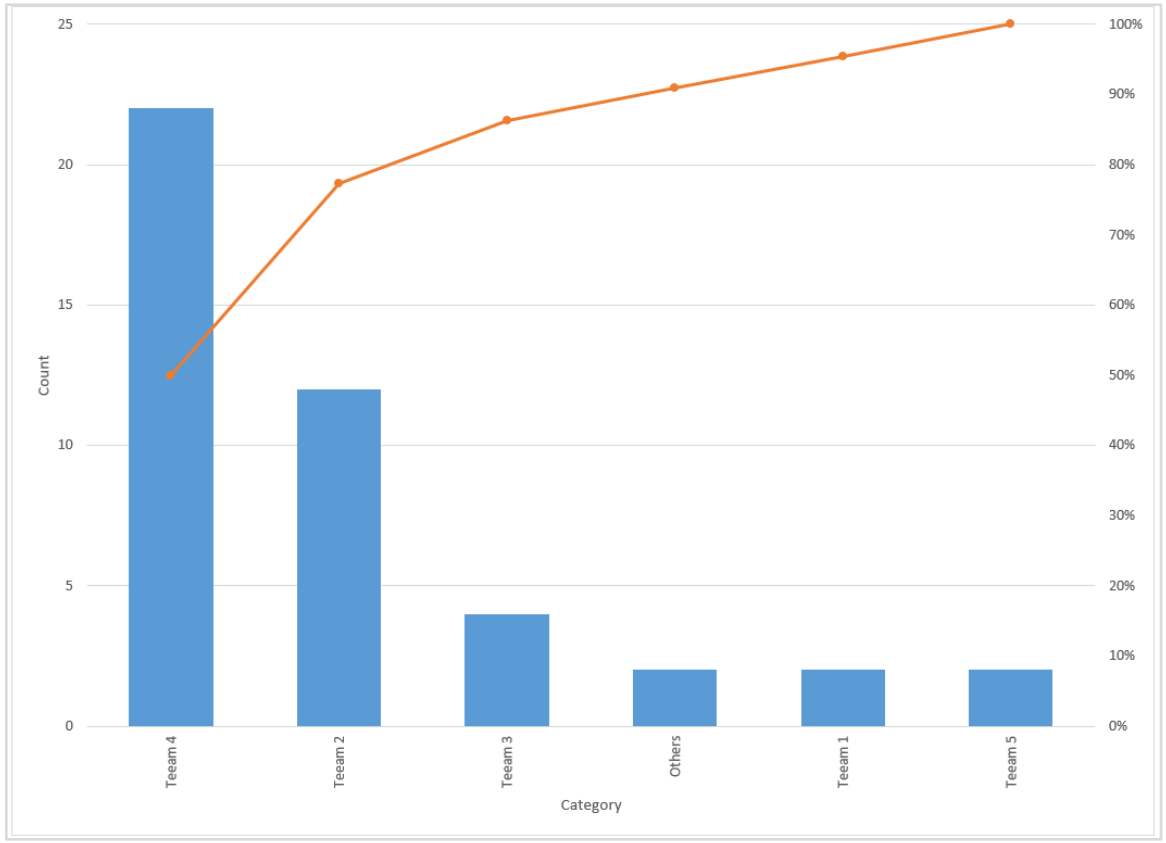
- The Pareto chart at right generated in SigmaXL presents the count of defective products by team.
- The bars are descending on a scale with the peak at 25, which is approximately the size of the largest bar.
- Compared with Minitab, it is a bit more difficult to ascertain the total number of defective items in the Pareto chart created in SigmaXL.



Pareto Analysis

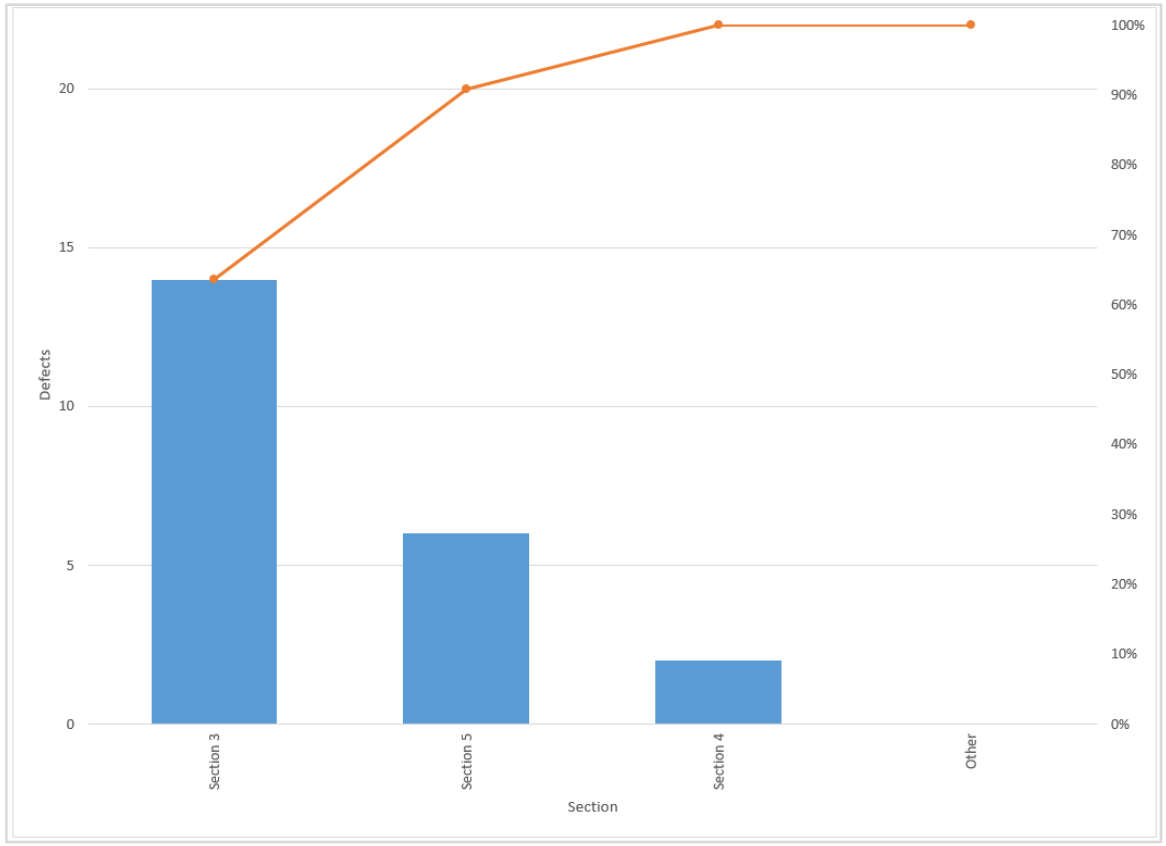
- The **Pareto analysis** is used to identify the root causes by using multiple Pareto charts.
- In Pareto analysis, we drill down into the bigger buckets of defects and identify the root causes of defects that contribute heavily to total defects.
- This "drill down" approach effectively solves a significant portion of the problem.
- Next you will see an example of three-level Pareto analysis.
  - The second-level Pareto is a Pareto chart that is a subset of the tallest bar on the first Pareto.
  - The third-level Pareto is a subset of the tallest bar of the second-level Pareto.

Pareto Analysis: First Level



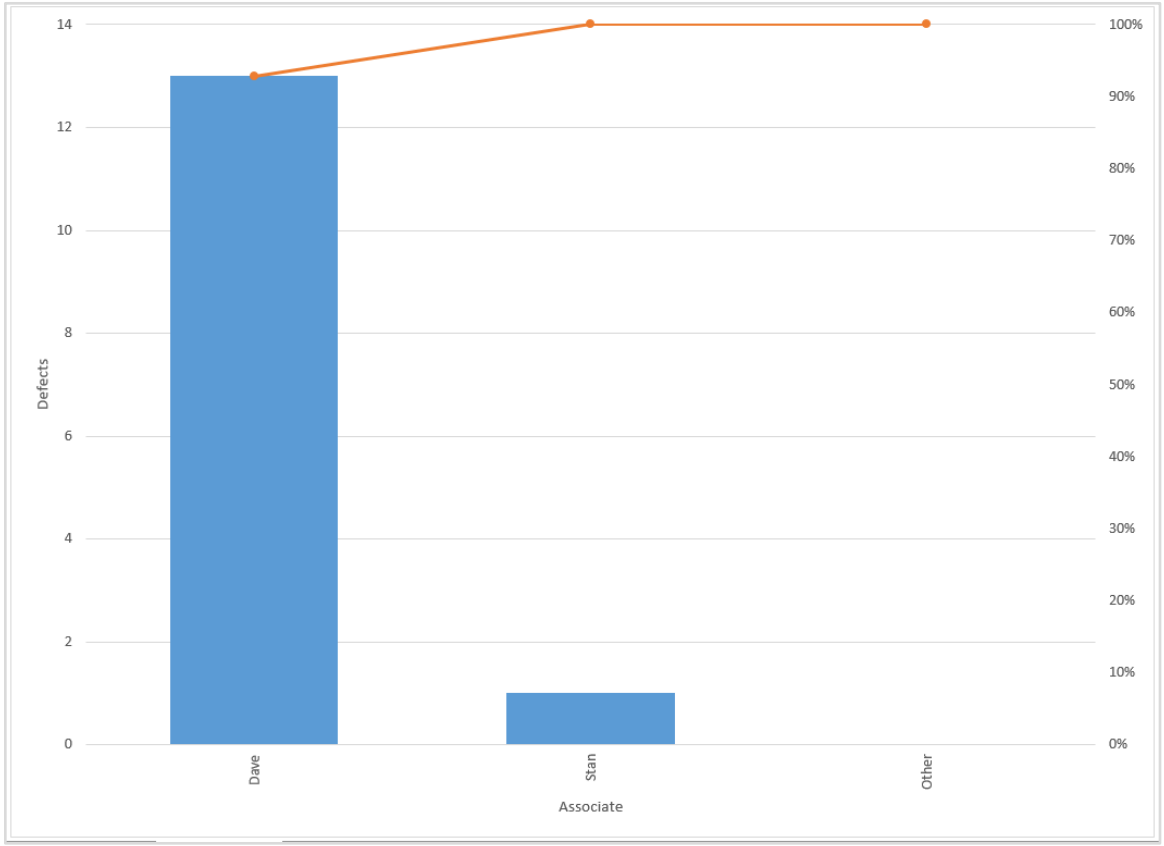
- First-level Pareto
- Shows the count of defective items by team
- Next level will only show the defective items of team 4

Pareto Analysis: Second Level



- Second-level Pareto
- Shows the count of the defective items by section for only team 4
- Next level will only show the defective items of section 3

Pareto Analysis: Third Level

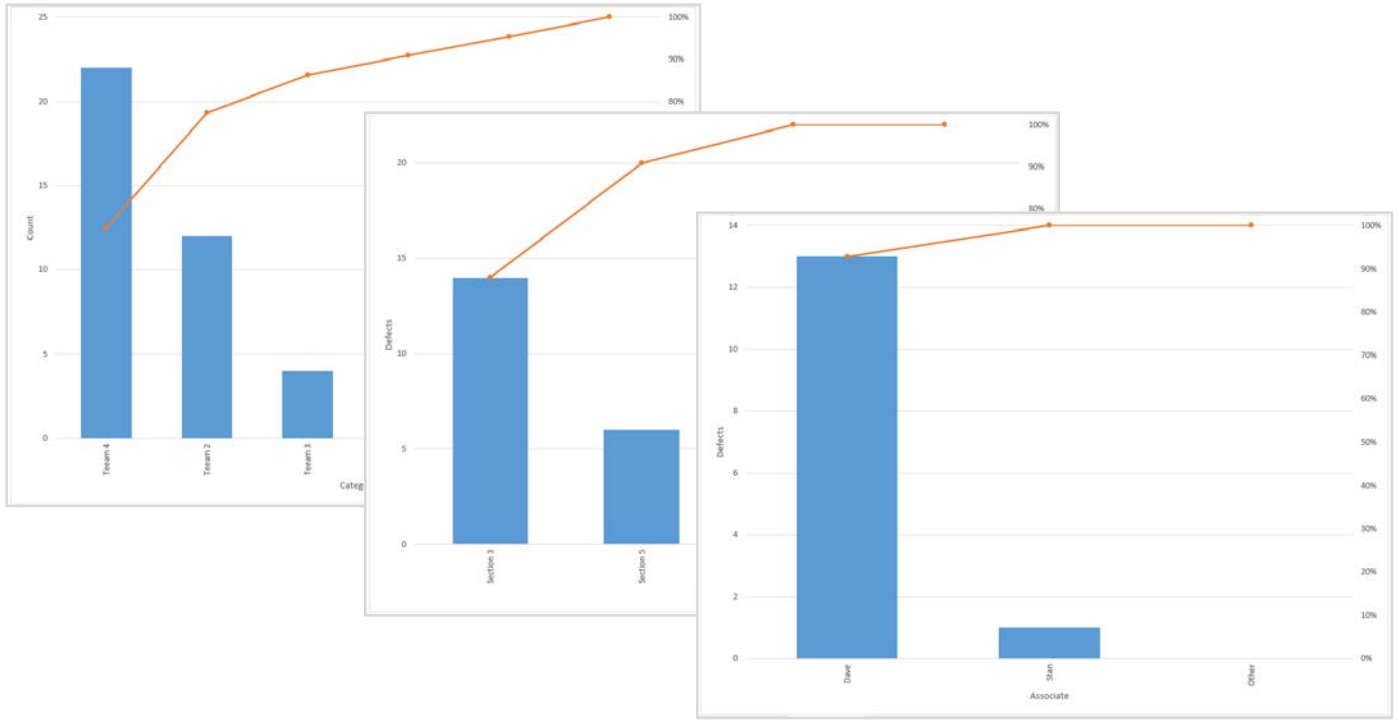


- Third-level Pareto
- Shows the count of defective items by associate for only section 3 of team 4
- Next level will only show the defective items of Dave

Pareto Analysis: Conclusion

- After drilling down three levels we find that most of the defective products are from Dave who is in Section 3 of Team 4.

- Determining what Dave might be doing differently and solving that problem can potentially fix about 30% of the entire defective products (13/44).



1.3 Six Sigma Projects

Black Belt Training: Define Phase

1.1 Six Sigma Overview

- 1.1.1 What is Six Sigma
- 1.1.2 Six Sigma History
- 1.1.3 Six Sigma Approach Y = f(x)
- 1.1.4 Six Sigma Methodology
- 1.1.5 Roles and Responsibilities

1.2 Six Sigma Fundamentals

- 1.2.1 Defining a Process
- 1.2.2 VOC and CTQs
- 1.2.3 QFD
- 1.2.4 Cost of Poor Quality (COPQ)
- 1.2.5 Pareto Analysis (80 : 20 rule)

1.3 Lean Six Sigma Projects

- 1.3.1 Six Sigma Metrics
- 1.3.2 Business Case and Charter
- 1.3.3 Project Team Selection
- 1.3.4 Project Risk Management
- 1.3.5 Project Planning

1.4 Lean Fundamentals

- 1.4.1 Lean and Six Sigma
- 1.4.2 History of Lean
- 1.4.3 The Seven Deadly Muda
- 1.4.4 Five-S (5S)

1.3.1 Six Sigma Metrics

Six Sigma Metrics

- There are many Six Sigma metrics and/or measures of performance used by Six Sigma practitioners.
- In addition to the ones we will cover here, several others (Sigma level, Cp, Cpk, Pp, Ppk, takt time, cycle time, utilization etc.) will be covered in other modules throughout this training.
- The Six Sigma metrics of interest here in the **define phase** are:
  - Defects per Unit (DPU)
  - Defects per Million Opportunities (DPMO)
  - Yield (Y)
  - Rolled Throughput Yield (RTY).

Defects per Unit: DPU

- **DPU** stands for “Defects per Unit“
- DPU is the basis for calculating DPMO and RTY, which we will cover in the next few pages.
- DPU is found by dividing total defects by total units.
  - ***DPU = D/U***
- For example, if you have a process step that produces an average of 65 defects for every 598 units, then your DPU = 65/598 = 0.109.

DPMO: Defects per Million Opportunities

- **DPMO** is one of the few important Six Sigma metrics that you should get comfortable with if you are associated with Six Sigma.

- In order to understand DPMO it is best if you first understand both the nomenclature and the nuances such as the difference between defect and defective.

- **Nomenclature**
  - **Defects** = **D**
  - **Unit** = **U**
  - **Opportunity to have a defect** = **O**

DPMO: Defects per Million Opportunities

- In order to properly discuss DPMO, we must first explore the differences between "defects" and "defective."
  - **Defective**
    - Defective suggests that the value or function of the entire unit or product has been compromised.
    - Defective items will always have at least one defect. Typically, however, it takes multiple defects and/or critical defects to cause an item to be defective.
  - **Defect**
    - A defect is an error, mistake, flaw, fault, or some type of imperfection that reduces the value of a product or unit.
    - A single defect may or may not render the product or unit "defective" depending on the specifications of the customer.
  - **Summary**
    - Defect means that part of a unit is bad.
    - Defective means that the whole unit is bad.

DPMO: Defects per Million Opportunities

- Now let us turn our attention to defining "opportunities" so that we can fully understand Defects per Million Opportunities (DPMO).

  - **Opportunities**
    - Opportunities are the total number of possible defects.
    - Therefore, if a unit has 6 possible defects, then each unit produced is equal to 6 defect opportunities.
    - If we produce 100 units, then there are 600 defect opportunities.

DPMO: Defects per Million Opportunities

- Calculating Defects per Million Opportunities
- **The equation is DPMO = (D/(U × O)) × 1,000,000**
- Example: Let us assume:
  - There are 6 defect opportunities per unit
  - There are an average of 4 defects every 100 units.

  - Opportunities = 6 × 100 = 600
  - Defect rate = 4/600
  - DPMO = 4/600 × 1,000,000 = 6,667

DPMO: Defects per Million Opportunities

- **What is the reason or significance of 1,000,000?**
- Converting defect rates to a per million value becomes necessary when the performance of your process approaches Six Sigma.

- When this happens, the number of defects shrinks to virtually nothing. In fact, if you recall from the “What is Six Sigma” module, sigma is 3.4 defects per million opportunities.

- By using 1,000,000 opportunities as the barometer we have the resolution in the measurement to count defects all the way up to Six Sigma.

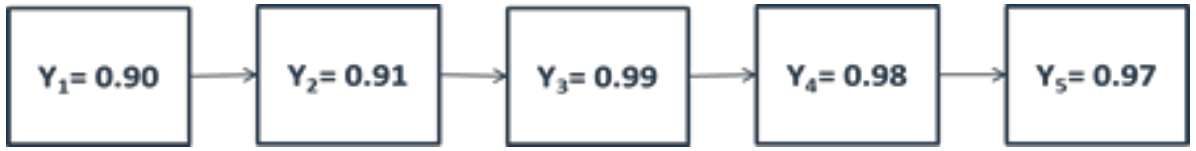
RTY: Rolled Throughput Yield

- **Rolled Throughput Yield (RTY)** is a process performance measure that provides insight into the cumulative effects of an entire process.
- RTY measures the yield for each of several process steps and provides the probability that a unit will come through that process defect-free.
- RTY allows us to expose the "hidden factory" by providing visibility into the yield of each process step.
- This helps us identify the poorest performing process steps and gives us clues into where to look to find the most impactful process improvement opportunities.

RTY: Rolled Throughput Yield

- **Calculating RTY:**
- RTY is found by multiplying the yields of each process step.

- Let us take the 5-step process below and calculate the RTY using the multiplication method mentioned above.



- The calculation is: RTY = 0.90 × 0.91 × 0.99 × 0.98 × 0.97 = 0.77

- Therefore, RTY = 77%.

RTY: Rolled Throughput Yield

- You may have noticed that in order to calculate RTY we must determine the yield for each process step.

- Before we get into calculating yield, there are a few abbreviations that need to be declared.

  - Abbreviations
    - Defects = **D**
    - Unit = **U**
    - Defects per Unit = **DPU**
    - Yield = **Y**
    - e = 2.71828 *(mathematical constant)*

RTY: Rolled Throughput Yield

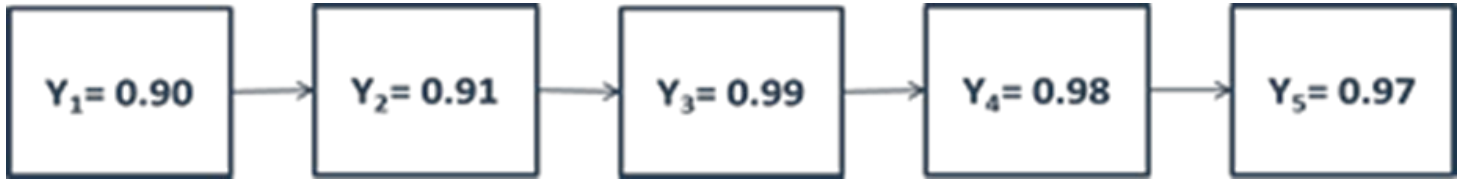
- **Calculating Yield**
- The **yield** of a process step is the success rate of that step or the probability that the process step produces no defects.

- In order to calculate the yield, we need to know the DPU and then we can apply it to the yield equation below.

$$Y = e^{-dpu}$$

- Example
  - Let us assume a process step has a DPU of 0.109 (65/598)
  - Y = 2.718 ^ -0.109 = 0.8967. Rounded, Y = 90%.

RTY: Rolled Throughput Yield



- Below is a table using the above process yield data that we used in the earlier RTY calculation.

- This table allows us to see the DPU and yield of each step as well as the RTY for the whole process.

Process Step	Defects	Units	DPU	Yield	RTY
1	65	598	0.10870	0.89701	0.90
2	48	533	0.09006	0.91389	0.82
3	5	485	0.01031	0.98974	0.81
4	10	480	0.02083	0.97938	0.79
5	14	471	0.02972	0.97072	0.77

RTY: Using an Estimate of Yield

Process Step	Defects	Units	DPU	Yield	RTY
1	65	598	0.10870	0.89701	0.90
2	48	533	0.09006	0.91389	0.82
3	5	485	0.01031	0.98974	0.81
4	10	480	0.02083	0.97938	0.79
5	14	471	0.02972	0.97072	0.77

- Calculating RTY using **yield estimation**
  - It is possible to "estimate" yield by taking the inverse of DPU or simply subtracting DPU from 1.
  - Yield Estimation = 1 − DPU
    - Yield Estimate for process step 1: 1 − 0.10870 = 0.90
    - Yield Estimate for process step 2: 1 − 0.09006 = 0.91
    - Yield Estimate for process step 3: 1 − 0.01031 = 0.99
    - Yield Estimate for process step 4: 1 − 0.02083 = 0.98
    - Yield Estimate for process step 5: 1 − 0.02972 = 0.97
  - RTY using the Yield Estimation Method
    - RTY = 0.90 × 0.91 × 0.99 × 0.98 × 0.97 = 0.77 = 77%

1.3.2 Business Case and Charter

Business Case and Project Charter

- Earlier we stated that DMAIC is a structured and rigorous methodology designed to be repeatedly applied to ***any*** process in order to achieve Six Sigma.

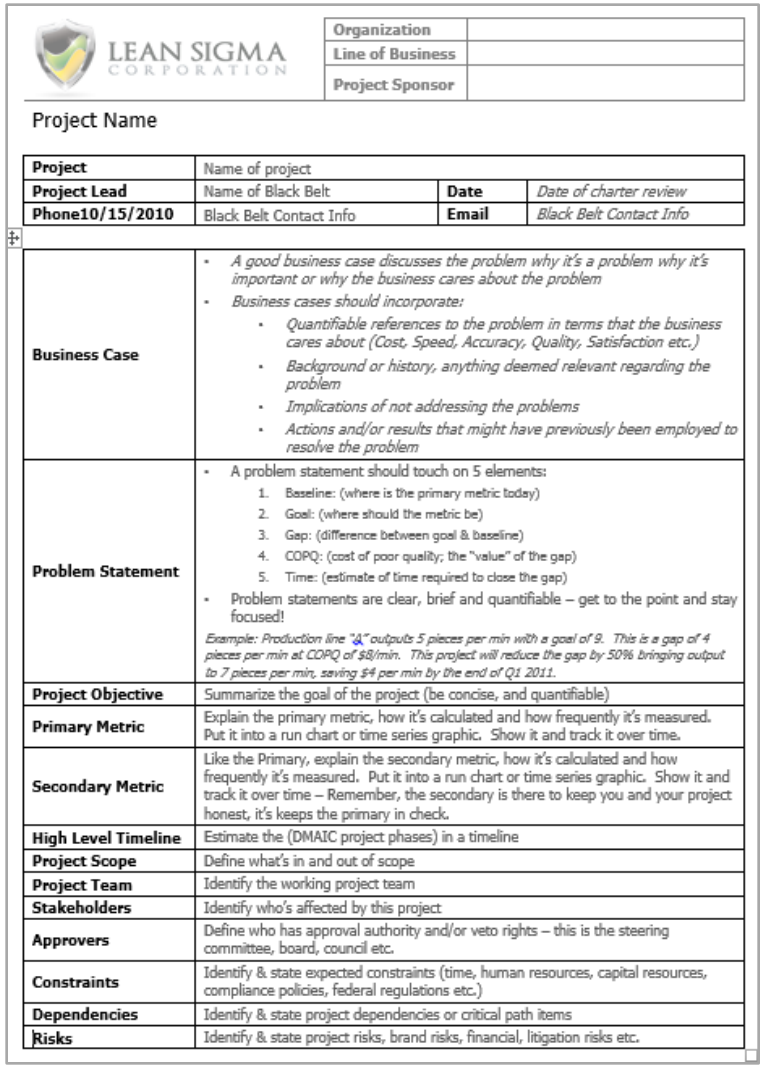
- We also stated that DMAIC was a methodology that refers to 5 phases of a project.
  - Define, Measure, Analyze, Improve, and Control

- Given that the premise of the DMAIC methodology is project-based, we must take the necessary steps to define and initiate a project, hence the need for. . .
  - **Project Charters**

Project Charter

- The purpose of a **project charter** is to provide vital information about a project in a quick and easy-to-comprehend manner.
- Project charters are used to get approval and buy-in for projects and initiatives as well as declaring:
  - The scope of work
  - Project teams
  - Decision authorities
  - Project lead
  - Success measures

Project Charter



LEAN SIGMA CORPORATION

Organization	
Line of Business	
Project Sponsor	

Project Name

Project	Name of project		
Project Lead	Name of Black Belt	**Date**	*Date of charter review*
Phone10/15/2010	Black Belt Contact Info	**Email**	*Black Belt Contact Info*

Business Case	• *A good business case discusses the problem why it's a problem why it's important or why the business cares about the problem*
	• *Business cases should incorporate:*
	  ▪ *Quantifiable references to the problem in terms that the business cares about (Cost, Speed, Accuracy, Quality, Satisfaction etc.)*
	  ▪ *Background or history, anything deemed relevant regarding the problem*
	  ▪ *Implications of not addressing the problems*
	  ▪ *Actions and/or results that might have previously been employed to resolve the problem*

Problem Statement	• A problem statement should touch on 5 elements: 1. Baseline: (where is the primary metric today) 2. Goal: (where should the metric be) 3. Gap: (difference between goal & baseline) 4. COPQ: (cost of poor quality; the "value" of the gap) 5. Time: (estimate of time required to close the gap) • Problem statements are clear, brief and quantifiable – get to the point and stay focused! *Example: Production line "A" outputs 5 pieces per min with a goal of 9. This is a gap of 4 pieces per min at COPQ of $8/min. This project will reduce the gap by 50% bringing output to 7 pieces per min, saving $4 per min by the end of Q1 2011.*
Project Objective	Summarize the goal of the project (be concise, and quantifiable)
Primary Metric	Explain the primary metric, how it's calculated and how frequently it's measured. Put it into a run chart or time series graphic. Show it and track it over time.
Secondary Metric	Like the Primary, explain the secondary metric, how it's calculated and how frequently it's measured. Put it into a run chart or time series graphic. Show it and track it over time – Remember, the secondary is there to keep you and your project honest, it's keeps the primary in check.
High Level Timeline	Estimate the (DMAIC project phases) in a timeline
Project Scope	Define what's in and out of scope
Project Team	Identify the working project team
Stakeholders	Identify who's affected by this project
Approvers	Define who has approval authority and/or veto rights – this is the steering committee, board, council etc.
Constraints	Identify & state expected constraints (time, human resources, capital resources, compliance policies, federal regulations etc.)
Dependencies	Identify & state project dependencies or critical path items
Risks	Identify & state project risks, brand risks, financial, litigation risks etc.

- Key Elements of Project Charters
  - Title
  - Project Lead
  - Business Case
  - Problem Statement
  - Project Objective
  - Primary and Secondary Metrics
  - Project Scope
  - Project Timeline
  - Project Constraints
  - Project Team
  - Stakeholders
  - Approvers
  - Constraints
  - Dependencies
  - Risks

Project Charter: Key Elements

- Title
  - Projects should have a name, title, or some reference that identifies them.
  - Branding can be an important ingredient in the success of a project so be sure your project has a reference name or title.

- Leader
  - Any projects needs a declared leader or someone who is responsible for project's execution and success.
  - You may hear references to RACI throughout in your Six Sigma journey.
  - **RACI** stands for **R**esponsible, **A**ccountable, **C**onsulted, **I**nformed and identifies the people that play those roles.
  - Every project must have declared leaders indicating who is responsible and who is accountable.

Project Charter: Key Elements

- Business Case
  - A **business case** is the quantifiable reason why the project is important.
  - Business cases help shed light on problems. They explain why a business should care.
  - Business cases must be quantified and stated succinctly.
  - COPQ is a key method of quantification for any business case.

Project Charter: Key Elements

- Problem Statement and Objective
  - A properly written problem statement has an objective statement woven into it.
  - There should be no question as to the current state or the goal.

  - A gap should be declared, the gap being the difference between the present state and the goal state.
  - The project objective should be to close the gap or reduce the gap by some reasonable amount.

  - Valuation or COPQ is the monetary value assigned to the gap.
  - Lastly, a well-written problem statement refers to a timeline expected to be met.

Project Charter: Problem Statement Examples

- Currently, process defect rates are 17% with a goal of 2%. This represents a gap of 15%, costing the business $7.4 million dollars. The goal of this project is to reduce this gap by 50% before Nov 2010 putting process defect rates at 9.5% and saving $3.7MM.

- Process cycle time has averaged 64 minutes since Q1 2009. However, production requirements put the cycle time goals at 48 min. This 16-min gap is estimated to cost the business $296,000. The goal of this project is reduce cycle time by 16 min. by Q4 2010 and capture all $296,000 cost savings.

Project Charter: Key Elements

- Metrics
  - A measure of success is an absolute for any project.
  - Metrics give clarity to the purpose of the work.
  - Metrics establish how the initiative will be judged.
  - Metrics establish a baseline or “starting point.”
  - For Six Sigma projects…**metrics are mandatory!**

Project Charter: Key Elements

- Primary Metric
  - The **primary metric** is a generic term for a Six Sigma project's most important measure of success. The primary metric is defined by the Black Belt, GB, MBB, or Champion.
  - A primary metric is an absolute MUST for any project and it should not be taken lightly. Here are a few characteristics of good primary metrics.
  - Primary metrics should be:
    - tied to the problem statement
    - measureable
    - expressed with an equation
    - aligned to business objectives
    - tracked at the proper frequency (hourly, daily, weekly, monthly etc.)
    - expressed pictorially over time with a run chart, time series, or control chart
    - validated with an MSA.

Project Charter: Key Elements

- The primary metric is the reason for your work.
- It is the success indicator.
- It is your beacon.
- The primary metric is of utmost importance and should be improved, **but** not at the expense of your secondary metric.

Project Charter: Key Elements

- Secondary Metric
  - The **secondary metric** is the thing you do not want sacrificed on behalf of a primary improvement.
  - A secondary metric is one that makes sure problems are not just "changing forms" or "moving around."
  - The secondary metric keeps us honest and ensures we are not sacrificing too much for our primary metric.
  - If your primary metric is a cost or speed metric, then your **secondary metric** should probably be some quality measure.
    - Example: If you were accountable for saving energy in an office building and your primary metric was energy consumption then you *could* shut off all the lights and the HVAC system and save tons of energy. . .except that your secondary metrics are probably comfort and functionality of the work environment.

Project Charter: Key Elements

- Elements of a Good Project Charters *(continued)*
  - Scope Statement – defined by high-level process map
  - Stakeholders Identified – who is affected by the project
  - Approval Authorities Identified – who makes the final call
  - Review Committees Defined – who is on the review team
  - Risks and Dependencies Highlighted – identify risks and critical path items
  - Project Team Declared – declare team members
  - Project Timeline Estimated – set high-level timeline expectations.

1.3.3 Project Team Selection

Project Team Selection

- Six Sigma project team selection is the cornerstone of a successful Six Sigma project.
- Teams and Team Success
  - A **team** is a group of people who share complementary skills and experience.
  - A team will be dedicated to consistent objectives.
  - Winning teams share similar and coordinated goals.
  - Teams often execute common methods or approaches.
  - Team members hold each other accountable for achieving shared goals.

Project Team Selection

- What makes a team successful?
  - Shared goals
  - Commitment
  - Leadership
  - Respect
  - Effective communication
  - Autonomy
  - Diversity (capabilities, knowledge, skills, experience etc.)
  - Adequate resources.

Project Team Selection

- Keys to Team Success
  - Agreed focus on the goal or the problem at hand
    - Focus on problems that have meaning to the business
    - Focus on solvable problems within the scope of influence; a successful team does not seek unattainable solutions.
  - Team Selection
    - Selected teammates have proper skills and knowledge
    - Adequately engaged management
    - Appropriate support and guidance from their direct leader
  - Successful teams use reliable methods
    - Follow the prescribed DMAIC methodology
    - Manage data, information, and statistical evidence
  - Successful teams always have exceeds players
    - Winning teams typically
      - Have unusually high standards.
      - Have greater expectations of themselves and each other.
      - Do not settle for average or even above average results.

Project Team Selection

- Principles of Team Selection:
  - Select team members based on
    - Skills required to achieve the objective
    - Experience (Subject Matter Expertise)
    - Availability and willingness to participate
    - Team size (usually 4–8 members)
      - Don’t go at it alone!
      - Don’t get too many cooks in the kitchen!
    - Members’ ability to navigate
      - The process
      - The company
      - The political landscape
  - Be sure to consider the inputs of others
    - Heed advice
    - Seek guidance

Project Team Development

- All teams experience the following four stages of development. It is helpful to understand these phases so that you can anticipate what your team is going to experience.

- The four stages of team development process:
  - Forming
  - Storming
  - Norming
  - Performing

- Teammates seek something different at each stage:
  - In the forming stage they seek inclusion
  - In the storming stage they seek direction and guidance
  - In the norming stage they seek agreement
  - In the performing stage they seek results.

Project Team Development

- Patterns of a team in the **Forming** stage:
  - Roles and responsibilities are unclear
  - Process and procedures are ignored
  - Scope and parameter setting is loosely attempted
  - Discussions are vague and frustrating
  - There is a high dependence on leadership for guidance
- Patterns of a team in the **Storming** stage:
  - Attempts to skip the research and jump to solutions
  - Impatience for some team members regarding lack of progress
  - Arguments about decisions and actions of the team
  - Team members establish their position
  - Subgroups or small teams form
  - Power struggles exist and resistance is present

Project Team Development

- Patterns of a team in the **Norming** stage:
  - Agreement and consensus start to form
  - Roles and responsibilities are accepted
  - Team members' engagement increases
  - Social relationships begin to form
  - The leader becomes more enabling and shares authority

- Patterns of a team in the **Performing** stage:
  - Team is directionally aware and agrees on objectives
  - Team is autonomous
  - Disagreements are resolved within the team
  - Team forms above average expectations of performance

Project Team Development

- Well-structured and energized project teams are the essential components of any successful Six Sigma project.

- To have better chances of executing the project successfully, you will need to understand and effectively manage the team development process.

1.3.4 Project Risk Management

Risk

Risk is defined as a future event that *can* impact the task/project if it occurs.

What is Project Risk Management?

- The main purpose of **risk management** is to foresee potential risks that may inhibit the project deliverables from being delivered on time, within budget, and at the appropriate level of quality, and then to mitigate these risks by creating, implementing, and monitoring *contingency plans*.

- Risk management is concerned with identifying, assessing, and monitoring project risks before they develop into issues and impact the project.

- **Risk analysis** helps to identify and manage potential problems that could impact key business initiatives or project goals.

Three Basic Parameters of Risk Analysis

- Risk Assessment:

  The process of identifying and evaluating risks, whether in absolute or relative terms.

- Risk Management:

  Project risk management is the effort of responding to risks throughout the life of a project and in the interest of meeting project goals and objectives.

- Risk Communication:

  Communication plays a vital role in the risk analysis process because it leads to a good understanding of risk assessment and management decisions.

Why is Risk Analysis Necessary?

What can happen if you omit the risk analysis?

- Vulnerabilities cannot be detected
- Mitigation plans are introduced without proper justification
- Customer dissatisfaction
- Not meeting project goals
- Remake the whole system
- Huge cost and time loss

Project Risk Analysis Steps

The project risk analysis process consists of the following steps that evolve through the life cycle of a project.

- **Risk Identification**:
  - Identify risks and risk categories, group risks, and define ownership.
- **Risk Assessment**:
  - Evaluate and estimate the possible impacts and interactions of risks.
- **Response Planning**:
  - Define mitigation and reaction plans.
- **Mitigation Actions**:
  - Implement action plans and integrate them into the project.
- **Tracking and Reporting**:
  - Provide visibility to all risks.
- **Closing**:
  - Close the identified risk.

Risk Identification

The first action of risk management is the identification of individual events that the project may encounter during its lifecycle.

The identification step comprises:

- Identify the risks
- Categorize the risks
- Match the identified risks to categories
- Define ownership for managing the risks.

Risk Identification

- Source of Risk:
- Identification of risk sources provides a basis for systematically examining changing situations over time to uncover circumstances that impact the ability of the project to meet its objectives.

Risk Identification

Source of Risk	Description
Human Resources	The risks originated from human resources (e.g., availability, skill etc.)
Physical Resources	The risks originated from physical resources (e.g., hardware or software, availability of the required number at the right time etc.)
Technology	The risks originated from technology (e.g., development environment, new or complex technologies, performance requirements, tools etc.)
Suppliers	The risks are associated with a supplier (e.g., delays in supplies, capability of suppliers etc.)
Customer	The risks derived from the customer (e.g., unclear requirements, requirement volatility, change in project scope, delays in response etc.)
Security	The risks are associated with information security, security of personnel, security of assets, and security of intellectual property
Legal	The risks are associated with legal issues that may impact the project
Project management	The risks are associated with project management processes, organizational maturity, and ability

Risk Identification

Risk Parameters:

Parameters for evaluating, categorizing, and prioritizing risks include the following:

- Risk likelihood (i.e., probability of risk occurrence)
- Risk consequence (i.e., impact and severity of risk occurrence)
- Thresholds to trigger management activities.

Risk Assessment

The **risk assessment** consists of evaluating the range of possible impacts should the risk occur.

Follow these steps when assessing risks:

1) Define the various impacts of each risk
2) Rate each impact based on a logical severity level
3) Sort and evaluate risks by severity level
4) Determine if any controls already exist
5) Define potential mitigation actions.

Risk Mitigation Planning

The risk owners are responsible for planning and implementing mitigation actions with support from the project team.

- All team members, inclusive of partners and suppliers, may be requested to identify and develop mitigation measures for identified risks.
- The project core team members are responsible for identifying an appropriate action owner for each identified risk.
- After mitigation actions are defined, the project core team will review the actions.
- The risk owner must track all mitigation actions and expected completion dates.
- The risk owner and the project core team members must hold all action owners accountable for the risk mitigation planning.

Risk Mitigation Action Implementation

- The **action implementation** is the responsibility of the risk owner.
- The action owners are responsible for the execution of the tasks or activities necessary to complete the mitigation action and eliminate or minimize the risk.
- The risk owner or the project manager will monitor completion dates of the mitigation action implementation.

Risk Occurrence and Contingency Plans

- Whenever any risk occurs, the project team should implement **contingency plans** to ensure that project deliverables can be met.

- The details of each occurrence should be recorded in the risk register or other tracking tool.

- The **risk register** or **risk management plan** *(see next slide)* will be maintained by the project manager and reviewed on a regular basis.

Risk Tracking and Reporting

- **Risk tracking and reporting** provides critical visibility to all risks.
- Risk owners must report on the status of their mitigation actions.
- Depending on the risk severity, project managers need to report the risk status of each category of risk to senior management.

This template is available in the "Lean Sigma Corporation Templates.xls" file

LEAN SIGMA CORPORATION

Risk Management Plan

Company	Project/Program Name	Project Lead	Project Sponsor/Champion	Last Updated

Risk ID	Risk Category	Risk Description	Risk Impact	Impact Rating	Mitigation Action	Responsible	Status

Risk Closure

- The risk owners are responsible for recommending the risk closure to the project manager.
- A risk is *closed* only when the item is not considered a risk to the project anymore.
- When a risk is closed, the project manager needs to update the risk status in the lessons learned document.

Risk Analysis Features

The risk analysis should be:

- Systematic
- Comprehensive
- Data driven
- Adherent to evidence
- Logically sound
- Practically acceptable
- Open to critique
- Easy to understand.

Project Risk Analysis Advantages

- Helps strategic and business planning
- Meets customer requirements
- Reduces schedule slips and cost overruns
- Promotes an effective usage of resources
- Promotes continuous improvement
- Helps to achieve project goals
- Minimizes surprises from customers and stakeholders
- Allows a quick grasp of new opportunities
- Enhances communication
- Reassures stakeholders that the project stays on track.

1.3.5 Project Planning

What is Project Management?

- **Project management** is the process of defining, planning, organizing, managing, leading, securing, and optimizing the resources to achieve a set of planned goals and objectives.

- *Project Management is the application of knowledge, skills, tools, and techniques to project activities in order to meet project requirements.

*This definition was taken from the Glossary of the Project Management Institute, A Guide to the Project Management Body of Knowledge, (PMBOK®Guide) –Fifth Edition, Project Management Institute, Inc., 2008

What is a Project Plan?

- A **project plan** is a crucial step in project management for achieving a project's goals.
- A project plan is a formal approved document used to guide and execute project tasks.
- It provides an overall framework for managing project tasks, schedules, and costs.
- A project plan is a coordinating tool and communication device that helps teams, contractors, customers, and organizations define the crucial aspects of a project or program.

Project Planning Stages

1. **Determine project scope and objectives**: Explore opportunities, identify and prioritize needs, consider project solutions.
2. **Plan the project**: Identify input and resources requirements such as human resources, materials, software, hardware, and budgets.
3. **Prepare the project proposal:** Based on stakeholder feedback, plan the necessary resources, timeline, budget etc.
4. **Implement the project:** Implement the project by engaging responsible resources and parties. Ensure execution and compliance of the defined plans.
5. **Evaluate the project:** Regularly review progress and results. Measure the project's effectiveness against *quantifiable* requirements.

Planning and Scheduling Objectives

- To optimize the use of resources (both human and other resources).
- To increase productivity
- To achieve desired schedules and deliverables
- To establish an approach to minimize long-term maintenance costs
- To minimize the chaos and productivity losses resulting from planned production schedules, priority changes, and non-availability of resources.
- To assess current needs and future challenges.

Project Planning Activities

- Statement of work (SOW)
- Work breakdown structure (WBS)
- Resource estimation plan
- Project schedule
- Budget or financial plan
- Communication plan
- Risk management plan

Project Planning Activities

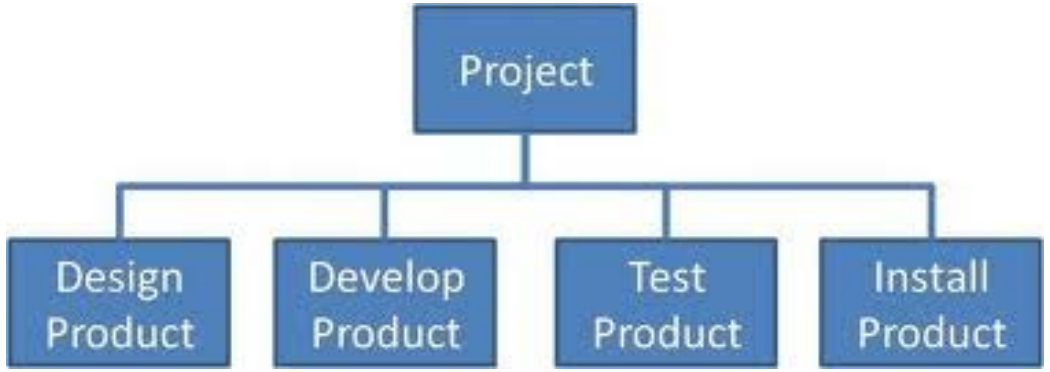
Statement of Work (SOW)

- Define the scope of the project.
- Establish customer expectations.
- Identify technical requirements for the project.

Project Planning Activities

Work Breakdown Structure (WBS)

- Identify all the tasks that need to be done in order to complete the project.
- Structure the tasks into small logical components and subcomponents.
- Define each task in detail so that each person responsible understands what is expected of them.
- Summarize and report project progress and results.



Project Planning Activities

Resource Estimation Plan

- Estimate resources
  - Human resources
  - Hardware and software
- Plan resources

Index	Resources Required	No. Required	By date	Responsibility
1	RAN access, SSE file transfer, office	1	03 Jan 11	John
2	Access to ITS	1	03 Jan 11	Bob
3	Software Engineers	4	12 Jan 11	Dave
4	Onwing-EgtMargin-V1, RCCMacro-V2,	1	03 Jan 11	Michael

Project Planning Activities

Project Schedule

- Assign time estimates to each activity in the WBS.
- Create each task start and end dates.
- Represent schedules as Gantt charts or network diagrams (PERT/CPM) charts.
- Identify critical dependencies between tasks.

Project Planning Activities

Project Schedule – Gantt Chart:

- The advantage of a Gantt chart is its ability to display the status of each task/activity at a glance.
- Because it is a graphic representation, it is easy to demonstrate the schedule to all the stakeholders.

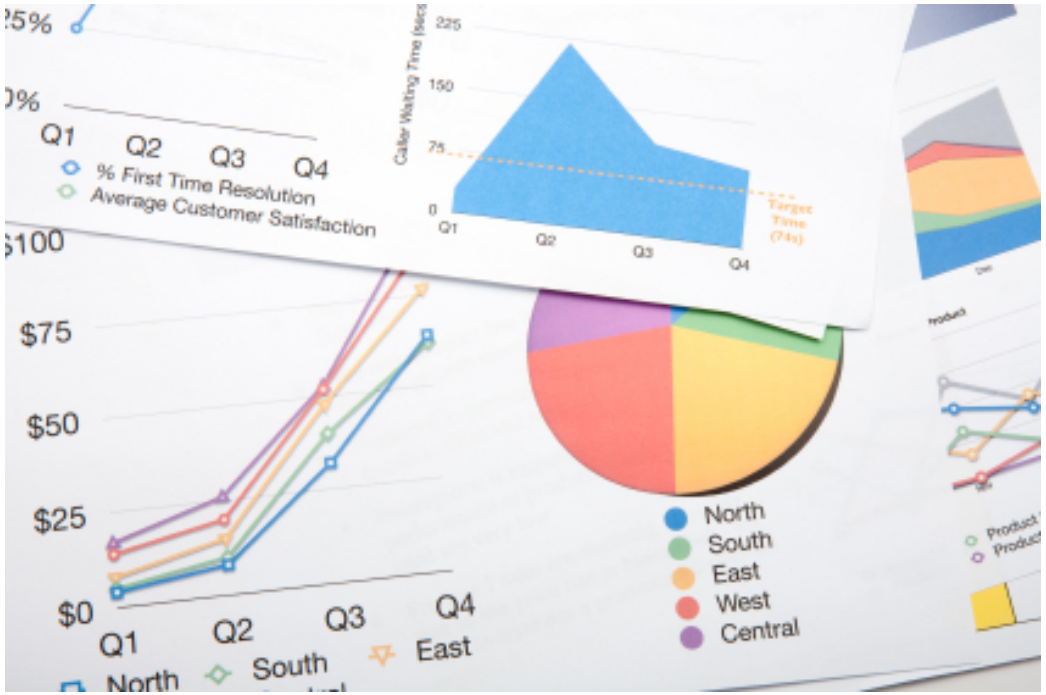
- Project Name : Six Sigma Digest
- Project Description : Launching Training Products
- Project Length :
- Start Date : 13-Jun-11 — End Date 31-Jul-11 / Number of Weeks 7
- Working Days : Monday - Sunday
- Today's Marker : Yes
- Holiday's Marker : No

Level	Task	PIC	Start Date	Finish Date	WD	DC	DR
1	**Online Training Products**	**Marketing Dept**					
1.1	Interactive Overviews	Richard	1-Jun-11	19-Jun-11	19	12	7
1.2	Interactive Quizzes	Michael	1-Jun-11	19-Jun-11	19	12	7
1.3	Editing Review	Kovach	1-Jun-11	19-Jun-11	19	12	7
1.4	Edit all Versions	Richard	20-Jun-11	26-Jun-11	7	-7	14
2	**Learning Management System**	**Marketing Dept**					
2.1	Set-up Course Structures in eFront	Michael	1-Jun-11	26-Jun-11	26	12	14
2.2	Set-up Profiles & Course Key Logistics	Michael	1-Jun-11	26-Jun-11	26	12	14
2.3	Publish Courses to LMS (Scorm format)	Richard	27-Jun-11	3-Jul-11	7	-14	21
2.4	Add Exam Questions to LMS	Michael	27-Jun-11	3-Jul-11	7	-14	21
2.5	Set up 5 Exams in LMS	Michael	4-Jul-11	10-Jul-11	7	-21	28
3	**Shopping Cart Functionality**						
3.1	Test product-option & download modules	Michael	27-Jun-11	3-Jul-11	7	-14	21
3.2	Test Authroize.net ARB & eCheck modules	Michael	27-Jun-11	3-Jul-11	7	-14	21
3.3	Install Modules	Michael	4-Jul-11	10-Jul-11	7	-21	28
3.4	Purchase & Load License keys (SXL & MTB)	Michael	11-Jul-11	11-Jul-11	1	-28	29
4	**Beta Testing Online Courses**						
4.1	Test purchase processes	Team	12-Jul-11	24-Jul-11	13	-29	42
4.2	Test login & activation processes	Team	12-Jul-11	24-Jul-11	13	-29	42
4.3	Test course navigation & function	Team	12-Jul-11	24-Jul-11	13	-29	42
4.4	Test exams	Team	12-Jul-11	24-Jul-11	13	-29	42
4.5	Final edits & adjustments	Team	25-Jul-11	30-Jul-11	6	-42	48
5	**Live Launch of Online Training**						
6	**Courseware Products**						
6.1	Package and Publish zip files to downloads	Michael	11-Jul-11	17-Jul-11	7	-28	35
7	**Live Launch of Courseware Products**						

Project Planning Activities

Budget or Financial Plan

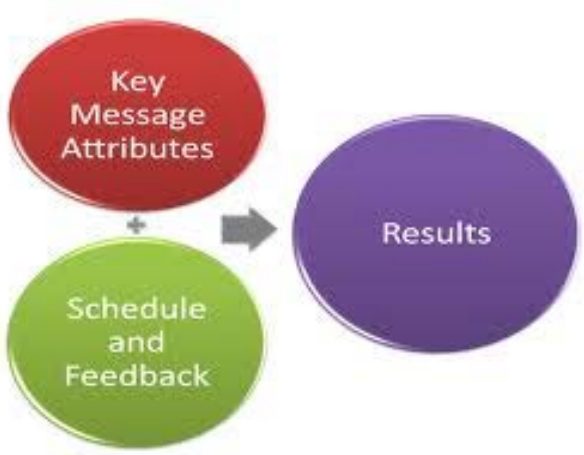
- Planned expenses
- Planned revenues
- Budget forecast



Project Planning Activities

Communication Plan

- Establish communication procedures among management, team members, and relevant stakeholders.
- Determine the communication schedule.
- Define the acceptable modes of communication.



Project Planning Activities

Risk Management Plan

- Identify the sources of project risks and estimate the effects of those risks.
- Risks might arise from new technology, availability of resources, lack of inputs from customers, business risks etc.
- Assess the impact of risk to the customers/stakeholders.
- Calculate the probability of risk occurrence based on previous similar projects or industry benchmarks
- Initiate mitigation and contingency plans
- Review risks on a periodic basis

Project Planning Tools Advantages

- Project planning tools are very useful to organize and communicate project plans, status, and projections.
- They help link tasks and sub-tasks or other work elements to get a whole view of what needs to be accomplished.
- They allow a more objective comparison of alternative solutions and provide consistent coverage of responsibilities.
- They allow for effective scope control and change management.
- They facilitate effective communication with all project participants and stakeholders.
- They help define management reviews.
- They act as an effective monitoring mechanism for the project.
- They establish project baselines for progress reviews and control points.

Project Planning Tools Disadvantages

- Project planning tools can sometimes take too much time to maintain.
- Data updating and accuracy can be cumbersome.
- Too much documentation can cause version control to be challenging.
- Ineffective use of tools, especially risk management tools or project plans, can bring unwarranted project risks because bad decisions can be made on inaccurate information.
- Understanding of tools and usage of the tools may require training, hence additional costs and time.

1.4 Lean Fundamentals

Black Belt Training: Define Phase

1.1 Six Sigma Overview

- 1.1.1 What is Six Sigma
- 1.1.2 Six Sigma History
- 1.1.3 Six Sigma Approach Y = f(x)
- 1.1.4 Six Sigma Methodology
- 1.1.5 Roles and Responsibilities

1.2 Six Sigma Fundamentals

- 1.2.1 Defining a Process
- 1.2.2 VOC and CTQs
- 1.2.3 QFD
- 1.2.4 Cost of Poor Quality (COPQ)
- 1.2.5 Pareto Analysis (80 : 20 rule)

1.3 Lean Six Sigma Projects

- 1.3.1 Six Sigma Metrics
- 1.3.2 Business Case and Charter
- 1.3.3 Project Team Selection
- 1.3.4 Project Risk Management
- 1.3.5 Project Planning

1.4 Lean Fundamentals

- 1.4.1 Lean and Six Sigma
- 1.4.2 History of Lean
- 1.4.3 The Seven Deadly Muda
- 1.4.4 Five-S (5S)

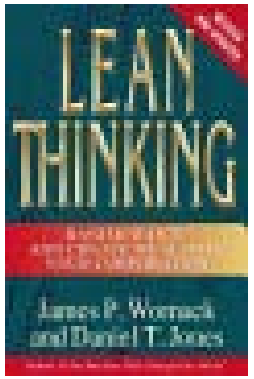
1.4.1 Lean and Six Sigma

What is Lean?

- A **lean enterprise** intends to eliminate waste and allow only value to be pulled through its system.
- **Lean manufacturing** is characterized by:
  - Identifying and driving value
  - Establishing flow and pull systems
  - Creating production availability and flexibility
  - Zero waste
- Waste Elimination
  - Waste identification and elimination is critical to any successful lean enterprise.
  - Elimination of waste enables flow, drives value, cuts cost, and provides flexible and available production.

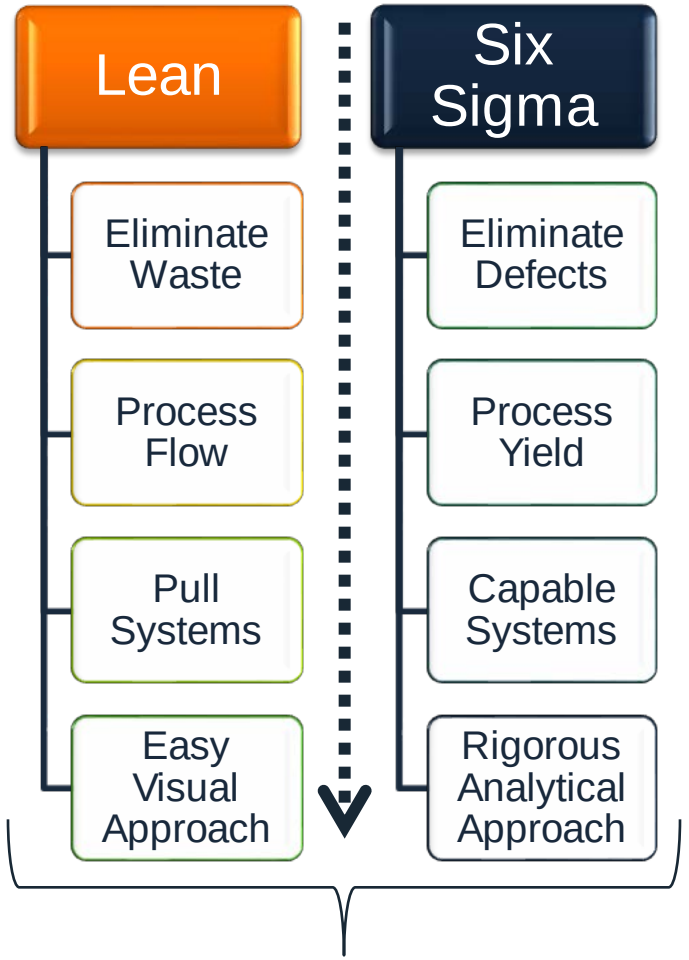
The 5 Lean Principles

- The following 5 principles of lean are taken from the book *Lean Thinking* (1996) by James P. Womack and Daniel T. Jones.



1. Specify value desired by customers.
2. Identify the value stream.
3. Make the product flow continuous.
4. Introduce pull systems where continuous flow is possible.
5. Manage toward perfection so that the number of steps and the amount of time and information needed to serve the customer continually falls.

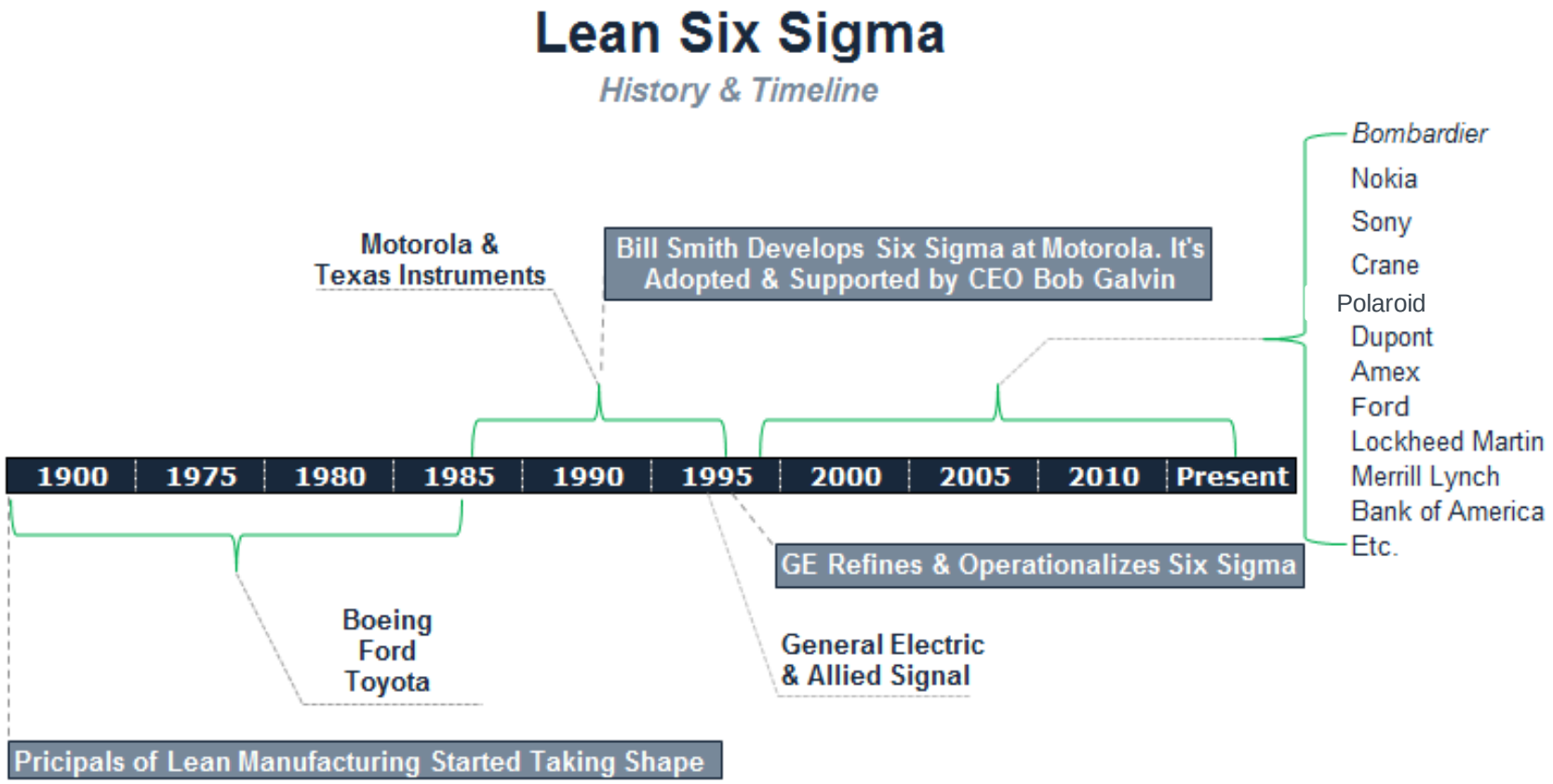
Lean & Six Sigma



- Lean and Six Sigma both have the objectives of producing high value (**quality**) at lower costs (**efficiency**).
- They approach these objectives in somewhat different manners but in the end, both Lean and Six Sigma drive out waste, reduce defects, improve processes, and stabilize the production environment.
- Lean and Six Sigma are a perfect combination of tools for improving quality and efficiency.

1.4.2 History of Lean

History of Lean



History of Lean

- Lean thinking originated, as far as is known, the 1400s.
- Henry Ford established the first mass production system in 1913 by combining standard parts, conveyors, and work flow.
- Decades later, Kiichiro Toyoda and Taiichi Ohno at Toyota improved and implemented various new concepts and tools (e.g., value stream, takt time, kanban etc.) based on Ford's effort.
- Toyota developed what is known today as the Toyota Production System (TPS) based on lean principles.

History of Lean

- Starting in the mid 1990s, Lean became extensively recognized and implemented when more and more Fortune 100 companies began to adopt Lean and Six Sigma.

- The term “Lean manufacturing” was introduced by James Womack in the 1990s.

- Lean and Six Sigma share similar objectives, work hand in hand, and have benefited from one another in the past 30 years.

1.4.3 Seven Deadly Muda

The 7 Deadly Muda

- The Japanese word for waste is “muda.”

- There are 7 commonly recognized forms of waste, often referred to as the “7 deadly muda.”

1. Defects
2. Overproduction
3. Over-Processing
4. Inventory
5. Motion
6. Transportation
7. Waiting

The 7 Deadly Muda: Defects

- Defects or defectives are an obvious waste for any working environment or production system.

- Defects require rework during production and/or after the product is returned from an unhappy customer.
- Some defects are difficult to solve and they create “workarounds” and hidden factories.
- Eliminating defects is a sure way to improve product quality, customer satisfaction, and production costs.

The 7 Deadly Muda: Overproduction

- Overproduction is wasteful because your system expends energy and resources to produce more materials than the customer or next function requires.

- Overproduction is one of the most detrimental of the seven deadly muda because it leads to many others:
- Inventory
- Transportation
- Waiting etc.

The 7 Deadly Muda: Over-processing

Over-processing occurs any time more work is done than is required by the next process step, operation, or consumer.

- Over-processing also includes being over capacity (scheduling more workers than required or having more machines than necessary).

- Another form of over processing can be buying tools or software that are overkill (more precise, complex, or expensive than required).

The 7 Deadly Muda: Inventory

- **Inventory** is an often overlooked waste. Look at the picture above and imagine all the time, materials, and logistics that went into establishing such an abundance of inventory.

- If this were your personal business, and inventory velocity was not matched with production, how upset would you be?

The 7 Deadly Muda: Motion

- Motion is another form of waste often occurring as a result of poor setup, configuration, or operating procedures.
- Wasted motion can be experienced by machines or humans.
- Wasted motion is very common with workers who are unaware of the impact of small unnecessary movements in repetitive tasks.
- Wasted motion is exaggerated by repetition or recurring tasks.

The 7 Deadly Muda: Transportation

- **Transportation** is considered wasteful because it does *nothing* to add value or transform the product.
- Imagine for a moment driving to and from work twice before getting out of your car to go into work. . .

- That is waste in the form of transportation.
- The less driving you have to do, the better.
- In a similar way, the less transportation a product has to endure, the better. There would be fewer opportunities for delay, destruction, loss, damage etc.

The 7 Deadly Muda: Waiting

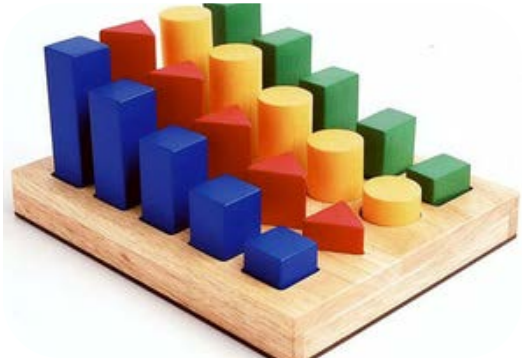
- **Waiting** is an obvious form of waste and is typically a symptom of an upstream problem.
- Waiting is usually caused by inefficiency, bottlenecks, or poorly-designed work flows within the value stream.
- Waiting can also be caused by inefficient administration.
- Reduction in waiting time will require thoughtful applications of lean and process improvement.

1.4.4 Five-S (5S)

What is 5S?

- **5S** is systematic method to organize, order, clean, and standardize a workplace…and to keep it that way!
  - 5S is a methodology of organizing and improving the work environment.
- 5S is summarized in five Japanese words all starting with the letter S:
  - **Seiri** (sorting)
  - **Seiton** (straightening)
  - **Seiso** (shining)
  - **Seiketsu** (standardizing)
  - **Shisuke** (sustaining)
- 5S was originally developed in Japan and is widely used to optimize the workplace to increase productivity and efficiency.

Five-S (5S)

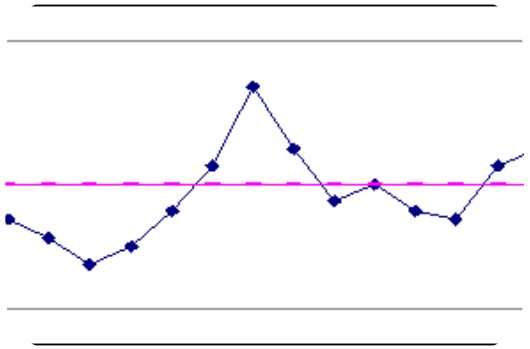


Sort

Set in Order

Shine

Standardize



Sustain

Goals of 5S

- Reduced waste
- Reduced cost
- Establish a work environment that is:
  - self-explaining
  - self-ordering
  - self-regulating
  - self-improving.
  - Where there is/are **no more:**
    - Wandering and/or searching
    - Waiting or delaying
    - Secret hiding spots for tools
    - Obstacles or detours
    - Extra pieces, parts, materials etc.
    - Injuries
    - Waste.

Benefits of 5S Systems

- Reduced changeovers
- Reduced defects
- Reduced waste
- Reduced delays
- Reduced injuries
- Reduced breakdowns
- Reduced complaints
- Reduced red ink
- Higher quality
- Lower costs
- Safer work environment
- Greater associate and equipment capacity.

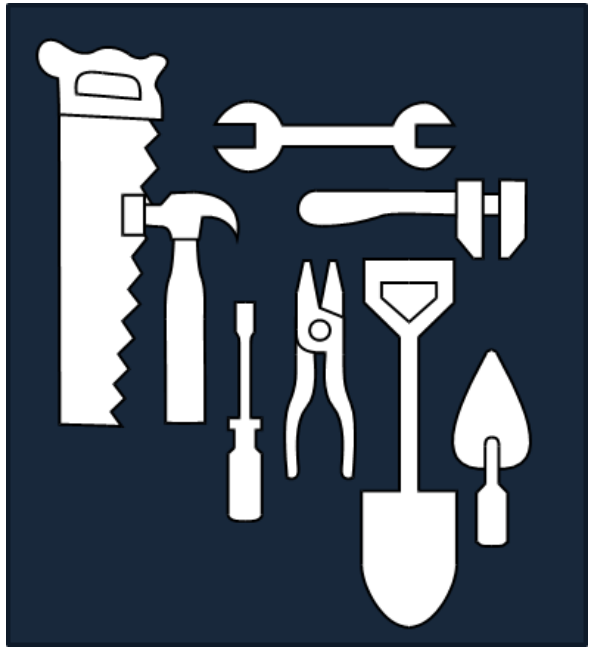
Reported Results of 5S Systems

- Cut in floor space: 60%
- Cut in flow distance: 80%
- Cut in accidents: 70%
- Cut in rack storage: 68%
- Cut in number of forklifts: 45%
- Cut in machine changeover time: 62%
- Cut in annual physical inventory time: 50%
- Cut in classroom training requirements: 55%
- Cut in nonconformance in assembly: 96%
- Increase in test yields: 50%
- Late deliveries: 0%
- Increase in throughput: 15%

Sorting (Seiri)

- Go through all the tools, parts, equipment, supply, and material in the workplace.
- Categorize them into two major groups: needed and unneeded.
- Eliminate the unneeded items from the workplace. Dispose of or recycle those items.
- Keep the needed items and sort them in the order of priority. **When in doubt…throw it out!**

Straightening (Seiton)



- **Straightening** in 5S is also called **setting in order.**
- Label each needed item.
- Store items at their best locations so that the workers can find them easily whenever they needed any item.
- Reduce the motion and time required to locate and obtain any item whenever it is needed.
- Promote an efficient work flow path.
- Use visual aids like the tool board image on this page.

Shining (Seiso)

- **Shining** in 5S is also called **sweeping**.
- Clean the workplace thoroughly.
- Maintain the tidiness of the workplace.
- Make sure every item is located at the specific location where it should be.
- Create the ownership in the team to keep the work area clean and organized.

Standardizing (Seiketsu)

- **Standardize** the workstation and the layout of tools, equipment and parts.

- Create identical workstations with a consistent way of storing the items at their specific locations so that workers can be moved around to any workstation any time and perform the same task.

Sustaining (Shisuke)

- **Sustaining** in 5S is also called **self-discipline**.
- Create the culture in the team to follow the first four S's consistently.
- Avoid falling back to the old ways of cluttered and unorganized work environment.
- Keep the momentum of optimizing the workplace.
- Promote innovations of workplace improvement.
- Sustain the first fours S's using:
  - 5S Maps
  - 5S Schedules
  - 5S Job cycle charts
  - Integration of regular work duties
  - 5S Blitz schedules
  - Daily workplace scans.

Simplified Summary of 5S

1. **Sort** – “when in doubt, move it out.”

2. **Set in Order** – Organize all necessary tools, parts, and components of production. Use visual ordering techniques wherever possible.

3. **Shine** – Clean machines and/or work areas. Set regular cleaning schedules and responsibilities.

4. **Standardize** – Solidify previous three steps, make 5S a regular part of the work environment and everyday life.

5. **Sustain** – Audit, manage, and comply with established 5S guidelines for your business or facility.

Five-S (5S)

- A few words about 5S and the Lean Enterprise
  - As a method, 5S generates immediate improvements.
  - 5S is one of many effective lean methods that create observable results.
  - It is tempting to implement 5S alone without considering the entire value stream.
  - However, it is advisable to consider a well-planned lean manufacturing approach to the entire production system.

2.0 Measure Phase

Black Belt Training: Measure Phase

2.1 Process Definition
- 2.1.1 Cause and Effect Diagrams
- 2.1.2 Cause and Effects Matrix
- 2.1.3 Process Mapping
- 2.1.4 FMEA: Failure Modes and Effects Analysis
- 2.1.5 Theory of Constraints

2.2 Six Sigma Statistics
- 2.2.1 Basic Statistics
- 2.2.2 Descriptive Statistics
- 2.2.3 Distributions and Normality
- 2.2.4 Graphical Analysis

2.3 Measurement System Analysis
- 2.3.1 Precision and Accuracy
- 2.3.2 Bias, Linearity, and Stability
- 2.3.3 Gage R&R
- 2.3.4 Variable and Attribute MSA

2.4 Process Capability
- 2.4.1 Capability Analysis
- 2.4.2 Concept of Stability
- 2.4.3 Attribute and Discrete Capability
- 2.4.4 Monitoring Techniques

2.1 Process Definition

Black Belt Training: Measure Phase

2.1 Process Definition
- 2.1.1 Cause and Effect Diagrams
- 2.1.2 Cause and Effects Matrix
- 2.1.3 Process Mapping
- 2.1.4 FMEA: Failure Modes and Effects Analysis
- 2.1.5 Theory of Constraints

2.2 Six Sigma Statistics
- 2.2.1 Basic Statistics
- 2.2.2 Descriptive Statistics
- 2.2.3 Distributions and Normality
- 2.2.4 Graphical Analysis

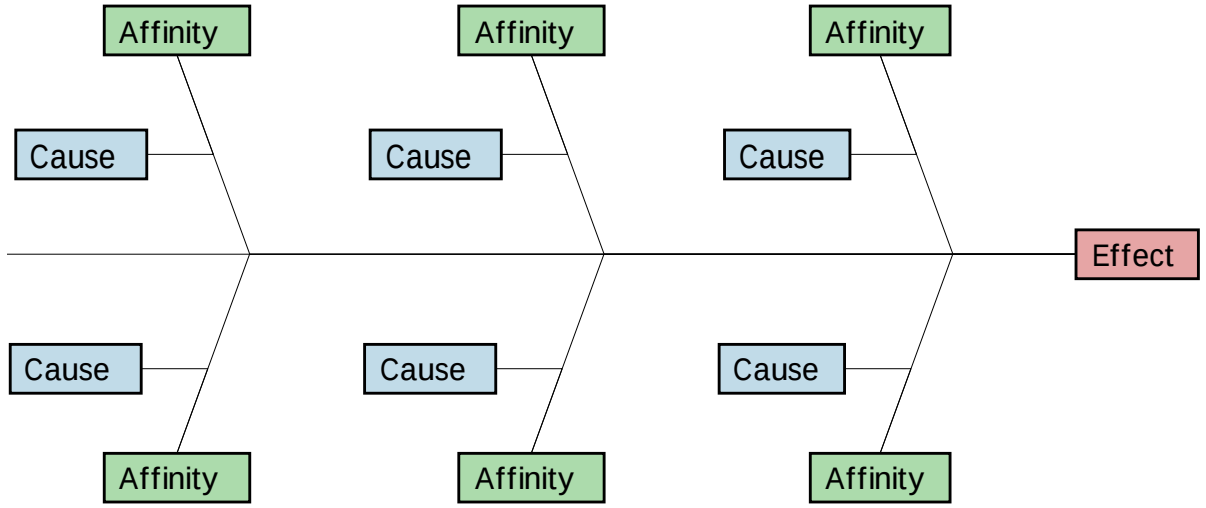
2.3 Measurement System Analysis
- 2.3.1 Precision and Accuracy
- 2.3.2 Bias, Linearity, and Stability
- 2.3.3 Gage R&R
- 2.3.4 Variable and Attribute MSA

2.4 Process Capability
- 2.4.1 Capability Analysis
- 2.4.2 Concept of Stability
- 2.4.3 Attribute and Discrete Capability
- 2.4.4 Monitoring Techniques

2.1.1 Cause and Effect Diagram

What is a Cause and Effect Diagram?

- A **cause and effect diagram** is also called a *Fishbone Diagram* or *Ishikawa Diagram*. It was created by Kaoru Ishikawa and is used to identify, organize, and display the potential causes of a specific effect or event in a graphical way similar to a fishbone.
- It illustrates the relationship between one specified event (output) and its categorized potential causes (inputs) in a visual and systematic way.



Major Categories of Potential Causes

- P4ME
  - **People**: People who are involved in the process
  - **Methods**: How the process is completed (e.g., procedures, policies, regulations, laws)
  - **Machines**: Equipment or tools needed to perform the process
  - **Materials**: Raw materials or information needed to do the job
  - **Measurements**: Data collected from the process for inspection or evaluation
  - **Environment**: Surroundings of the process (e.g., location, time, culture).

How to Plot a Cause and Effect Diagram

- Step 1: Identify and define the effect/event being analyzed.
  - Clearly state the operational definition of the effect/event of interest.
  - The event can be the positive outcome desired or negative problem targeted to solve.
  - Enter the effect/event in the end box of the Fishbone diagram and draw a spine pointed to it.

How to Plot a Cause and Effect Diagram

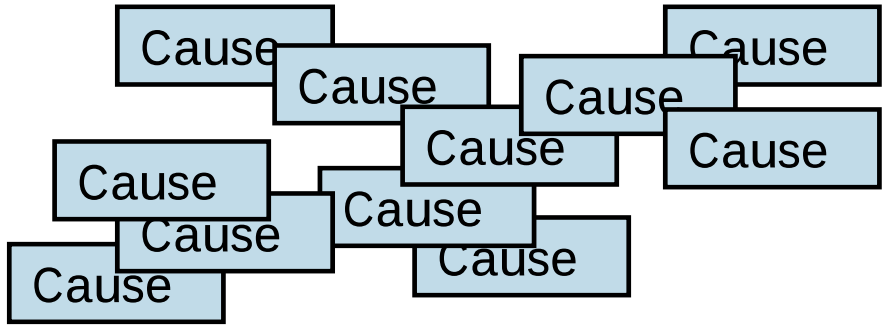
- Step 1

How to Plot a Cause and Effect Diagram

- Step 2: Brainstorm the potential causes or factors of the effect/event occurring.
  - Identify any factors with a potential impact on the effect/event and include them in this step.
  - Put all the identified potential causes aside for use later.

How to Plot a Cause and Effect Diagram

- Step 2

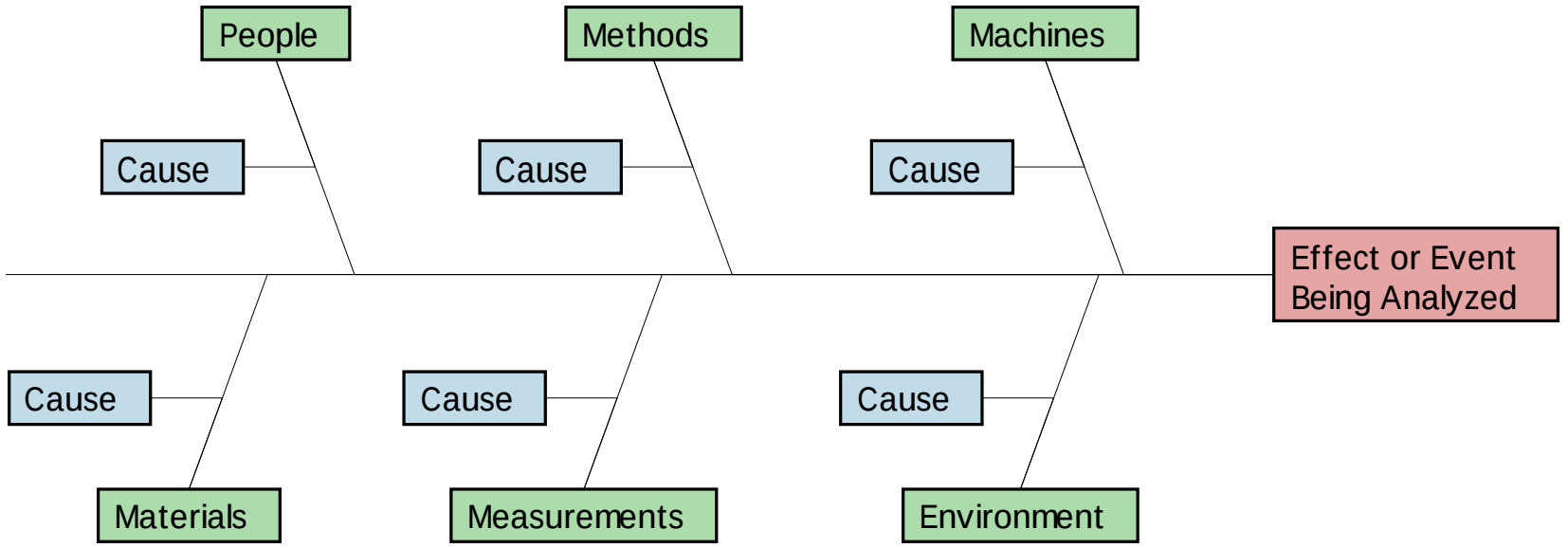


How to Plot a Cause and Effect Diagram

- Step 3: Identify the main categories of causes and group the potential causes accordingly.
  - Besides P4ME (i.e., people, methods, machines, materials, measurements, and environment), you can group potential causes into other customized categories.
  - Below each major category, you can define sub-categories and then classify them to help you visualize the potential causes.
  - Enter each cause category in a box and connect the box to the spine. Link each potential cause to its corresponding cause category.

How to Plot a Cause and Effect Diagram

- Step 3



How to Plot a Cause and Effect Diagram

- Step 4: Analyze the cause and effect diagram.
  - A cause and effect diagram includes all the possible factors of the effect/event being analyzed.
  - Use a Pareto chart to filter causes the project team needs to focus on.
  - Identify causes with high impact that the team can take action upon.
  - Determine how to measure causes and effects quantitatively. Prepare for further statistical analysis.

Benefits to Using Cause and Effect Diagram

- Helps to quickly identify and sort the potential causes of an effect.
- Provides a systematic way to brainstorm potential causes effectively and efficiently.
- Identifies areas requiring data collection for further quantitative analysis.
- Locates “low-hanging fruit.”

Limitation of Cause and Effect Diagrams

- A cause and effect diagram only provides qualitative analysis of correlation between each cause and the effect.
- One cause and effect diagram can only focus on *one* effect or event at a time.
- Further statistical analysis is required to quantify the relationship between various factors and the effect and identify the root causes.

Cause and Effect Diagram Example

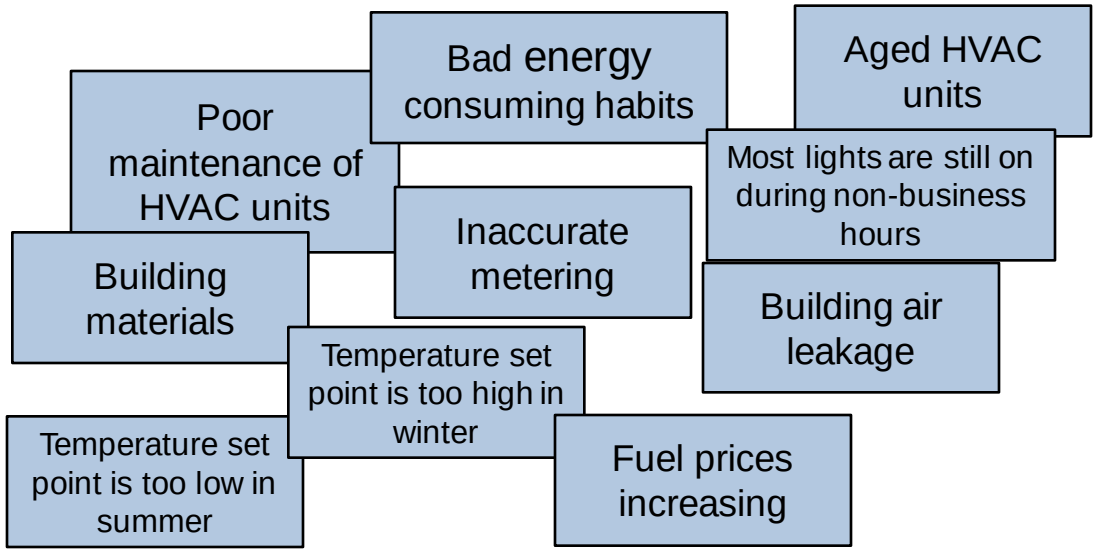
- *Case study*:
  - A real estate company is interested to find the root causes of high energy costs of its properties.
  - The cause and effect diagram is used to identify, organize, and analyze the potential root causes.

Cause and Effect Diagram Example

- Step 1: Identify and define the effect/event being analyzed: high energy costs of buildings.

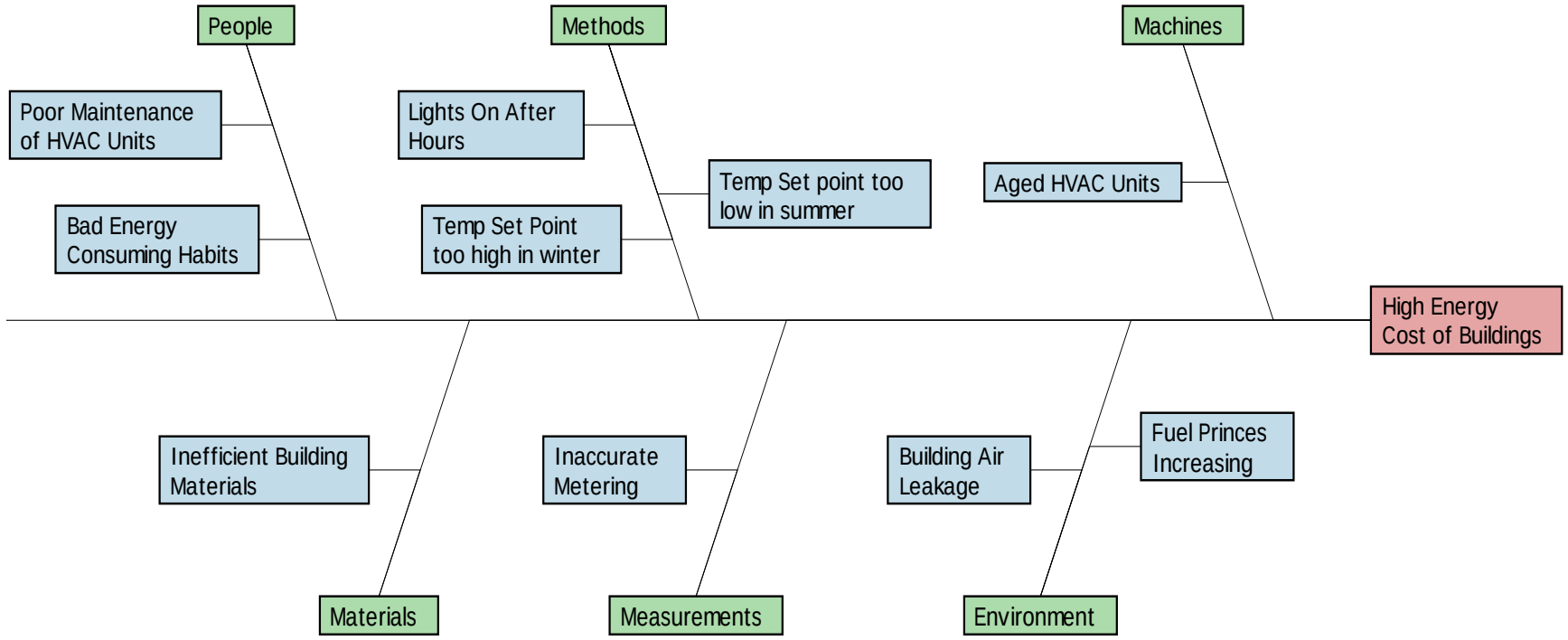
Cause and Effect Diagram Example

- Step 2: Brainstorm the potential causes or factors of the high energy costs.



Cause and Effect Diagram Example

- Step 3: Identify the main categories of causes and group the potential causes accordingly.



Cause and Effect Diagram Example

- Step 4: Analyze the cause and effect (C&E) diagram.
  - After completing the C&E diagram, the real estate company conducts further research on each potential root cause.
  - It is discovered that:
    - The utility metering is accurate
    - The building materials are fine and there is not significant amount of air leakage from the building
    - The fuel prices increased recently but were negligible
    - Most lights are off during the non-business hours except that some lights have to be on for security purposes
    - The temperature set points in the summer and winter are both adequate and reasonable
    - The high energy costs are probably caused by the poor HVAC maintenance on aged units and the wasteful energy consuming habits.
  - Next, the real estate company needs to collect and analyze the data to check whether root causes identified in the C&E diagram are statistically the causes of the high energy costs.

2.1.2 Cause and Effects Matrix

What is a Cause and Effect Matrix?

- The **cause and effect matrix** (XY Matrix) is a tool to help subjectively quantify the relationship of several X's to several Y's.

- Among the Y's under consideration, two important ones should be the *primary* and *secondary metrics* of your Six Sigma project.

- The X's should be derived from your cause and effect diagram. Let us take a peek as what it looks like on the next page.

Cause and Effects Matrix

Lean Six Sigma *XY Matrix*

LEAN SIGMA CORPORATION

Date:

Project:

XY Matrix Owner:

Output Measures (Y's)*	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	
Weighting (1-10):											
Input Variables (X's)#	*For each X, score its impact on each Y listed above (use a 0,3,5,7 scale)*										Score
X_1											0
X_2											0
X_3											0
X_4											0
X_5											0
X_6											0
X_{27}											0
X_{28}											0
X_{29}											0
X_{30}											0

XY Matrix Premis: The XY Matrix or "Cause & Effect Matrix functions on the premis of the Y=f(x) equation.

*Rate each "Y" on a scale of 1 to 10 with 1 being the least important output measure

#For each X rate its impact on each Y using a 0,3,5,7 scale (0=No impact, 3=Weak impact, 5=Moderate impact, 7=Strong

How to Use a Cause and Effect Matrix

1. Across the top enter your output measures. These are the Y's that are important to your project.

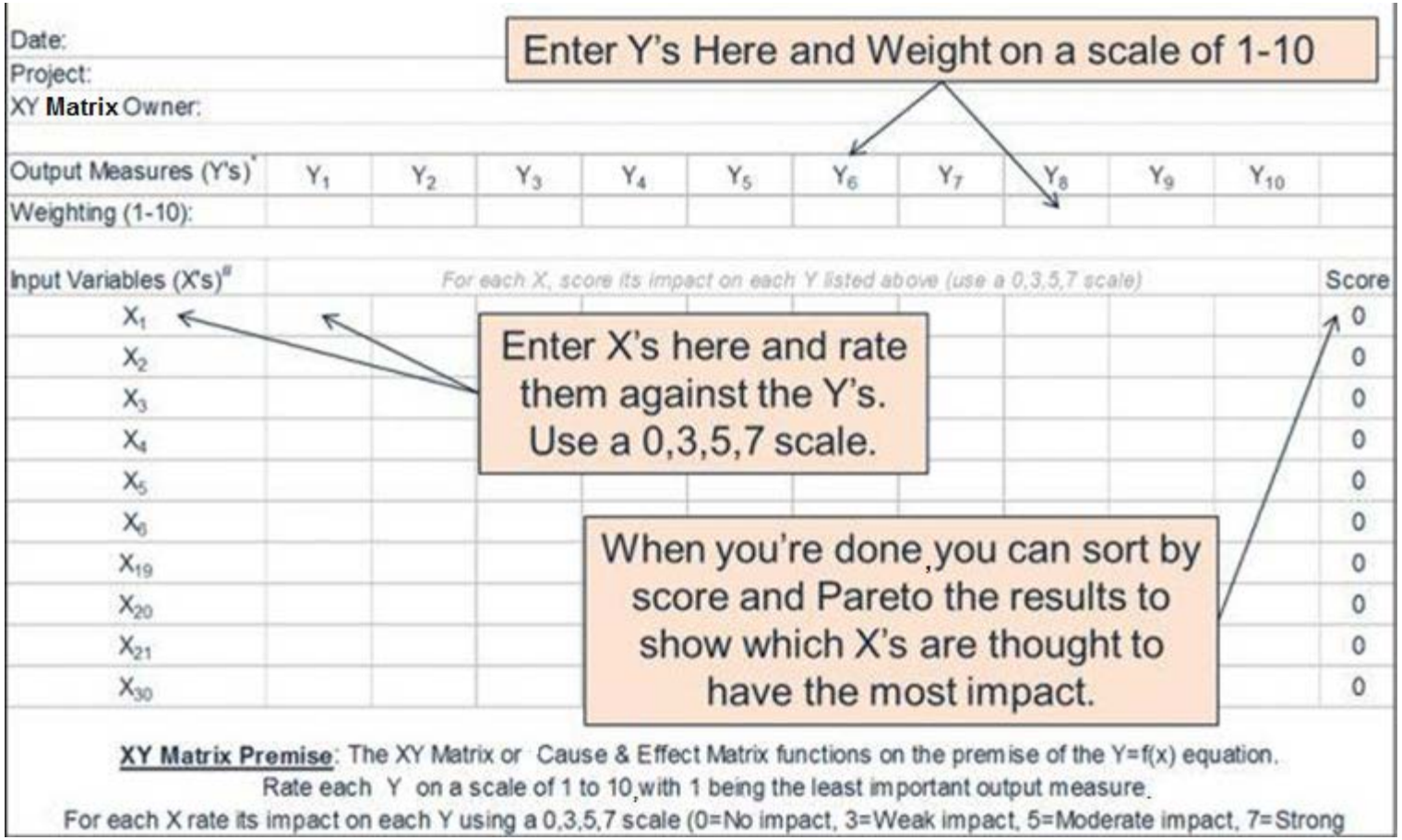
2. Next, give each Y a weight. Use a 1–10 scale, 1 being least important and 10 most important.

3. Below, in the leftmost column, enter all the variables you identified with your cause and effect diagram.

4. Within the matrix itself, rate the strength of the relationship between the X in the row and the corresponding Y in that column. Use a scale of 0, 3, 5, and 7.

5. Lastly, sort the "Score" column to order the most important X's first.

Cause and Effect Matrix Notes



After You Have Completed the C&E Matrix

After you have completed your cause and effects matrix, build a strategy for validating and/or eliminating the x's as significant variables to the Y=f(x) equation.

- Build a data collection plan
- Prepare and execute planned studies
- Perform analytics
- Review results with SMEs
- etc.

2.1.3 Process Mapping

What is a Process Map?

- A **process map** is a graphical representation of a process flow.

- It visualizes how the business process is accomplished step by step.
- It describes how the information or materials sequentially flow from one business entity to the next.
- It illustrates who is responsible for what between the process boundaries.
- It depicts the input and output of each individual process step.

- In the Measure phase, the project team should map the current state of the process instead of the ideal state.

Process Map Basic Symbols

- The following four symbols are the most commonly used symbols in a process map.

Terminator (Oval):
Shows the start/end points in the process.

Process (Rectangle):
Indicates a single process step.

Decision (Diamond):
Indicates a question with two choices (e.g., Yes/No).

Flow Line (Arrow):
Shows the direction of the process flow.

Additional Process Symbols

Alternative Process:
Indicates a process step as the alternate of a normal process step.

Predefined Process:
Indicates a formally defined process step.

Manual Operation:
Indicates a process step conducted manually.

Preparation:
Indicates a preparation step.

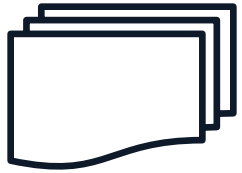
Delay:
Indicates a waiting period in the process.

Additional Process Symbols

- Additional file- and information-related symbols:

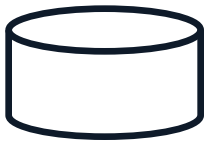
Data (I/O):
Shows the inputs and outputs of a process.

Document:
Indicates a process step that results in a document.



Multi-Document:
Indicates a process step that results in multiple documents.

Stored Data:
Indicates a process step that stores data.

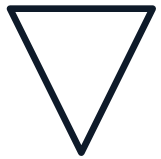


Magnetic Disk:
Indicates a database.

Additional Process Symbols

- Additional control of flow symbols:

Off-Page Connector:
Indicates the process flow continues onto another page.



Merge:
Indicates multiple processes merge into one.

Extract:
Indicates a process splits into multiple parallel processes.

Or:
Indicates a single data processing flow diverges to multiple branches with different criteria requirements.

Summing Junction:
Indicates multiple data processing flows converge into one.

How to Plot a Process Map

- Step 1: Define boundaries of the process you want to map.
  - A process map can depict the flow of an entire process or a segment of it.
  - You need to identify and define the beginning and ending points of the process before starting to plot.
  - Use operational definition.

How to Plot a Process Map

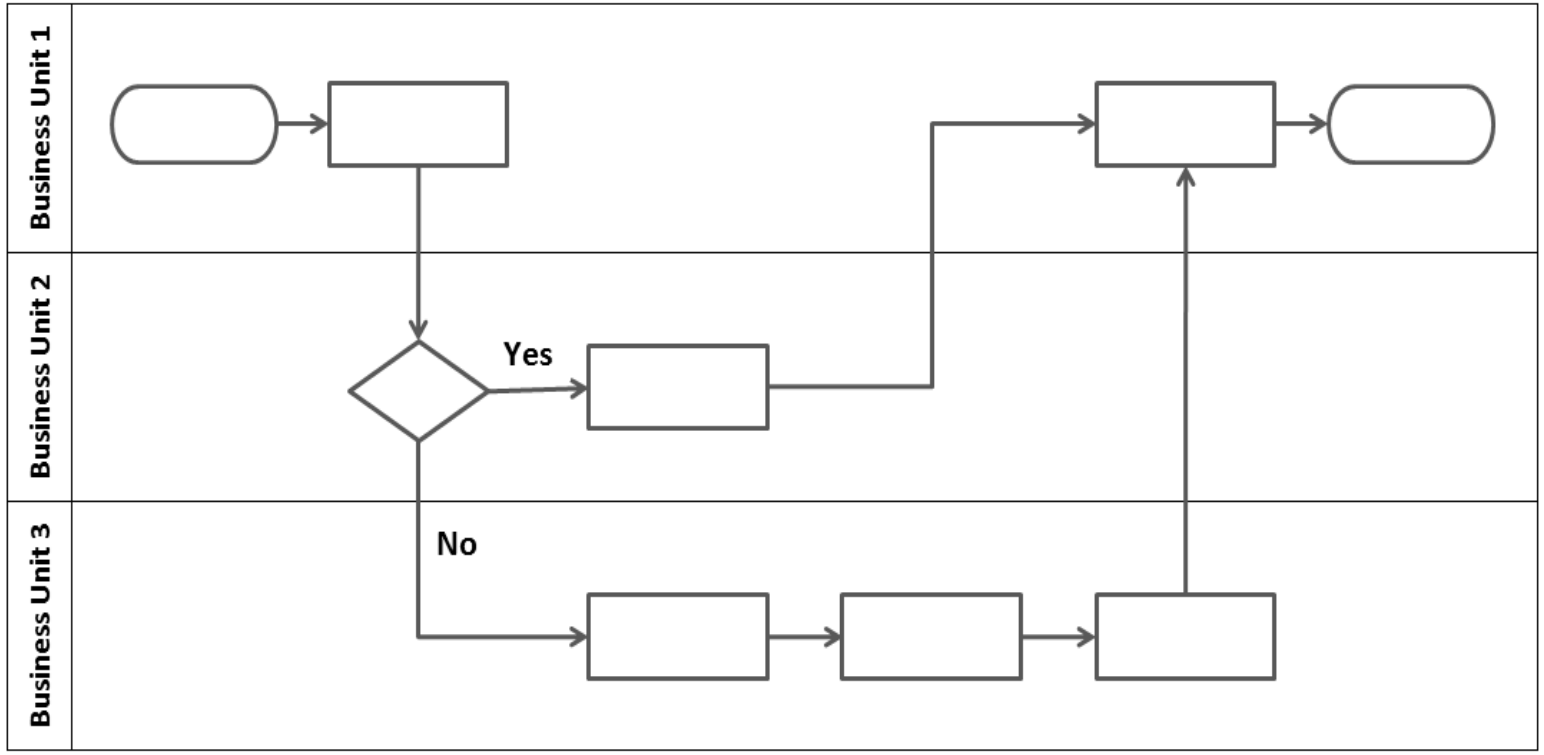
- Step 2: Define and sort the process steps with the flow.
  - Consult with process owners and subject matter experts or observe the process in action to understand how the process is actually performed.
  - Record the process steps and sort them according to the order of their occurrence.

How to Plot a Process Map

- Step 3: Fill the step information into the appropriate process symbols and plot the diagram.
  - In the team meeting of process mapping, place the sticky notes with different colors on a white board to flexibly adjust the under-construction process map.
  - The flow lines are plotted directly on the white board. For the decision step, rotate the sticky note by 45°.
  - When the map is completed on the white board, record the map using Excel, PowerPoint, or Visio.

How to Plot a Process Map

- Step 3:
  - To illustrate the responsibility of different organizations involved in the process, use a Swim Lane Process Map.



How to Plot a Process Map

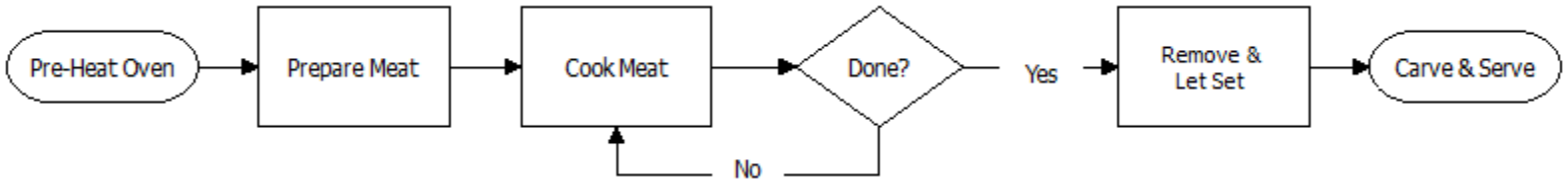
- Step 4: Identify and record the inputs/outputs and their corresponding specifications for each process step.
  - The process map helps in understanding and documenting Y=f(x) of a process where Y represents the outputs and x represents the inputs.
  - The inputs of each process step can be controllable or non-controllable, standardized operational procedure, or noise. They are the source of variation in the process and need to be analyzed qualitatively and quantitatively in order to identify the vital few inputs that have significant effect on the outcome of the process.
  - The outputs of each process step can be products, information, services, etc. They are the little Y's within the process.

How to Plot a Process Map

- Step 5: Evaluate the process map and adjust it if needed.
  - If the process is too complicated to be covered in one single process map, you may create additional detailed sub-process maps for further information.
  - Number the process steps in the order of their occurrence for clarity.

High Level Process Map

- Most high-level business process maps are also referred to as **flow charts**.
- The key to a high-level process map is to over-simplify the process being depicted so that it can be understood in its most generic form.
- As a general rule, high-level process maps should be 4–6 steps and no more.
- Below is an oversimplified version of a high-level process map for cooking a 10lb prime rib for a dozen holiday guests.

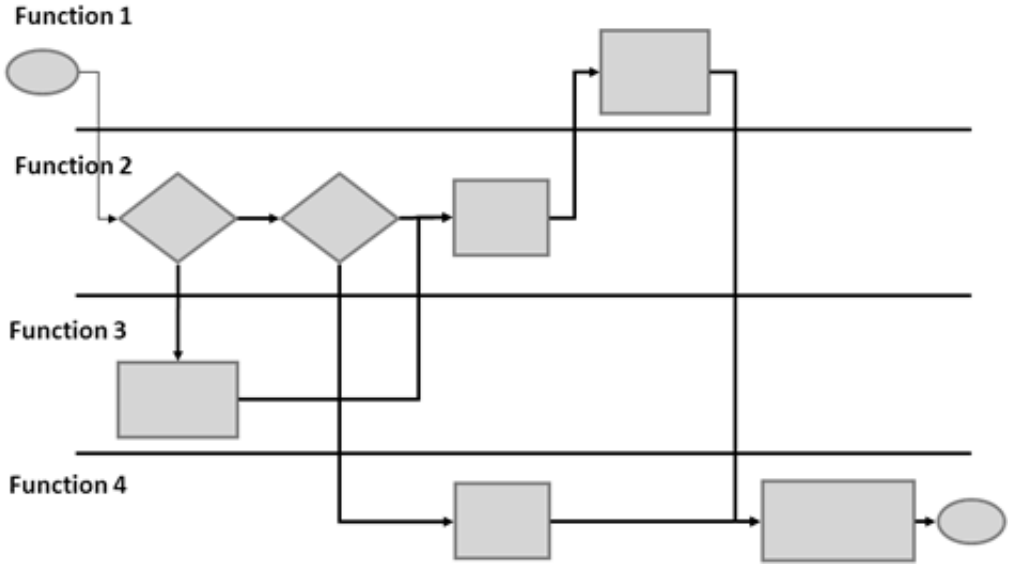


Detailed Process Map

- Detailed process maps or multi-level maps take the high-level map much further.
- Detailed maps can be two, three, or more levels deeper than your high-level process map.
- A good guideline used to help create the second level is to take each step in the high-level map and break it down into another two to four steps each (no more).
- Repeat this process (level 3, level 4 etc.) until reaching the desired level of detail.
- Some detailed maps are two or three levels deep, others can be five or six levels deep. Obviously, the deeper the levels, the more complex and the more burdensome.

Functional Process Map

- The functional map adds dimension to the high-level or detailed map.
- The dimension added is identifying which function or job performs the step or makes the decision.
- Below is a generic example of a functional map. Note that functions are identified in horizontal "lanes" and each process step is placed in the appropriate lane based on which function performs the step.



What is SIPOC?

- A SIPOC (Suppliers-Input-Process-Output-Customers) is a high-level visualization tool to help identify and link the different components in a process.

- It is usually applied in the Measure phase in order to better understand the current state of the process and define the scope of the project.

Key Components of a SIPOC

- **S**uppliers: vendors who provide the raw material, services, and information. Customers can also be suppliers sometimes.
- **I**nput: the raw materials, information, equipment, services, people, environment involved in the process.
- **P**rocess: the high-level sequence of actions and decisions that results in the services or products delivered to the customers.
- **O**utput: the services or products delivered to the customers and any other outcomes of the process.
- **C**ustomers: the end users or recipients of the services or products.

How to Plot a SIPOC Diagram

- The first method:
  - Step 1: Create a template that can contain the information of the five key components in a clear way.
  - Step 2: Plot a high-level process map that covers five steps at maximum.
  - Step 3: Identify the outputs of the process.
  - Step 4: Identify the receipt of the process.
  - Step 5: Brainstorm the inputs required to run each process step.
  - Step 6: Identify the suppliers who provide the inputs.

How to Plot a SIPOC Diagram

- The second method:
  - Step 1: Create a template that can contain the information of the five key components in a clear way.
  - Step 2: Identify the receipt of the process.
  - Step 3: Identify the outputs of the process.
  - Step 4: Plot a high-level process map that covers five steps at maximum.
  - Step 5: Brainstorm the inputs required to run each process step.
  - Step 6: Identify the suppliers who provide the inputs.

Benefits of SIPOC Diagrams

- A SIPOC diagram provides more detailed information than process maps and it demonstrates how each component gets involved in the process.
- It helps visualize and narrow the project scope.
- It serves as a great communication tool to help different process owners understand the entire process, their specific roles and responsibilities.

SIPOC Diagram Example

- Example of plotting a SIPOC diagram for Mom cooking scrambled eggs for two kids

S	I	P	O	C
Supermarket Mom	Eggs Oil Pan Stove Electricity Plate Utensils Salt	Purchase the goods from the supermarket ↓ Crack 6 eggs in a bowl and whisk them well ↓ Heat up the pan over medium heat ↓ Add two teaspoons of oil to the heated pan ↓ Pour the whisked eggs into the pan ↓ Scramble the eggs when they become less liquid ↓ Add 1/2 teaspoon salt to the egges and mix them well ↓ Move the eggs from the pan to a plate and serve.	One plate of scrambled egges	Two kids

Creating a SIPOC

- Step 1: Vertically List High-Level Process
  - If you followed the general rules for a high-level process map, then you should have no more than 4–6 steps for your process.
  - List those steps in a vertical manner as depicted below.

SUPPLIERS	INPUTS	PROCESS	OUTPUTS	CUSTOMERS
		Start		
		Step 1		
		Step 2		
		Step 3		
		Last Step		

Creating a SIPOC

- Step 2: List Process Outputs

SUPPLIERS	INPUTS	PROCESS	OUTPUTS	CUSTOMERS
		Start		
		Step 1	Enter Step 1 Outputs	
		Step 2	Enter Step 2 Outputs	
		Step 3	Enter Step 3 Outputs	
		Last Step	Enter Step 4 Outputs	

Creating a SIPOC

- Step 3: List Output Customers

SUPPLIERS	INPUTS	PROCESS	OUTPUTS	CUSTOMERS
		Start		
		Step 1	Enter Step 1 Outputs	Enter Step 1 Customers
		Step 2	Enter Step 2 Outputs	Enter Step 2 Customers
		Step 3	Enter Step 3 Outputs	Enter Step 3 Customers
		Last Step	Enter Step 4 Outputs	Enter Step 4 Customers

Creating a SIPOC

- Step 4: List Process Inputs

SUPPLIERS	INPUTS	PROCESS	OUTPUTS	CUSTOMERS
		Start		
	Enter Step 1 Inputs	Step 1	Enter Step 1 Outputs	Enter Step 1 Customers
	Enter Step 2 Inputs	Step 2	Enter Step 2 Outputs	Enter Step 2 Customers
	Enter Step 3 Inputs	Step 3	Enter Step 3 Outputs	Enter Step 3 Customers
	Enter Step 4 Inputs	Last Step	Enter Step 4 Outputs	Enter Step 4 Customers

Creating a SIPOC

- Step 5: List Suppliers of Inputs

SUPPLIERS	INPUTS	PROCESS	OUTPUTS	CUSTOMERS
		Start		
Enter Step 1 Suppliers	Enter Step 1 Inputs	Step 1	Enter Step 1 Outputs	Enter Step 1 Customers
Enter Step 2 Suppliers	Enter Step 2 Inputs	Step 2	Enter Step 2 Outputs	Enter Step 2 Customers
Enter Step 3 Suppliers	Enter Step 3 Inputs	Step 3	Enter Step 3 Outputs	Enter Step 3 Customers
Enter Step 4 Suppliers	Enter Step 4 Inputs	Last Step	Enter Step 4 Outputs	Enter Step 4 Customers

SIPOC Benefits

- Visually communicate project scope
- Identify key inputs and outputs of a process
- Identify key suppliers and customers of a process
- Verify:
  - Inputs match outputs for upstream processes
  - Outputs match inputs for downstream processes.
- This type of mapping is effective for identifying opportunities for improvement of your process.

- If you have completed your high-level process map, follow the outlined steps to create a process map of **S**uppliers, **I**nputs, **P**rocess, **O**utputs, and **C**ustomer.

What is Value Stream Mapping?

- **Value stream mapping** is a method to visualize and analyze the path of how information and raw materials are transformed into products or services customers receive.

- It is used to identify, measure, and decrease the non-value-adding steps in the current process.

Non-Value-Added Activities

- **Non-value-adding activities** are activities in a process that do not add any other value to the products or services customers demand.

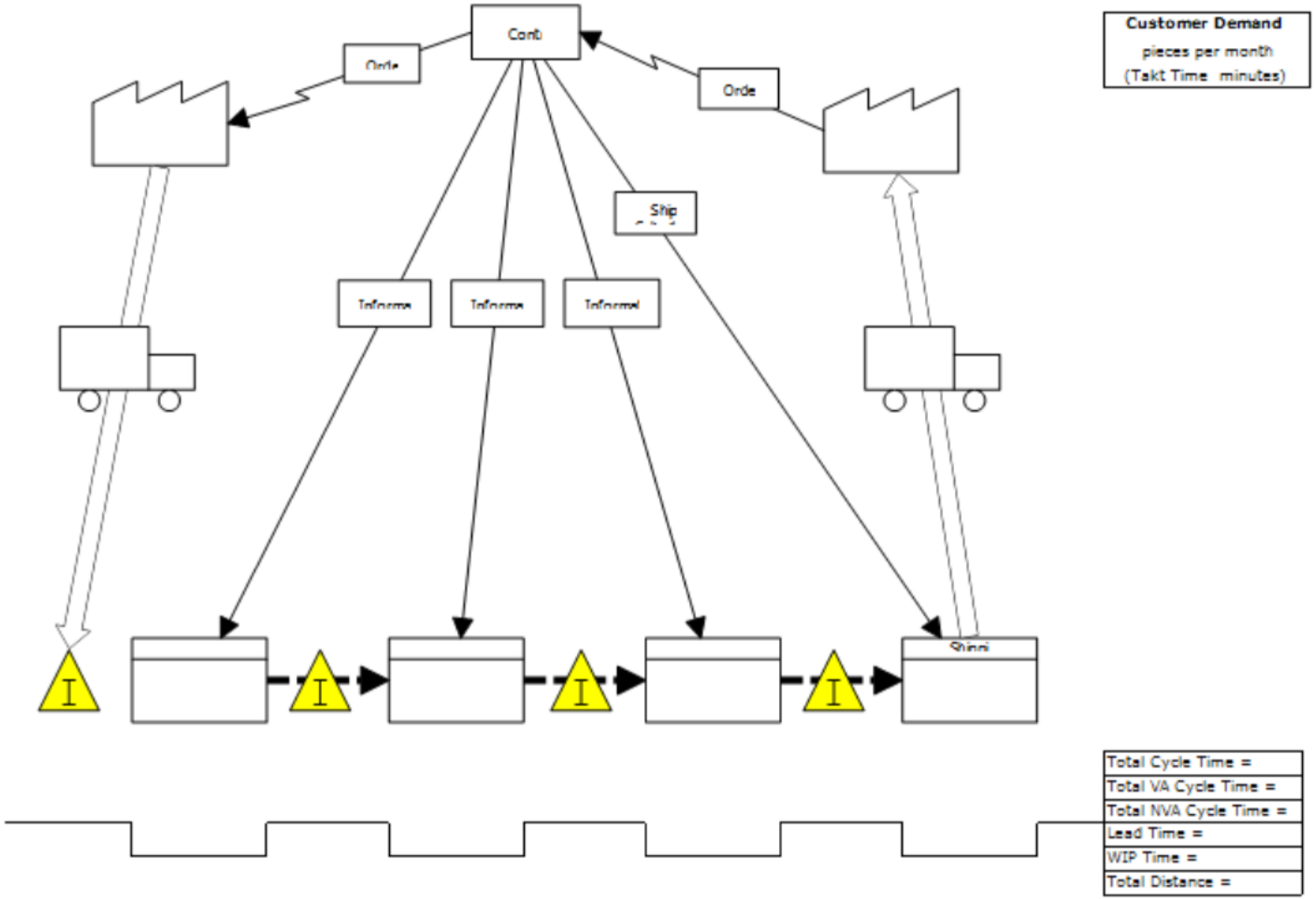
- Example of non-value-adding activities:
  - Rework
  - Overproduction
  - Excess transportation
  - Excess stock
  - Waiting
  - Unnecessary motion.

- Not all non-value-adding activities are unnecessary.

How to Plot a Value Stream Map

- Plot the entire high-level process flow from when the customer places the order to when the customer receives the products or services in the end.
- A value stream map requires more detailed information for each step than the standard process map.
  - Cycle time
  - Preparation time
  - Actual working time
  - Available time
  - Scrap rate
  - Rework rate
  - Number of operators
- Assess the value stream map of current process, identity and eliminate the waste.

Basic Value Stream Map Prototype



Additional Mapping Techniques

- Spaghetti Chart
- Thought Process Mapping

Spaghetti Chart

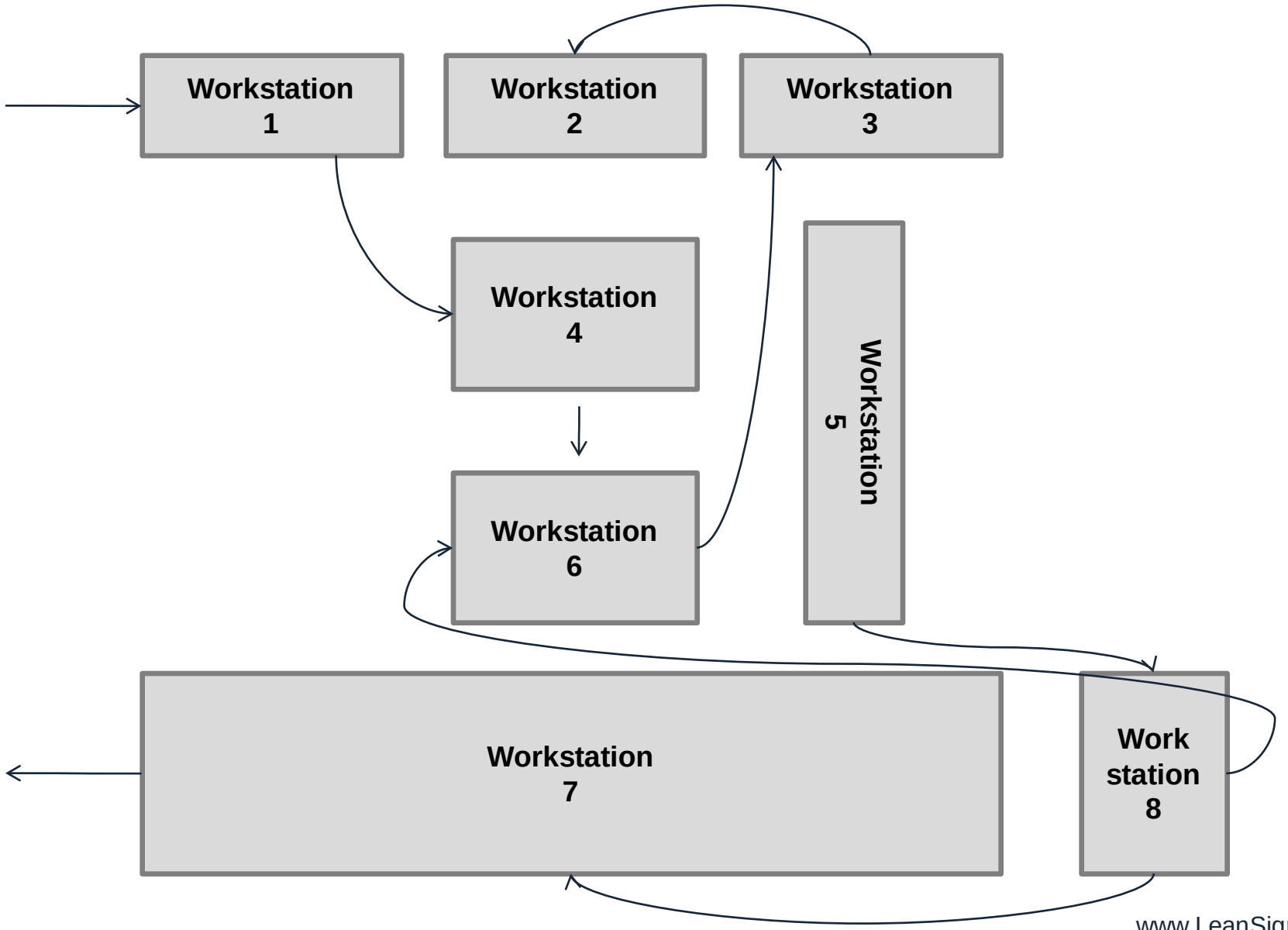
- A **spaghetti chart** is a graphical tool to map out the physical flow of materials, information, and people involved in a process. It can also reflect the distances between multiple workstations the physical flow has been through.

- A process that has not been streamlined has messy and wasteful movements of materials, information, and people, resembling a bowl of cooked spaghetti.

How to Plot a Spaghetti Chart

- Step 1: Create a map of the work area layout.
- Step 2: Observe the current work flow and draw the actual work path from the very beginning of work to the end when products exit the work area.
- Step 3: Analyze the spaghetti chart and identify improvement opportunities.

Spaghetti Chart Example



Thought Process Mapping

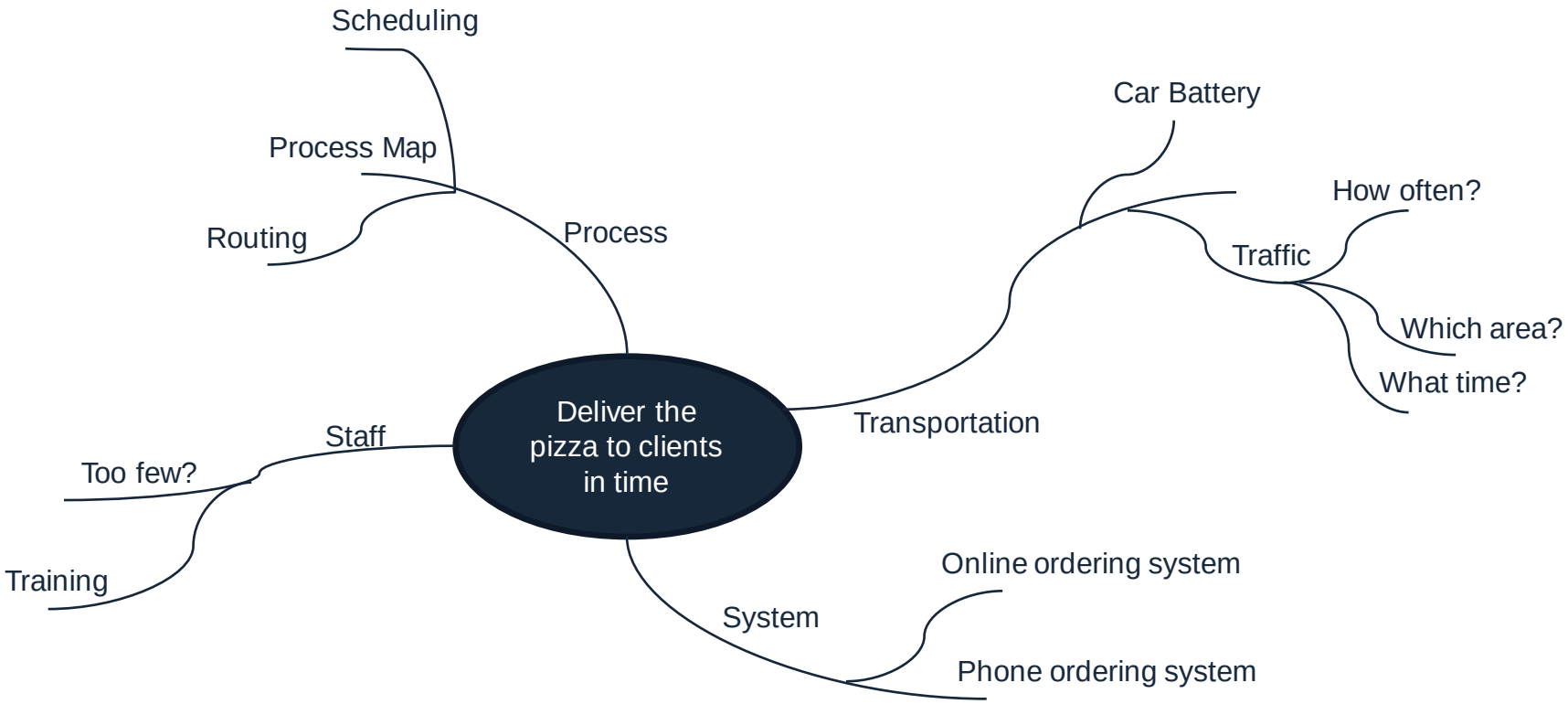
- A **thought process map** is a graphical tool to help brainstorm, organize, and visualize the information, ideas, questions, or thoughts regarding reaching the project goal.

- It is a popular tool generally used at the beginning of a project in order to:
  - identify knowns and unknowns
  - communicate assumptions and risks
  - discover potential problems and solutions
  - identify resources, information, and actions required to meet the goal
  - present relationship of thoughts.

How to Plot a Thought Process Map

- Step 1: Define the project goal.
- Step 2: Brainstorm knowns and unknowns about the project.
- Step 3: Brainstorm questions and group the unknowns and questions into five phases (**D**efine, **M**easure, **A**nalyze, **I**mprove, and **C**ontrol).
- Step 4: Sequence the questions below the project goal and link the related questions.
- Step 5: Identify tools or methods that would be used to answer the questions.
- Step 6: Repeat steps 3 to 5 as the project continues.

Thought Process Map Example



2.1.4 FMEA

What is FMEA?

- The FMEA (Failure Modes and Effects Analysis) is an analysis technique to identify, evaluate, and prioritize a potential deficiency in a process so that the project team can design action plans to reduce the probability of the failure/deficiency occurring.

LEAN SIGMA CORPORATION

Failure Modes & Effects Analysis - **FMEA**

Product or Process Step	Potential Failure Mode	Potential Failure Effects	S	Potential Causes	O	Current Controls	D	RPN	Recommended Actions	Responsible	Actions Taken	S	O	D	RPN
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0
								0							0

- FMEA is completed in cross-functional brainstorming sessions in which attendees have a good understanding of the entire process or of a segment of it.

Basic FMEA Terms

- **Process Functions**
  - Process steps depicted in the process map. FMEA is based on a process map and one step/function is analyzed at a time.
- **Failure Modes**
  - Potential and actual failure in the process function/step. It usually describes the way in which failure occurs. There might be more than one failure mode for one process function.
- **Failure Effects**
  - Impact of failure modes on the process or product. One failure mode might trigger multiple failure effects.
- **Failure Causes**
  - Potential defect of the design that might result in the failure modes occurring. One failure mode might have multiple potential failure causes.
- **Current Controls**
  - Procedures currently conducted to prevent failure modes from happening or to detect the failure mode occurring.

Basic FMEA Terms

- **Severity Score**
  - The seriousness of the consequences of a failure mode occurring.
  - Ranges from 1 to 10, with 10 indicating the most severe consequence.

- **Occurrence Score**
  - The frequency of the failure mode occurring.
  - Ranges from 1 to 10, with 10 indicating the highest frequency.

- **Detection Score**
  - How easily failure modes can be detected.
  - Ranges from 1 to 10, with 10 indicating the most difficult detection.

Basic FMEA Terms

- **RPN (Risk Prioritization Number)**
  - The product of the severity, occurrence, and detection scores.
  - Ranges from 1 to 1000.
  - The higher RPN is, the more focus the particular step/function needs.

- **Recommended Actions**
  - The action plan recommended to reduce the probability of failure modes occurring.

How to Conduct an FMEA

- Step 1: List the critical functions of the process based on the process map created.
- Step 2: List all potential failure modes that might occur in each function. One function may have multiple potential failures.
- Step 3: List all potential failure effects that might affect the process or product.
- Step 4: List all possible causes that may lead to the failure mode happening.
- Step 5: List the current control procedures for each failure mode.

How to Conduct an FMEA

- Step 6: Determine the severity rating for each potential failure mode.
- Step 7: Determine the occurrence rating for each potential failure cause.
- Step 8: Determine the detection rating for each current control procedure.
- Step 9: Calculate RPN (Risk Prioritization Number).
- Step 10: Rank the failures using RPN and determine the precedence of problems or critical inputs of the process. A Pareto chart might help to focus on the failure modes with high RPNs. The higher the RPN, the higher the priority the correction action plan.

How to Conduct an FMEA

- Step 11: Brainstorm and create recommended action plans for each failure mode.
- Step 12: Determine and assign the task owner and projected completion date to take actions.
- Step 13: Determine the new severity rating if the actions are taken.
- Step 14: Determine the new occurrence rating if the actions are taken.
- Step 15: Determine the new detection rating if the actions are taken.
- Step 16: Update the RPN based on new severity, occurrence, and detection ratings.

FMEA Example

Case study:

- Joe is trying to identify, analyze, and eliminate the failure modes he experienced in the past when preparing his work bag before heading to the office every morning. He decides to run an FMEA for his process of work bag preparation.

- There are only two steps involved in the process.
  - Putting the work files in the bag
  - Putting a water bottle in the bag.

FMEA Example

- Step 1: List the critical functions of the process based on the process map created.

Product or Process Step	Potential Failure Mode	Potential Failure Effects
Place files in bag		
Put water bottle in bag		

- Step 2: List all the potential failure modes for each function.

Product or Process Step	Potential Failure Mode	Potential Failure Effects
Place files in bag	Incorrect files put in the bag	
Put water bottle in bag	Water leaks	

- Step 3: List potential failure effects that might affect the process.

Product or Process Step	Potential Failure Mode	Potential Failure Effects
Place files in bag	Incorrect files put in the bag	Work is delyed
Put water bottle in bag	Water leaks	Files in bag damaged

FMEA Example

- Step 4: List all possible causes to the failure mode.

Potential Failure Effects	S	Potential Causes	O
Work is delyed		Files are not organized well	
Files in bag damaged		Cap on water bottle not tight	

- Step 5: List any control procedures for each failure mode.

Potential Causes	O	Current Controls	D
Files are not organized well		Check if files are needed	
Cap on water bottle not tight		Check bottle cap before inserting	

- Step 6: Determine the severity rating for each failure mode.

Potential Failure Effects	S	Potential Causes	O	Current Controls
Work is delyed	9	Files are not organized well		Check if files are needed
Files in bag damaged	7	Cap on water bottle not tight		Check bottle cap before inserting

FMEA Example

- Step 7: Determine the occurrence rating for each failure cause.

Potential Failure Effects	S	Potential Causes	O	Current Controls
Work is delyed	9	Files are not organized well	3	Check if files are needed
Files in bag damaged	7	Cap on water bottle not tight	5	Check bottle cap before inserting

- Step 8: Determine the detection rating for each control.

S	Potential Causes	O	Current Controls	D	RPN
9	Files are not organized well	3	Check if files are needed	5	135
7	Cap on water bottle not tight	5	Check bottle cap before inserting	5	175

- Step 9: Calculate the RPN (Risk Prioritization Number).

S	Potential Causes	O	Current Controls	D	RPN
9	Files are not organized well	3	Check if files are needed	5	135
7	Cap on water bottle not tight	5	Check bottle cap before inserting	5	175

FMEA Example

- Step 10: Rank the failures using the RPN and determine the precedence of problems or critical inputs of process.

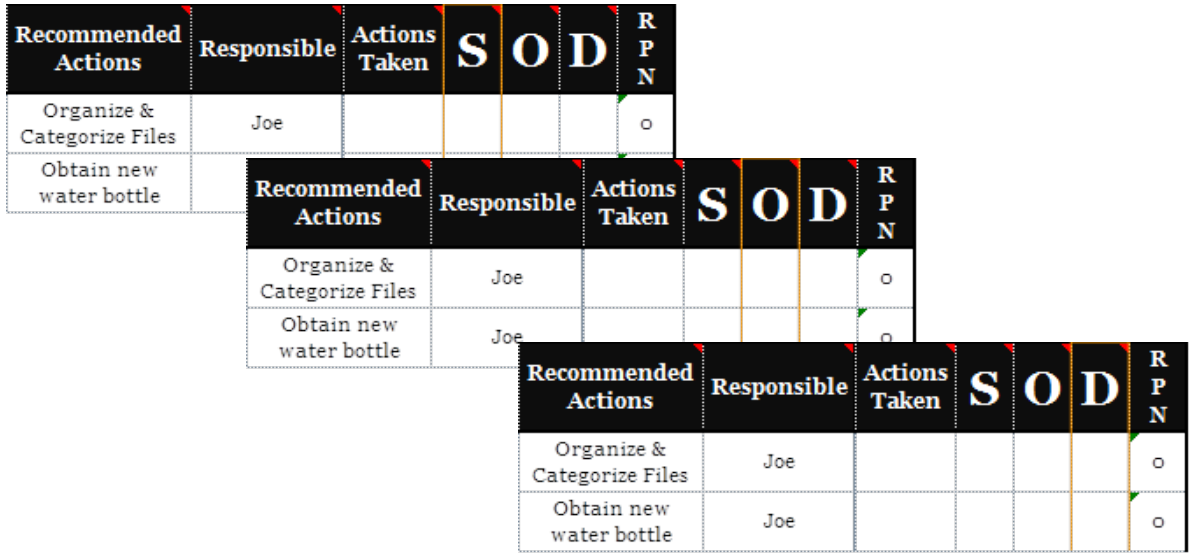
Current Controls	D	RPN	Recommended Actions
Check bottle cap before inserting	5	175	Organize & Categorize Files
Check if files are needed	5	135	Obtain new water bottle

- Step 11: Brainstorm and create recommended action plans.
- Step 12: Determine and assign owners with completion dates.

D	RPN	Recommended Actions	Responsible
5	175	Organize & Categorize Files	Joe
5	135	Obtain new water bottle	Joe

FMEA Example

- Steps 13-15: Determine new severity, occurrence and detection ratings if actions are taken.



Recommended Actions	Responsible	Actions Taken	S	O	D	RPN
Organize & Categorize Files	Joe					0
Obtain new water bottle	Joe					0

- Step 16: Update RPN based on new ratings.

Recommended Actions	Responsible	Actions Taken	S	O	D	RPN
Organize & Categorize Files	Joe					0
Obtain new water bottle	Joe					0

2.1.5 Theory of Constraints

What is the Theory of Constraints?

- Processes, systems, and organizations are vulnerable to their weakest part.
- Any manageable system is limited by constraints in its ability to produce more (and there is always at least one constraint).

Performance Measures

Making sound financial decisions based on these three measures is a critical requirement.

- **Throughput** – rate at which a system generates money through sales.
- **Operational Expense** – money spent by the system to turn inventory into throughput.
- **Inventory** – money the system has invested in purchasing things it intends to sell.

Five Focusing Steps

Objective: To ensure ongoing improvement efforts are focused on the constraints of a system.

1. Identify the system's constraint(s).
2. Decide how to exploit the constraint(s).
3. Subordinate everything else to the decision in step 2.
4. Elevate the constraint(s).
5. If in previous steps a constraint has been broken, return to step 1, but do not allow inertia to cause a system's constraint.

Logical Thinking Processes

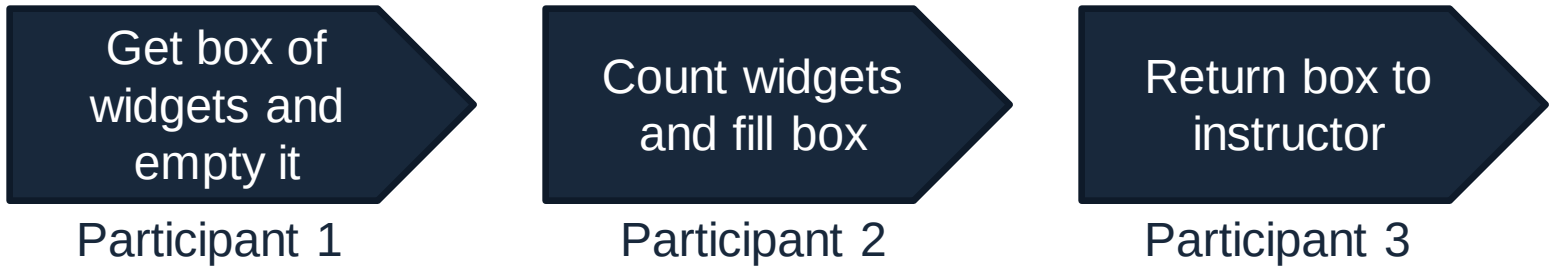
	Focusing Step	Thinking Process	Tools
1	Identify the system's constraint(s)	• Identify the problems • Find the root causes	Cause and effect diagram
2	Decide how to exploit the constraint(s)	• Develop a solution	Future reality tree
3	Subordinate everything else to the decision in step 2	• Identify the conflict preventing the solution • Remove the conflict	Evaporating cloud
4	Elevate the constraint	• Construct and execute an implementation plan	Prerequisite tree Transition tree
5	If in previous steps a constraint has been broken, return to step 1, but do not allow inertia to cause a system's constraint		

Simulation Exercise

Resources needed:

- 3 “production line” participants
- 1 timer per each “production line” participant
- 5 small boxes of 15 widgets each (paperclips, pens/pencils, candy, etc.)

Widget Value Chain:



Simulation Exercise

Widget Value Chain:

Five focusing steps:

1. Identify the system's constraint(s).
2. Decide how to exploit the constraint(s).
3. Subordinate everything else to the decision in step 2.
4. Elevate the constraint(s).
5. If in previous steps a constraint has been broken, return to step 1, but do not allow inertia to cause a system's constraint.

2.2 Six Sigma Statistics

Black Belt Training: Measure Phase

2.1 Process Definition
- 2.1.1 Cause and Effect Diagrams
- 2.1.2 Cause and Effects Matrix
- 2.1.3 Process Mapping
- 2.1.4 FMEA: Failure Modes and Effects Analysis
- 2.1.5 Theory of Constraints

2.2 Six Sigma Statistics
- 2.2.1 Basic Statistics
- 2.2.2 Descriptive Statistics
- 2.2.3 Distributions and Normality
- 2.2.4 Graphical Analysis

2.3 Measurement System Analysis
- 2.3.1 Precision and Accuracy
- 2.3.2 Bias, Linearity, and Stability
- 2.3.3 Gage R&R
- 2.3.4 Variable and Attribute MSA

2.4 Process Capability
- 2.4.1 Capability Analysis
- 2.4.2 Concept of Stability
- 2.4.3 Attribute and Discrete Capability
- 2.4.4 Monitoring Techniques

2.2.1 Basic Statistics

What is Statistics?

- **Statistics** is the science of collection, analysis, interpretation, and presentation of data.

- In Six Sigma, we apply statistical methods and principles to quantitatively measure and analyze the process performance to reach statistical conclusions and help solve business problems.

Types of Statistics

- Descriptive Statistics
  - Describing what was going on

- Inferential Statistics
  - Making inferences from the data at hand to more general conditions

Descriptive Statistics

- **Descriptive statistics** is applied to describe the main characteristics of a collection of data.
- Descriptive statistics summarizes the features of the data quantitatively.
- Descriptive statistics is descriptive only and it does not make any generalizations beyond the data at hand.
- The data used for descriptive statistics are for the purpose of representing or reporting.

Inferential Statistics

- **Inferential statistics** is applied to infer the characteristics or relationships of the populations from which the data are collected.

- Inferential statistics draws statistical conclusions about the population by analyzing the sample data subject to random variation.

- A complete data analysis includes both descriptive statistics and inferential statistics.

Statistics vs. Parameters

- The word *statistic* refers to a numeric measurement calculated using a sample data set, for example, sample mean or sample standard deviation. Its plural is *statistics* (the same spelling as “statistics” which refers to the scientific discipline).

- The *parameter* refers to a numeric metric describing the population, for example, population mean and population standard deviation. Unless you have the full data set of the population, you will not be able to know the population parameters.

Continuous Variable vs. Discrete Variable

- Continuous Variable
  - Measured
  - There is an infinite number of values possible
  - Examples: temperature, height, weight, money, time

- Discrete Variable
  - Counted
  - There is a finite number of values available
  - Examples: count of people, count of countries, count of defects, count of defectives

Types of Data

- Nominal
  - Categorical data
  - Examples: a set of colors, the social security number
- Ordinal
  - Rank-ordering data
  - Examples: the first, second place in a race, scores of exams
- Interval
  - Equidistant data
  - Examples: temperature with Fahrenheit or Celsius scale
- Ratio
  - The ratio between the magnitude of a continuous value and the unit value of the same category
  - Examples: weight, length, time

2.2.2 Descriptive Statistics

Basics of Descriptive Statistics

- Descriptive statistics provides a quantitative summary for the data collected.

- It summarizes the main features of the collection of data.
  - Shape
  - Location
  - Spread

- It is a presentation of data collected and it does *not* provide any inferences about a more general condition.

Shape of the Data

- **Distribution** is used to describe the shape of the data.

- Distribution (also called frequency distribution) summarizes the frequency of an individual value or a range of values of a variable (either continuous or discrete).

- Distribution is depicted as a table or graph.

Shape of the Data

- Simple example of distribution
  - We are tossing a fair die. The possible value we obtain from each tossing is a value between 1 and 6.
  - Each value between 1 and 6 has a 1/6 chance to be hit for each tossing.
  - The distribution of this game describes the relationship between every possible value and the percentage of times the value is being hit (or count of times the value is being hit).

Shape of the Data

- Examples of continuous distribution
  - Normal Distribution
  - T distribution
  - Chi-square distribution
  - F distribution

- Examples of discrete distribution
  - Binomial distribution
  - Poisson distribution

Location of the Data

- The **location** (i.e. central tendency) of the data describes the value where the data tend to cluster around.

- There are multiple measurements to capture the location of the data:
  - Mean
  - Median
  - Mode.

Mean

- The **mean** is the arithmetic average of a data set.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

n is the number of values in the data set

- For example, we have a set of data: 2, 3, 5, 8, 5, and 9. The arithmetic mean of the data set is

$$\frac{2+3+5+8+5+9}{6} = 5.33$$

Median

- The **median** is the middle value of the data set in numeric order.
- It separates the finite set of data into two parts: one with values higher that the median and the other with values lower than the median.
- For example, we have a set of data: 45, 32, 67, 12, 37, 54 and 28. The median is 37 since it is the middle value of the sorted list of values (i.e. 12, 28, 32, 37, 45, 54 and 67).

Mode

- The **mode** is the value that occurs most often in the data set.
- If no number is repeated, there is no mode for the data set.
- For example, we have a data set: 55, 23, 45, 45, 68, 34, 45, 55. The mode is 45 since it occurs most frequently.

Spread of the Data

- The **spread** (i.e. variation) of the data describes the degree of data dispersing around the center value.

- There are multiple measurements to capture the spread of the data:
  - Range
  - Variance
  - Standard Deviation.

Range

- The **range** is the numeric difference between the greatest and smallest values in a data set.

- Only two data values (i.e. the greatest and the smallest values) are accounted for calculating the range.

- For example, we have a set of data: 34, 45, 23, 12, 32, 78 and 23. The range of the data is 78−12 = 66.

Variance

- The **variance** measures how far on average the data points spread out from the mean.
- It is the average squared deviation of each value from its mean.
- All the data points are accounted for calculating the variance.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where

n is the number of values in the data set

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Standard Deviation

- **Standard deviation** describes how far the data points spread away from the mean.
- It is simply the square root of the variance.
- All the data points are accounted for calculating the standard deviation.

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where

n is the number of values in the data set

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2.2.3 Normal Distribution & Normality

What is Normal Distribution?

- The **normal distribution** is a probability distribution of a continuous random variable whose values spread symmetrically around the mean.

- A normal distribution can be completely described by using its mean (μ) and variance (σ^2).

- When a variable x is normally distributed, we denote $x \sim N(\mu, \sigma^2)$.

Z Distribution

- The **Z distribution** is the simplest normal distribution with the mean equal to zero and the variance equal to one.
- Any normal distribution can be transferred to a Z distribution by applying

$$z = \frac{x - \mu}{\sigma}$$

where

$$x \sim N\left(\mu, \sigma^2\right) \qquad \sigma \neq 0$$

Z Score

- The **Z Score** is the measure of how many standard deviations an observation is above or below the mean.
- Positive Z Scores indicate the observation is above the mean or “right of the mean”.
- Negative Z Scores indicate the observation is below the mean of “left of the mean”
- Calculate Z Score using the formula below:

$$z = \frac{x - \mu}{\sigma}$$

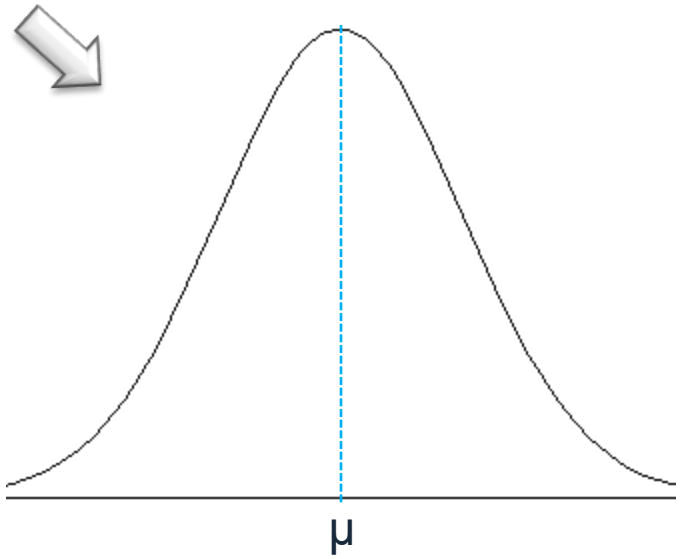
where

x is the observation
μ is the mean of the population
σ is the standard deviation of the population

Shape of Normal Distribution

- The probability density function curve of normal distribution is bell-shaped.
- Probability density function of normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Location of Normal Distribution

- If a variable is normally distributed, the mean, the median, and the mode have the same value.

- The probability density curve of normal distribution is symmetric around a center value which is the mean, the median, and the mode at the same time.

Spread of Normal Distribution

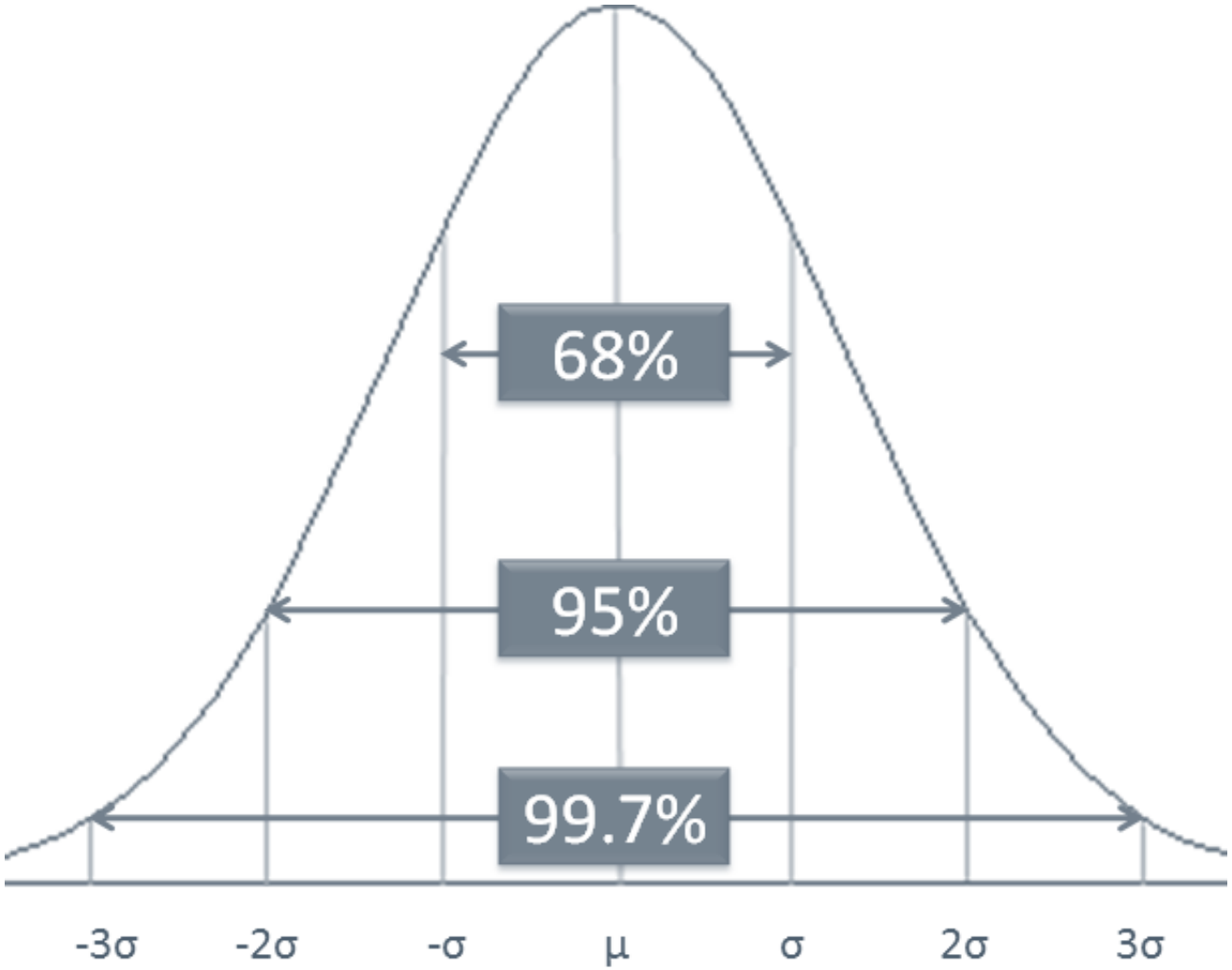
- The spread or variation of the normally-distributed data can be described using the variance or the standard deviation.

- The smaller the variance or the standard deviation, the less variability in the data set.

68-95-99.7 Rule

- The 68-95-99.7 rule or the *empirical rule in statistics* states that for a normal distribution:
  - About 68% of the data fall within one standard deviation of the mean
  - About 95% of the data fall within two standard deviations of the mean
  - About 99.7% of the data fall within three standard deviations of the mean.

68-95-99.7 Rule



Normality

- Not all the distributions with a bell shape are normal distributions.
- To check whether a group of data points are normally distributed, we need to run a *normality test*.
- There are different normality tests available:
  - Anderson-Darling test
  - Sharpiro-Wilk test
  - Jarque-Bera test.
- More details of normality test will be introduced in the Analyze module.

Normality Testing

- To check whether the population of our interest is normally distributed, we need to run normality test.
  - Null Hypothesis (H_0): The data points *are* normally distributed.
  - Alternative Hypothesis (H_a): The data points are *not* normally distributed.
- There are many normality tests available. For example, Anderson-Darling test, Sharpiro-Wilk test, Jarque-Bera test, and so on.

Use SigmaXL to Run a Normality Test

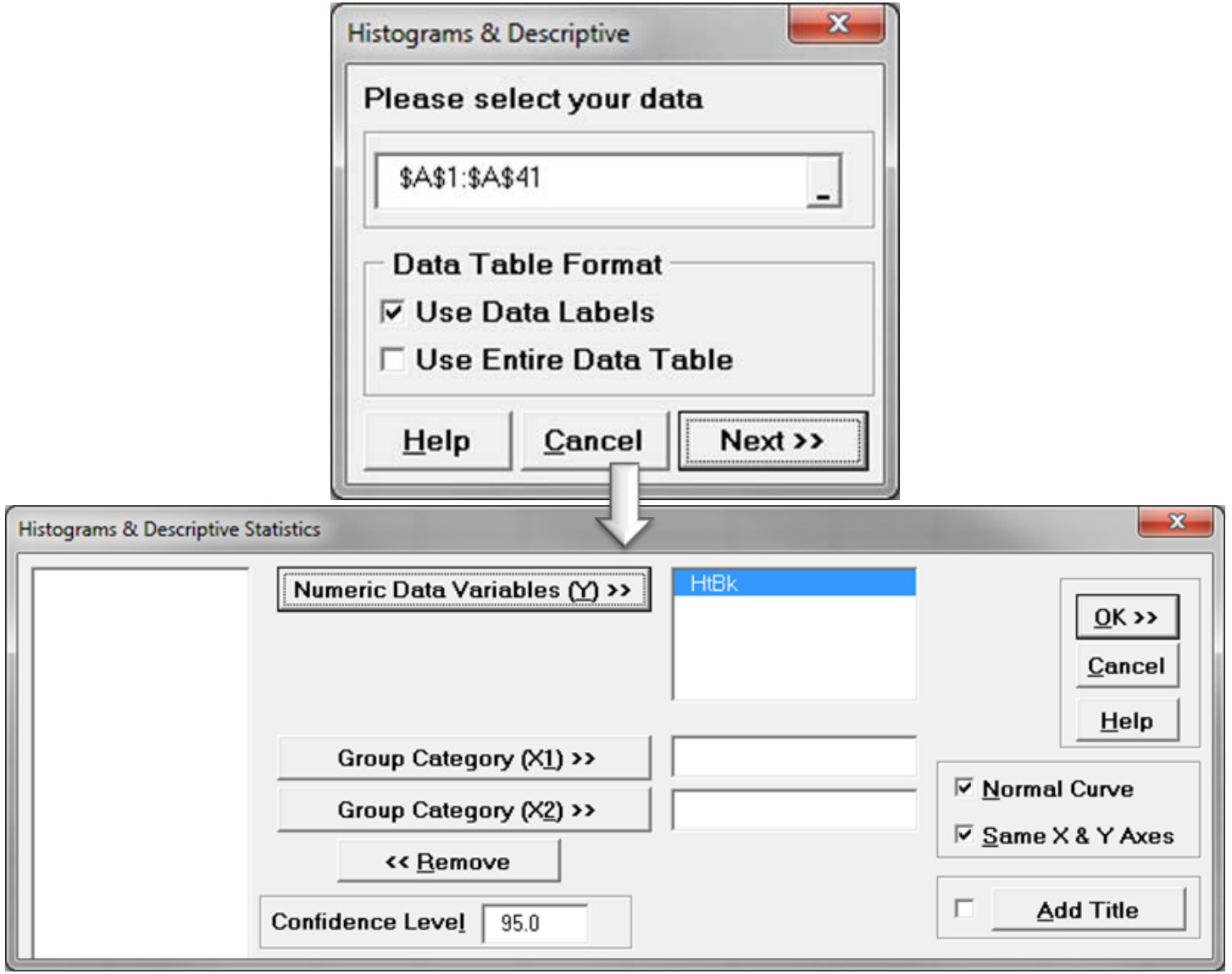
- Case Study: we are interested to know whether the height of basketball players is normally distributed.
- Data File: “One Sample T-Test” tab in “Sample Data.xlsx”

- Null Hypothesis (H0): the height of basketball players is normally distributed.
- Alternative Hypothesis (Ha): the height of basketball players is not normally distributed.

Use SigmaXL to Run a Normality Test

- Steps to run a normality test in SigmaXL
  - Select the entire range of data
  - Click SigmaXL -> Graphical Tools -> Histograms & Descriptive Statistics
  - A new window named “Histograms & Descriptive” pops up with the selected range appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Histograms & Descriptive Statistics” appears
  - Select “HtBk” as the “Numeric Data Variables (Y)”
  - Click “OK”
  - The normality test results appear in the newly generated tab “Hist Descript (1)”

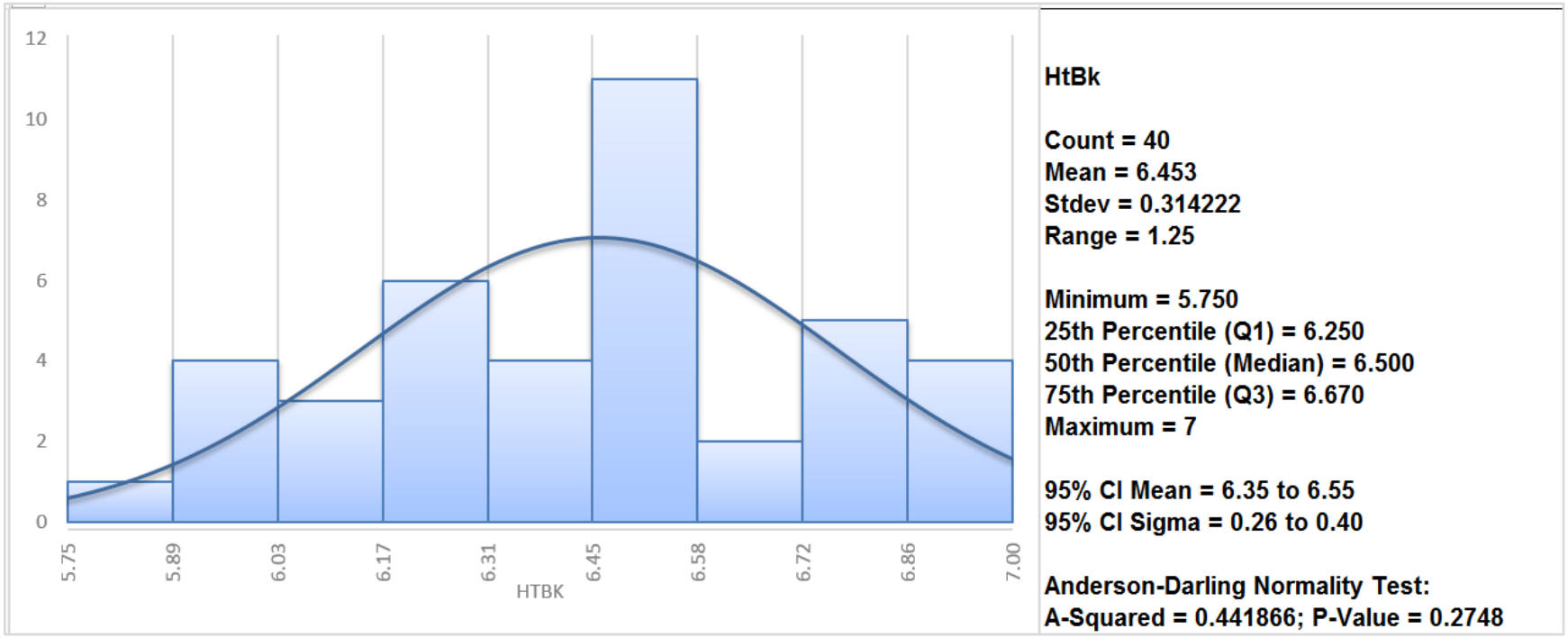
Use SigmaXL to Run a Normality Test



Use SigmaXL to Run a Normality Test

- Null Hypothesis (H_0): The data are normally distributed.
- Alternative Hypothesis (H_a): The data are not normally distributed.

- Since the p-value of the normality is 0.2748 greater than alpha level (0.05), we fail to reject the null and claim that the data are normally distributed.



2.2.4 Graphical Analysis

What is Graphical Analysis?

- In statistics, **graphical analysis** is a method to visualize the quantitative data.

- Graphical analysis is used to discover the structure and patterns in the data, explaining and presenting the statistical conclusions.

- A complete statistical analysis includes both quantitative analysis and graphical analysis.

Graphical Analysis Example

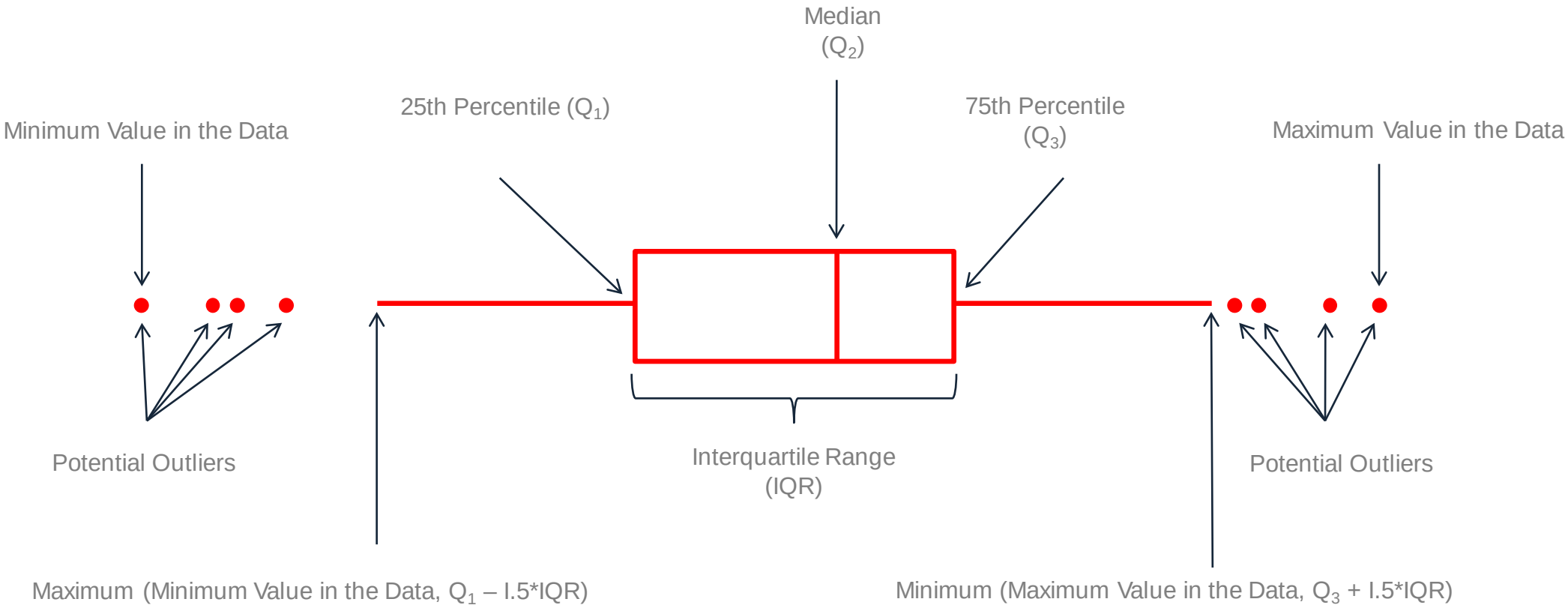
- There are various graphical analysis tools available. Here are four most commonly used examples:
  - Box Plot
  - Histogram
  - Scatter Plot
  - Run Chart.

Box Plot

- A **box plot** is a graphical method to summarize a data set by visualizing the minimum value, 25th percentile, median, 75th percentile, the maximum value, and potential outliers.

- A percentile is the value below which a certain percentage of data fall. For example, if 75% of the observations have values lower than 685 in a data set, then 685 is the 75th percentile of the data.

Box Plot

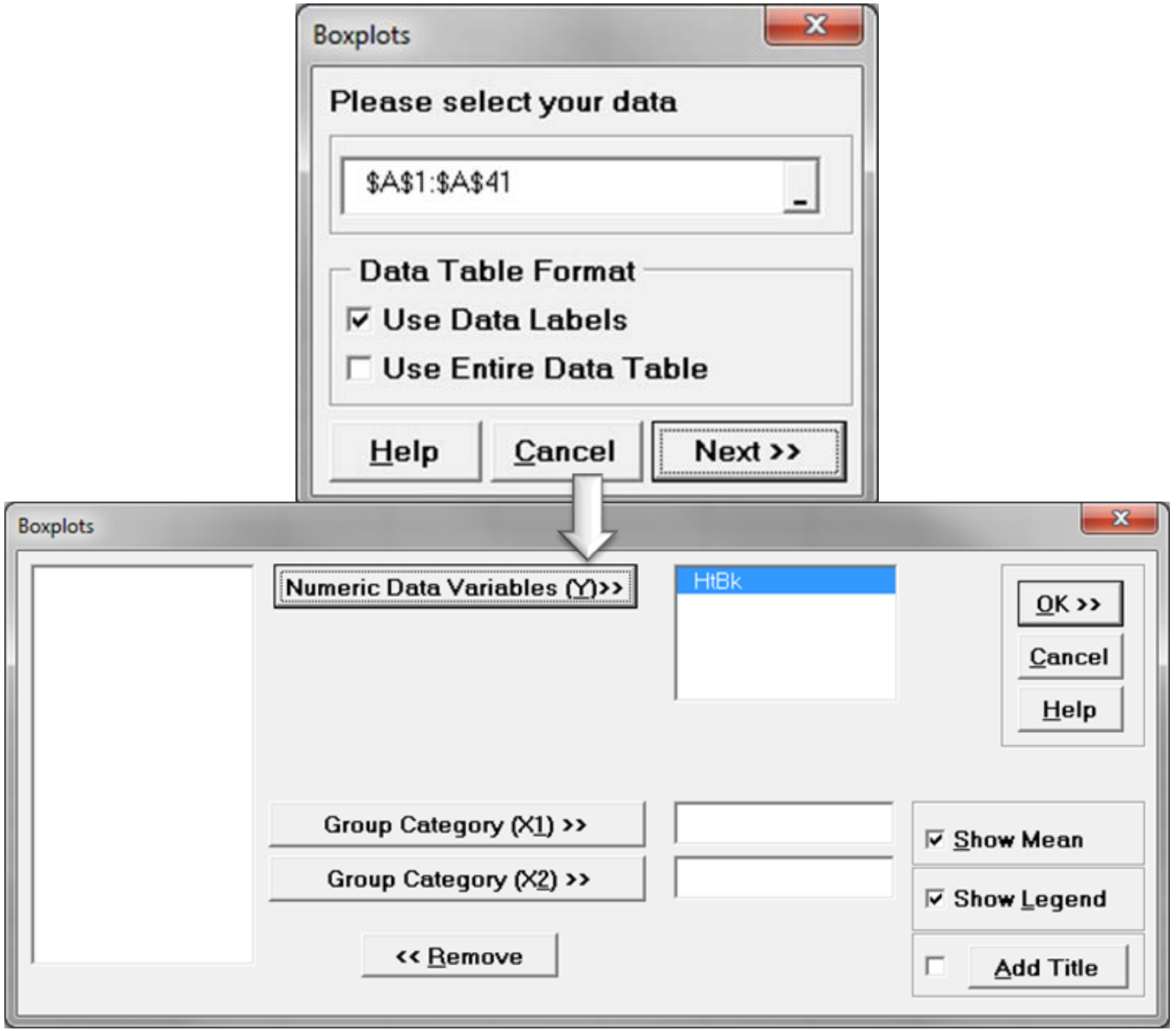


Interquartile Range = 75th Percentile – 25th Percentile

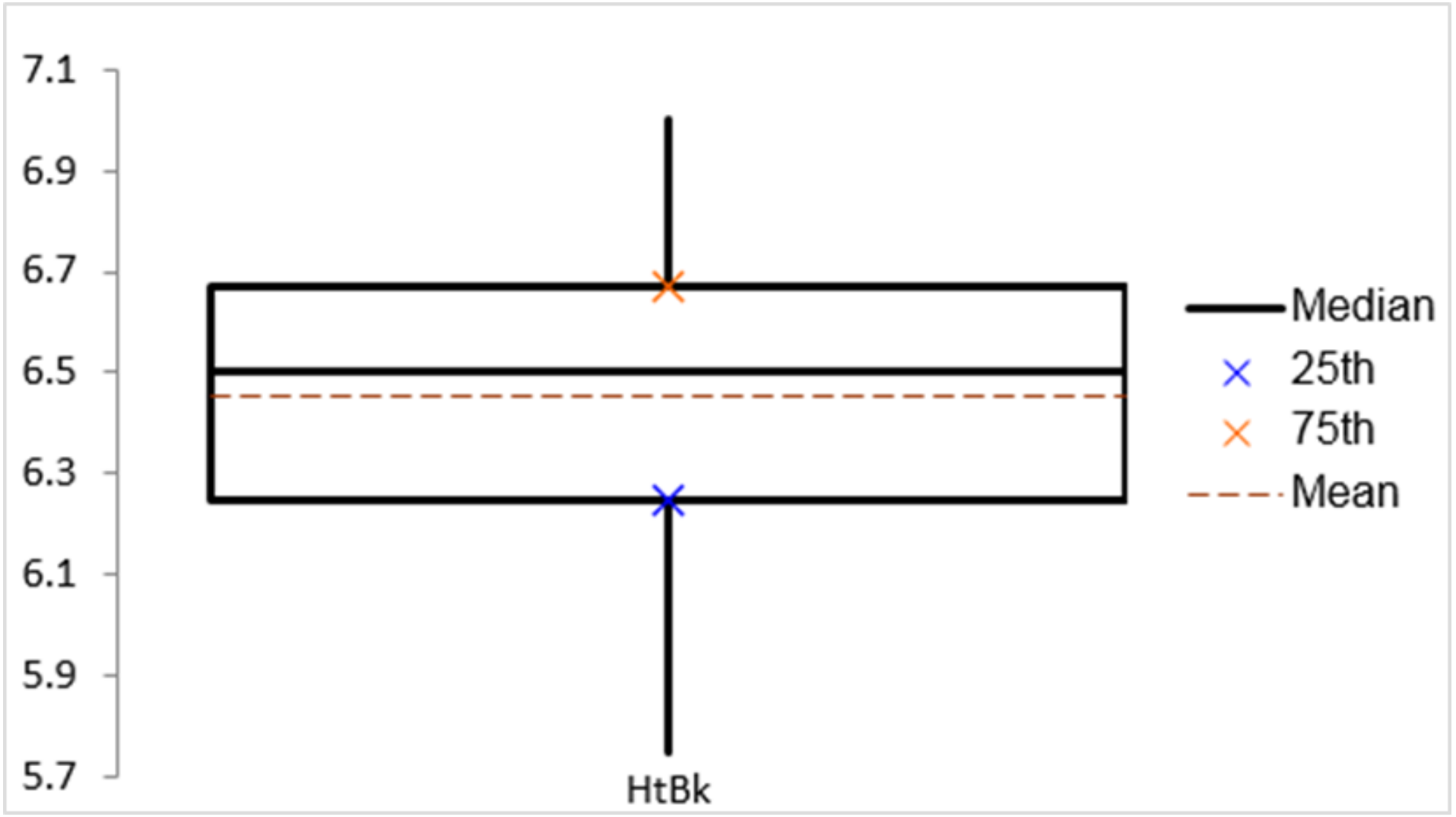
How to Use SigmaXL to Generate a Box Plot

- Data File: “Box Plot” tab in “Sample Data.xlsx”
- Steps to render a Box Plot in SigmaXL
  - Select the entire range of the data
  - Click SigmaXL -> Graphical Tools -> Boxplots
  - A new window named “Boxplots” pops up with the selected range appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window also named “Boxplots” appears
  - Select “HtBk” as the “Numeric Data Variables (Y)”
  - Check the check box “Show Legend”
  - Click “OK>>”
  - The Boxplot appears automatically in the new tab “Boxplot (1)”

How to Use SigmaXL to Generate a Box Plot



How to Use SigmaXL to Generate a Box Plot



Histogram

- A **histogram** is a graphical tool to present the distribution of the data.
- The X axis represents the possible values of the variable and the Y axis represents the frequency of the value occurring.
- A histogram consists of adjacent rectangles erected over intervals with heights equal to the frequency density of the interval.
- The total area of all the rectangles in a histogram is the number of data values.

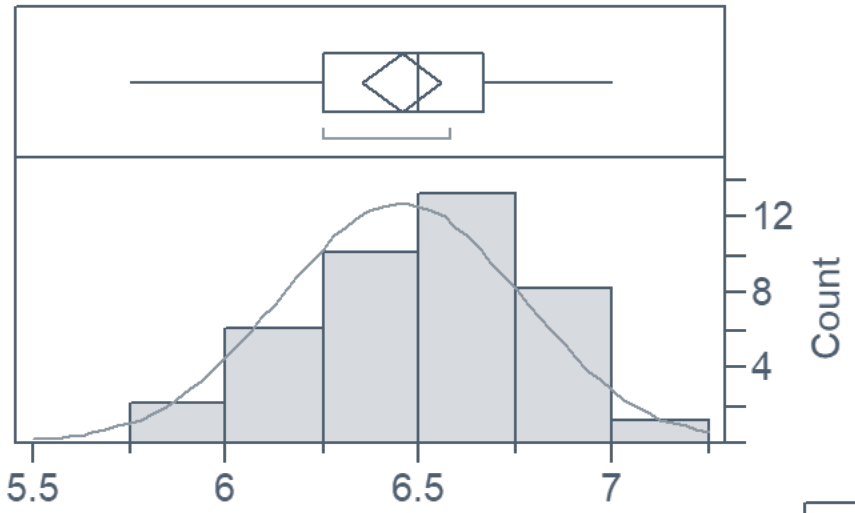
Histogram

- A histogram can also be normalized. In this case, the X axis still represents the possible values of the variable, but the Y axis represents the percentage of observations that fall into each interval on the X axis.

- The total area of all the rectangles in a normalized histogram is 1.

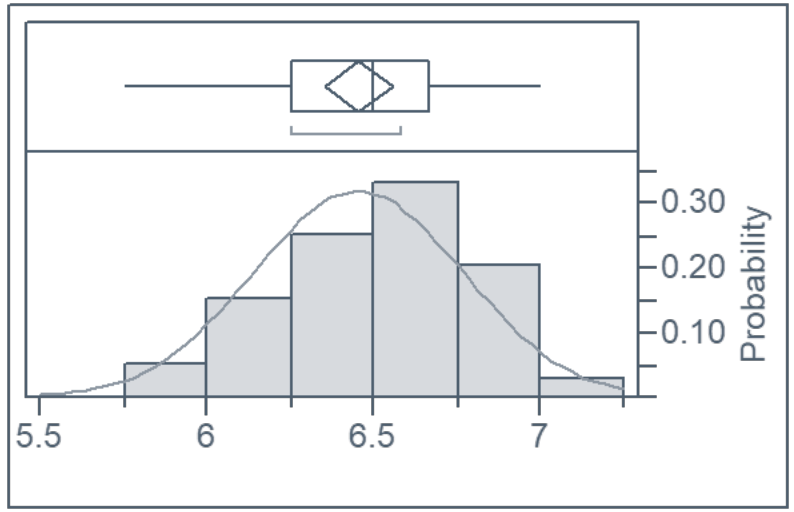
- With the histogram, we have a better understanding of the shape, location, and spread of the data.

Histogram



Histogram with frequency (count) as the Y axis

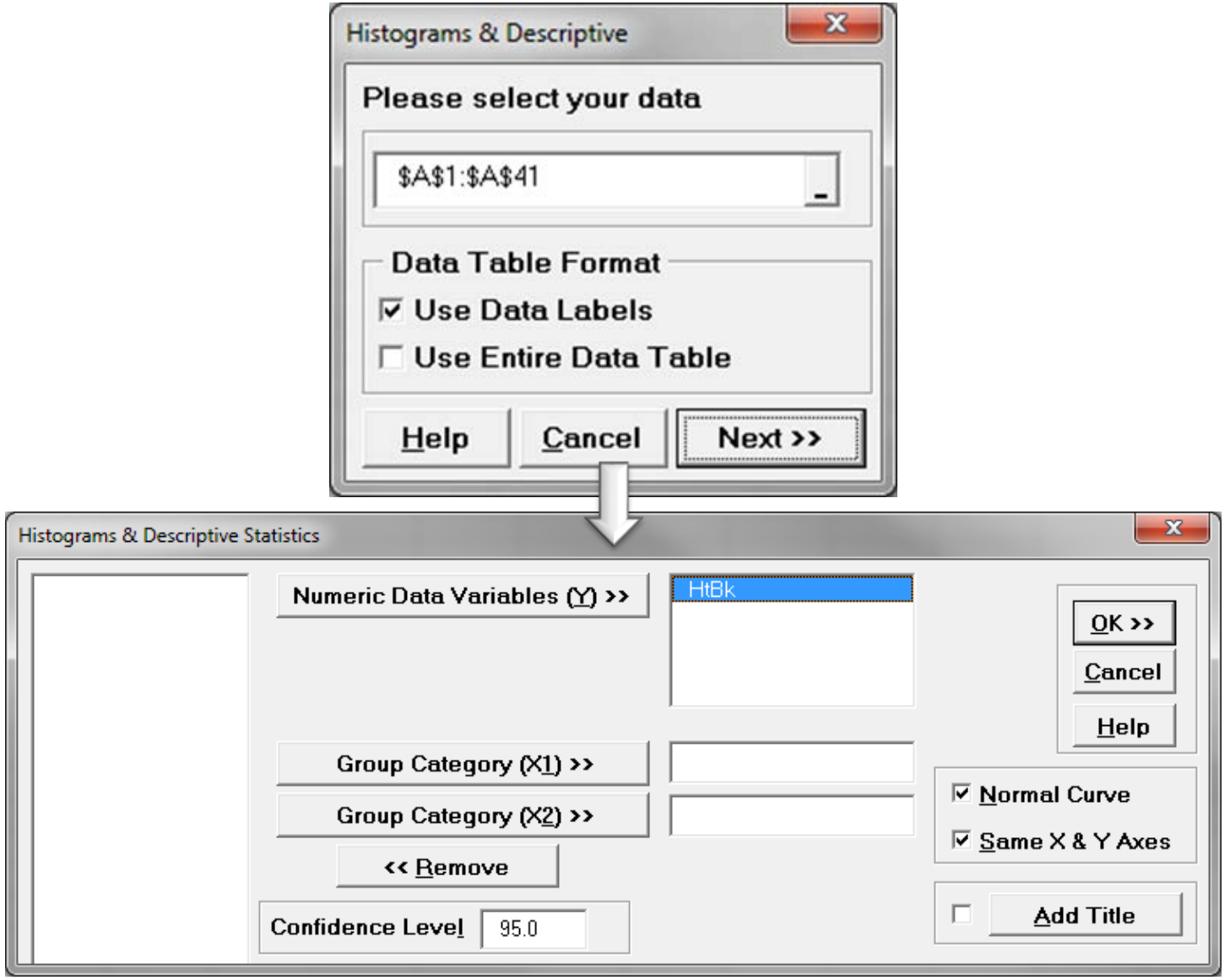
Normalized histogram with proportion (probability) as the Y axis



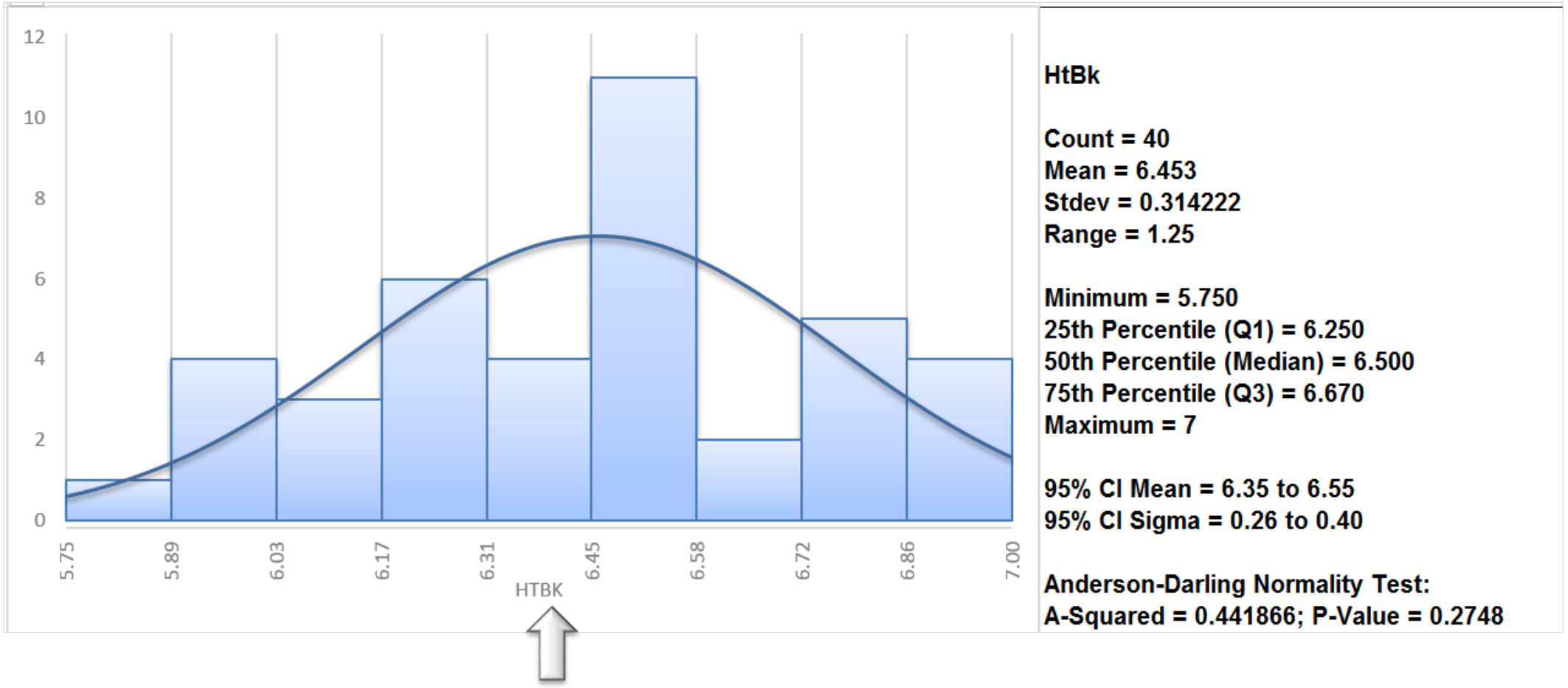
How to Use SigmaXL to Generate a Histogram

- Data File: “Histogram” tab in “Sample Data.xlsx”
- Steps to render a histogram in SigmaXL
  - Select the entire range of data
  - Click SigmaXL -> Graphical Tools
    -> Histograms & Descriptive Statistics
  - A new window named “Histograms & Descriptive” pops up with the selected range of data appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Histograms & Descriptive Statistics” appears
  - Select “HtBk” as the “Numeric Data Variables (Y)”
  - Click “OK>>”
  - The histogram appears in the new tab “Hist Descript (1)”

How to Use SigmaXL to Generate a Histogram



How to Use SigmaXL to Generate a Histogram



Histogram

Scatter Plot

- A **scatter plot** is a diagram to present the relationship between two variables of a data set.

- A scatter plot consists of a set of data points.

- On the scatter plot, a single observation is presented by a data point with its horizontal position equal to the value of one variable and its vertical position equal to the value of the other variable.

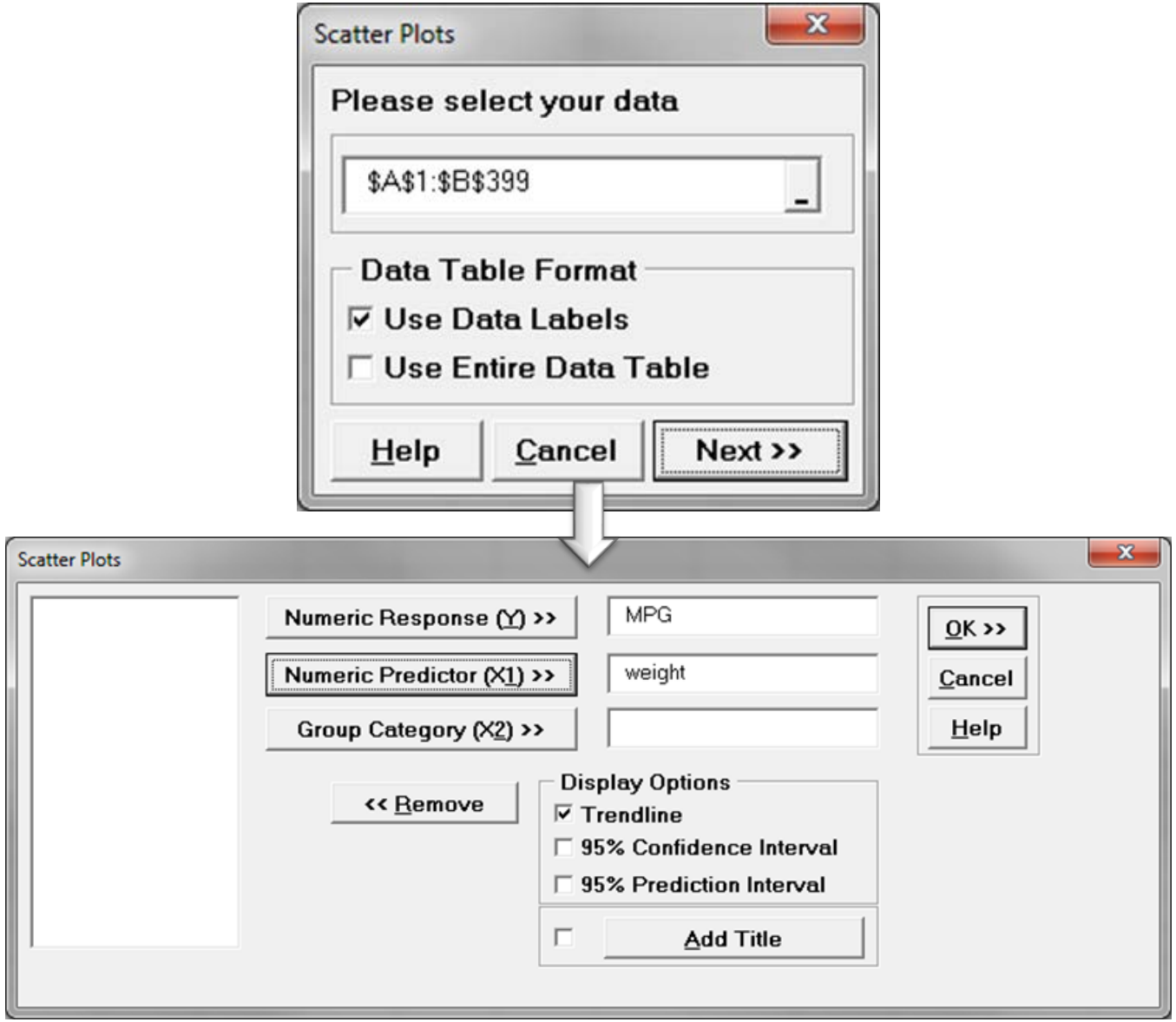
Scatter Plot

- A scatter plot helps to understand:
  - Whether the two variables are related to each other or not
  - How is the strength of their relationship
  - What is the shape of their relationship
  - What is the direction of their relationship
  - Whether outliers are present.

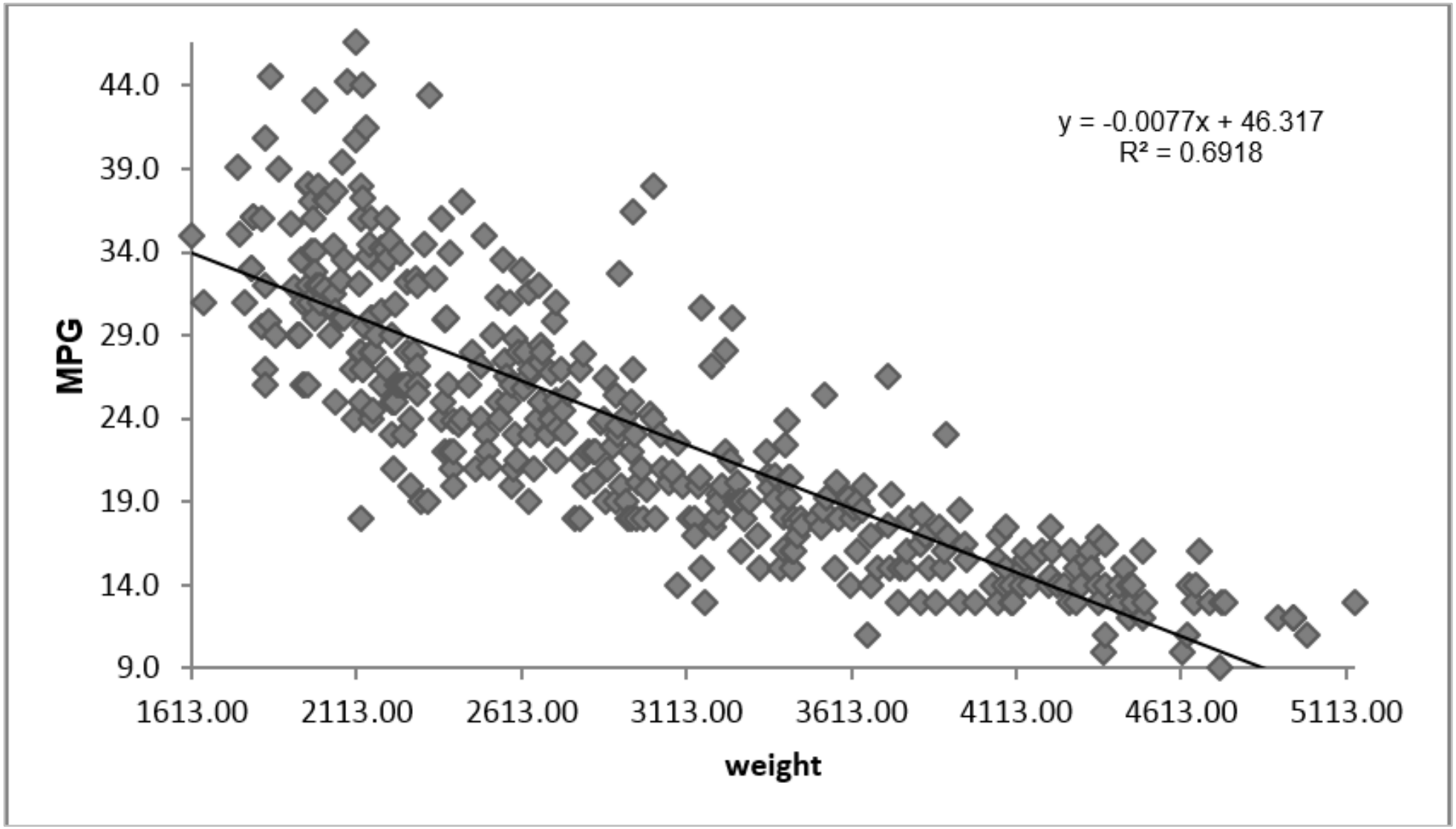
How to Use SigmaXL to Generate a Scatter Plot

- Data File: “Scatter Plot” tab in “Sample Data.xlsx”
- Steps to render a histogram in SigmaXL
  - Select the entire range of data (both “MPG” and “weight”)
  - Click SigmaXL -> Graphical Tools -> Scatter Plots
  - A new window named “Scatter Plots” pops up with the selected range of data appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window also named “Scatter Plots” appears
  - Select “MPG” as the “Numeric Response (Y)”
  - Select “weight” as the “Numeric Predictor (X1)”
  - Click “OK>>”
  - The scatterplot appears in the new tab “Scatterplot (1)”

How to Use SigmaXL to Generate a Scatter Plot



How to Use SigmaXL to Generate a Scatter Plot



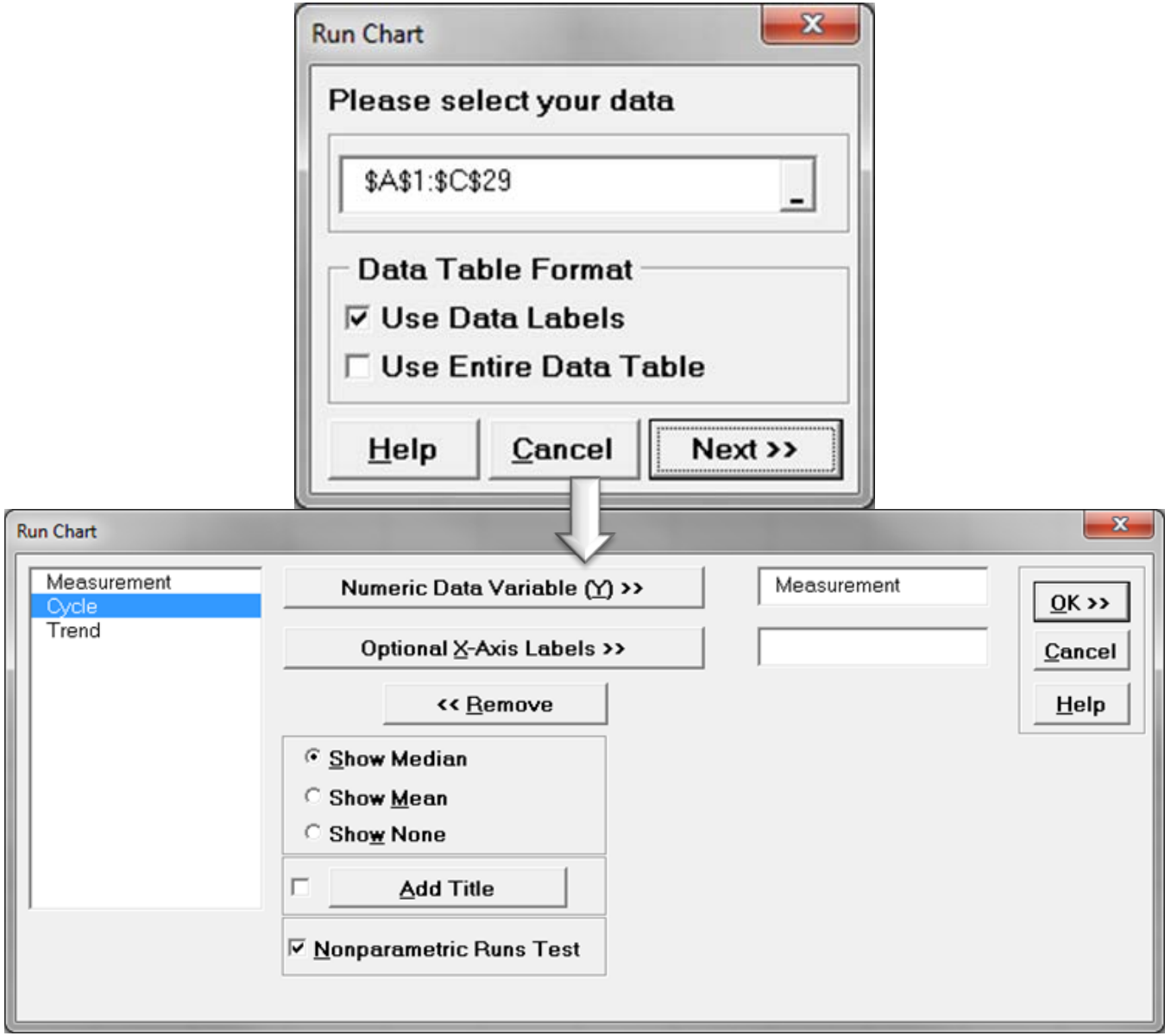
Run Chart

- A **run chart** is a chart used to present the data in time order. It captures the process performance over time.
- The X axis of the run chart indicates the time and the Y axis indicates the observed values.
- Run chart looks similar to control charts except that a run chart does not have control limits plotted. It is easier to produce a run chart than a control chart.
- It is often used to identify the anomalies in the data and discover the pattern of data changing over time.

How to Plot a Run Chart in SigmaXL

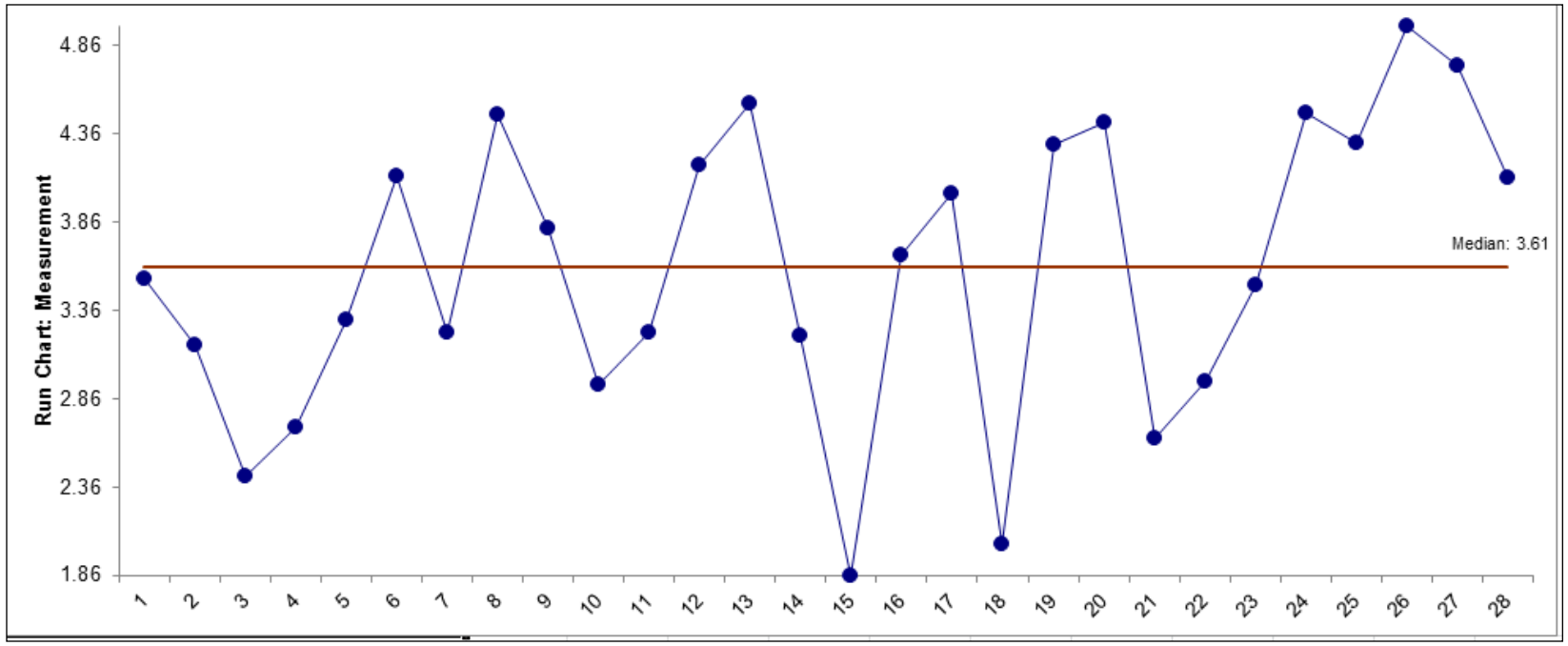
- Steps to plot a run chart in SigmaXL:
  - Data File: “Run Chart” tab in “Sample Data.xlsx”
  - Select the entire range of the data (“Measurement”, “Cycle” and “Trend”). In this first example, let’s select the data in column “Measurement’. We will use the other two columns later.
  - Click SigmaXL -> Graphical Tools -> Run Chart
  - A new window named “Run Chart” pops up with the selected range of data appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window also named “Run Chart” appears
  - Select “Measurement” as the “Numeric Data Variable (Y)”
  - Click “OK”
  - The run chart appears automatically in the tab “Run Chart (1)”

How to Plot a Run Chart in SigmaXL



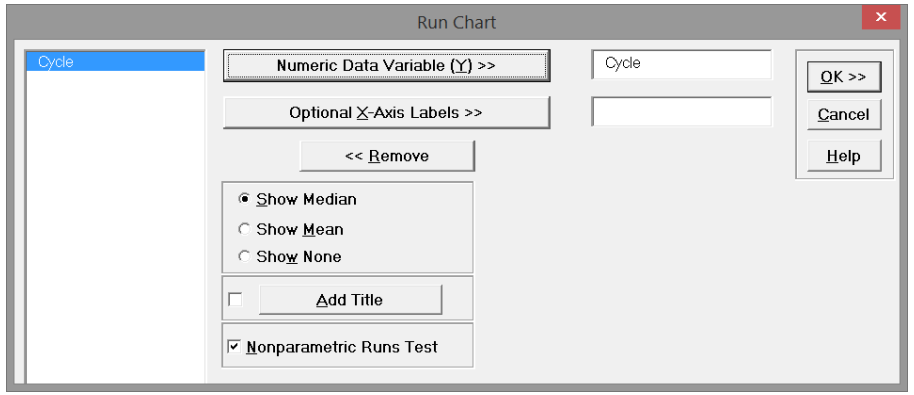
Run Chart Example

- Run chart is used to identify the trend, cycle, seasonal pattern, abnormality in the data.
  - The time series in this chart appear stable.
  - There are no extreme outliers, trending or seasonal patterns.

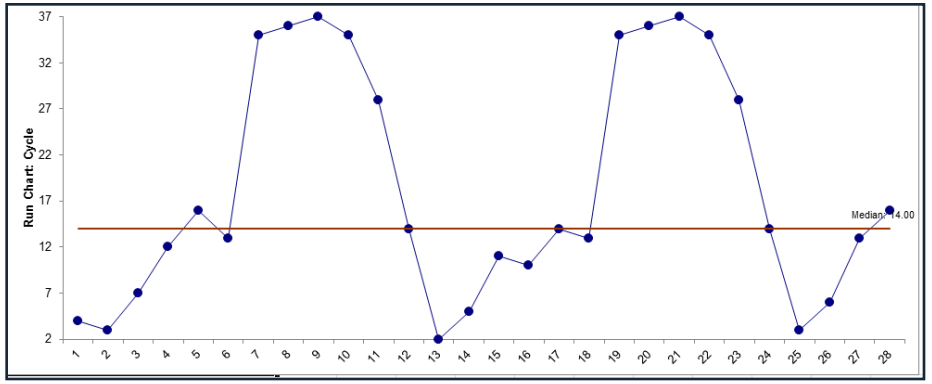


Run Chart Example

- Create another run chart using the data listed in column “Cycle” in the “Run Chart” tab of “Sample Data.xlsx”.
  - In this example, the data clearly is exhibiting a pattern. It could be something that is “seasonal”, or could be something cyclical in a process.
  - Imagine that the data points are monthly, and this is showing us a process performing over the period of 2.5 years. Perhaps this represent the number of customers buying new homes. The home buying market tends to peak in the summer months and dies down in the winter.

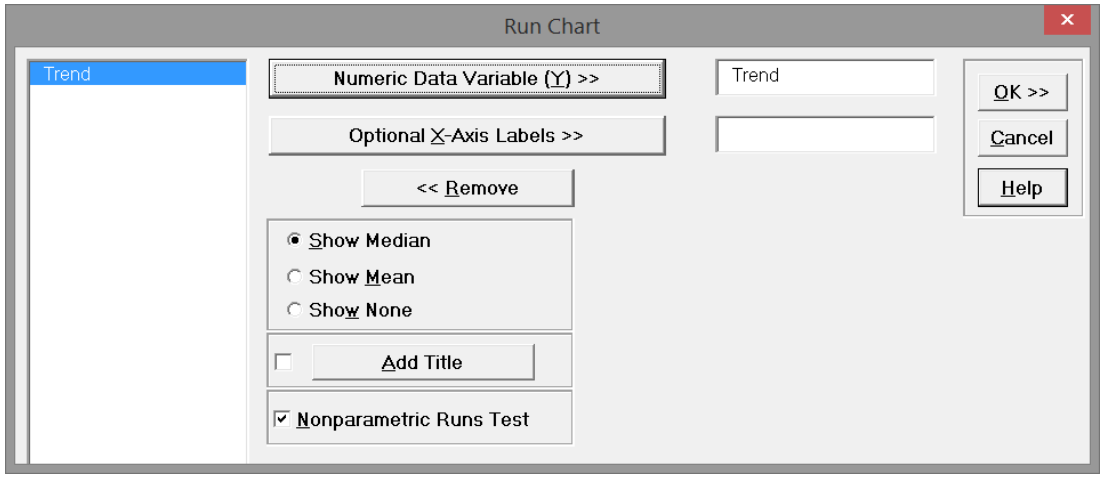


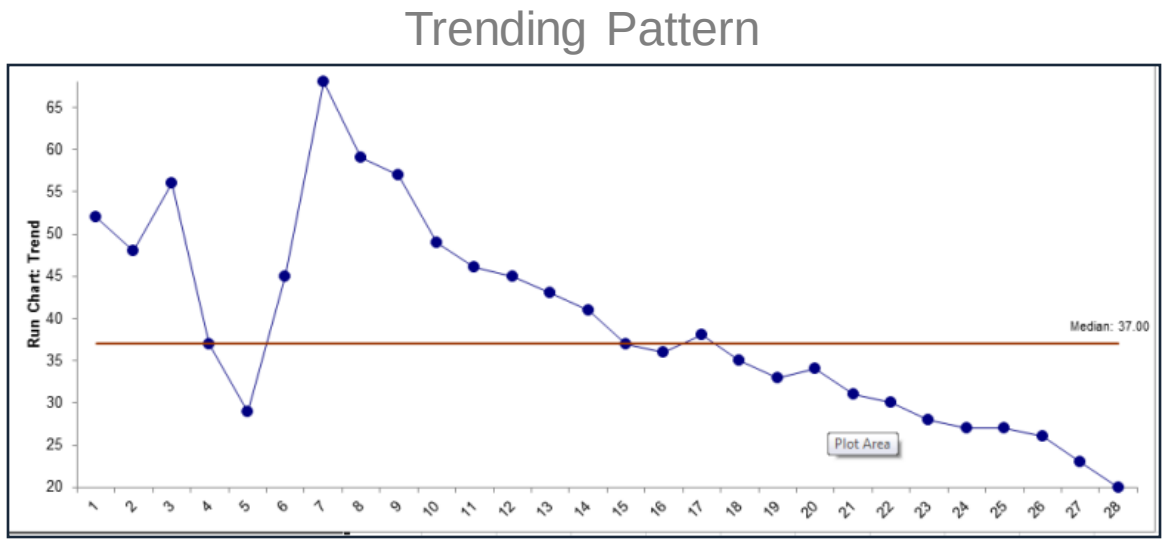
Seasonal Pattern



Run Chart Example

- Create another run chart using the data listed in column “Trend” in the “Run Chart” tab of “Sample Data.xlsx”.





2.3 MSA (Measurement System Analysis)

Black Belt Training: Measure Phase

2.1 Process Definition
- 2.1.1 Cause and Effect Diagrams
- 2.1.2 Cause and Effects Matrix
- 2.1.3 Process Mapping
- 2.1.4 FMEA: Failure Modes and Effects Analysis
- 2.1.5 Theory of Constraints

2.2 Six Sigma Statistics
- 2.2.1 Basic Statistics
- 2.2.2 Descriptive Statistics
- 2.2.3 Distributions and Normality
- 2.2.4 Graphical Analysis

2.3 Measurement System Analysis
- 2.3.1 Precision and Accuracy
- 2.3.2 Bias, Linearity, and Stability
- 2.3.3 Gage R&R
- 2.3.4 Variable and Attribute MSA

2.4 Process Capability
- 2.4.1 Capability Analysis
- 2.4.2 Concept of Stability
- 2.4.3 Attribute and Discrete Capability
- 2.4.4 Monitoring Techniques

2.3.1 Precision and Accuracy

What is Measurement System Analysis

- **Measurement System Analysis** (MSA) is a systematic method to identify and analyze the variation components in the measurement.

- It is a mandatory step in any Six Sigma project to ensure the data are reliable before making any data-based decisions.

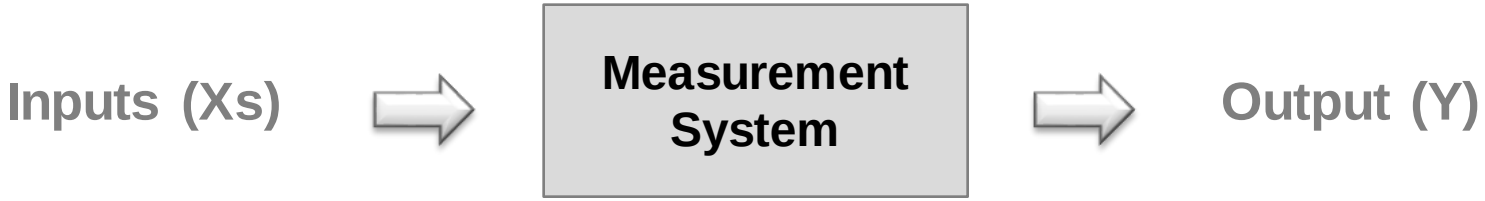
- The MSA is the check point of data quality before we start any further analysis and draw any conclusions from the data.

Data-Based Analysis

- Here are some examples of data-based analysis where MSA is the prerequisite:
  - Correlation analysis
  - Regression analysis
  - Hypothesis testing
  - Analysis of variance
  - Design of experiments
  - Multivariate analysis
  - Statistical process control.

Measurement System

- A measurement system is a process to obtain data.



- Y (output of the measurement system)
  - Observed values

- X's (inputs of the measurement system)
  - True values
  - Measurement errors

Measurement Errors

Observed Value = True Value + Measurement Error

- **True Value**
  - The actual value we are interested to measure
  - It reflects the true performance of the process we are measuring
- **Measurement Error**
  - The errors brought in by measurement system
- **Observed Value**
  - The observed/measured value obtained by the measurement system

Measurement Errors

- Types of Observed Values:
  - Continuous measurements
    - Weight
    - Height
    - Money
  - Discrete measurements
    - Red/Yellow/Green
    - Yes/No
    - Ratings of 1–10

- A *variable MSA* is designed for continuous measurements and an *attribute MSA* is for discrete measurements.

Measurement Errors

- Sources of measurement errors:
  - Human
  - Environment
  - Equipment
  - Sample
  - Process
  - Material
  - Method.

- Fishbone diagrams can help to brainstorm the potential factors affecting the measurement system.

Measurement Errors

- The more errors the measurement system brings in, the less reliable the observed values are.
- A valid measurement system brings in minimum amount of measurement errors.
- The goal of MSA is to qualify the measurement system by quantitatively analyzing its characteristics.

Characteristics of a Measurement System

- Any measurement systems can be characterized by two aspects:
  - Accuracy (location related)
  - Precision (variation related).

- A valid measurement system is *both* accurate and precise.
  - Being accurate does not guarantee the measurement system is precise.
  - Being precise does not guarantee the measurement system is accurate.

Accuracy vs. Precision

- Accuracy:
  - The level of closeness between the average observed value and the true value
  - How well the observed value reflects the true value.

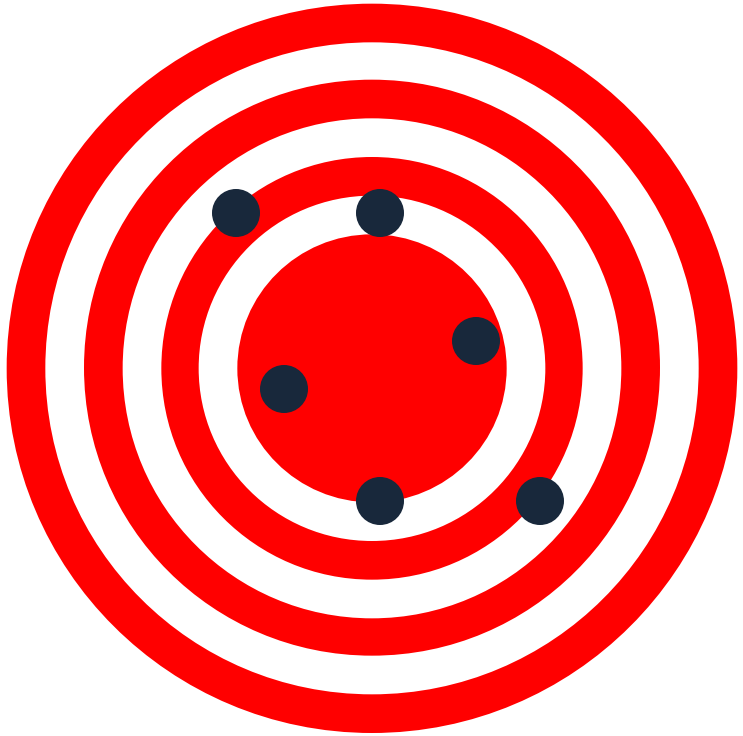
- Precision:
  - The spread of measurement values
  - How consistent the repeated measurements deliver the same values under the same circumstances.

Accuracy vs. Precision

- Accurate and precise
  - high accuracy and high precision

Accuracy vs. Precision

- Accurate and not precise
  - high accuracy and low precision



Accuracy vs. Precision

- Precise and not accurate
  - high precision and low accuracy

Accuracy vs. Precision

- Not accurate and not precise
  - low accuracy and low precision

MSA Conclusions

- If the measurement system is considered *both* accurate and precise, we can start the data-based analysis or decision making.

- If the measurement system is either not accurate or not precise, we need to identify the factor(s) affecting it and calibrate the measurement system until it is both accurate and precise.

Stratification of Accuracy and Precision

- Accuracy
  - Bias
  - Linearity
  - Stability

- Precision
  - Repeatability
  - Reproducibility

2.3.2 Bias, Linearity and Stability

Bias

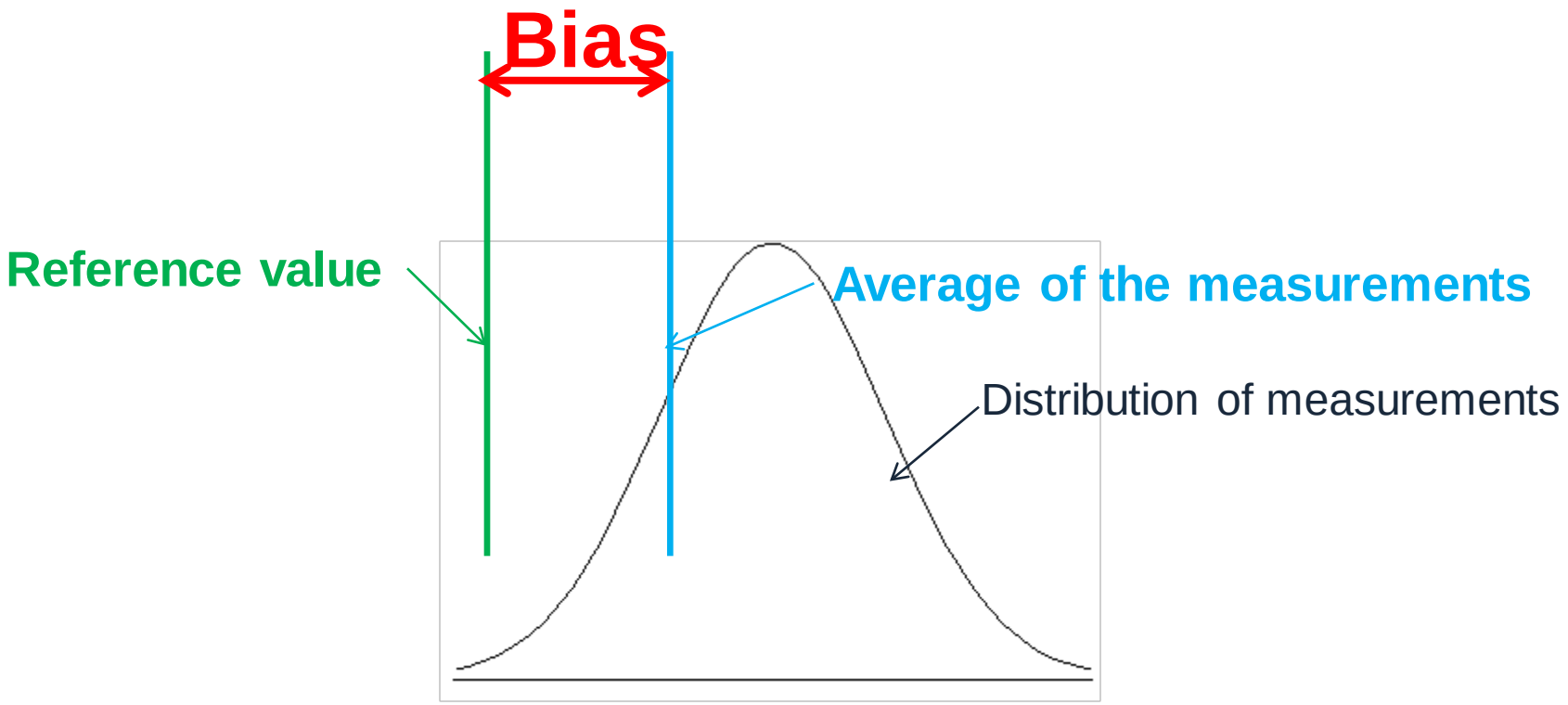
- **Bias** is the difference between the observed value and the true value of a parameter or metric being measured.

- It is calculated by subtracting the reference value from the average value of the measurements.

Bias = Grand Mean – Reference Value

where the reference value is a standard agreed upon

Bias



- The closer the average of all measurements is to the reference level, the smaller the bias.
- The reference level is the average of measurements of the same items using the master or standard instrument.

Bias

- To determine whether the difference between the average of observed measurement and the reference value is statistically significant (we will explain more details about statistical significance in the Analyze module), we can either conduct a *hypothesis testing* or *compare* the reference value against the confidence intervals of the average measurements.

- If the reference value falls into the confidence intervals, the bias is not statistically significant and can be ignored. Otherwise, the bias is statistically significant and must be fixed.

Bias

- Potential causes of bias:
  - Errors in measuring the reference value
  - Lack of proper training for appraisers
  - Damaged equipment or instrument
  - Measurement instrument not calibrated precisely
  - Appraisers read the data incorrectly.

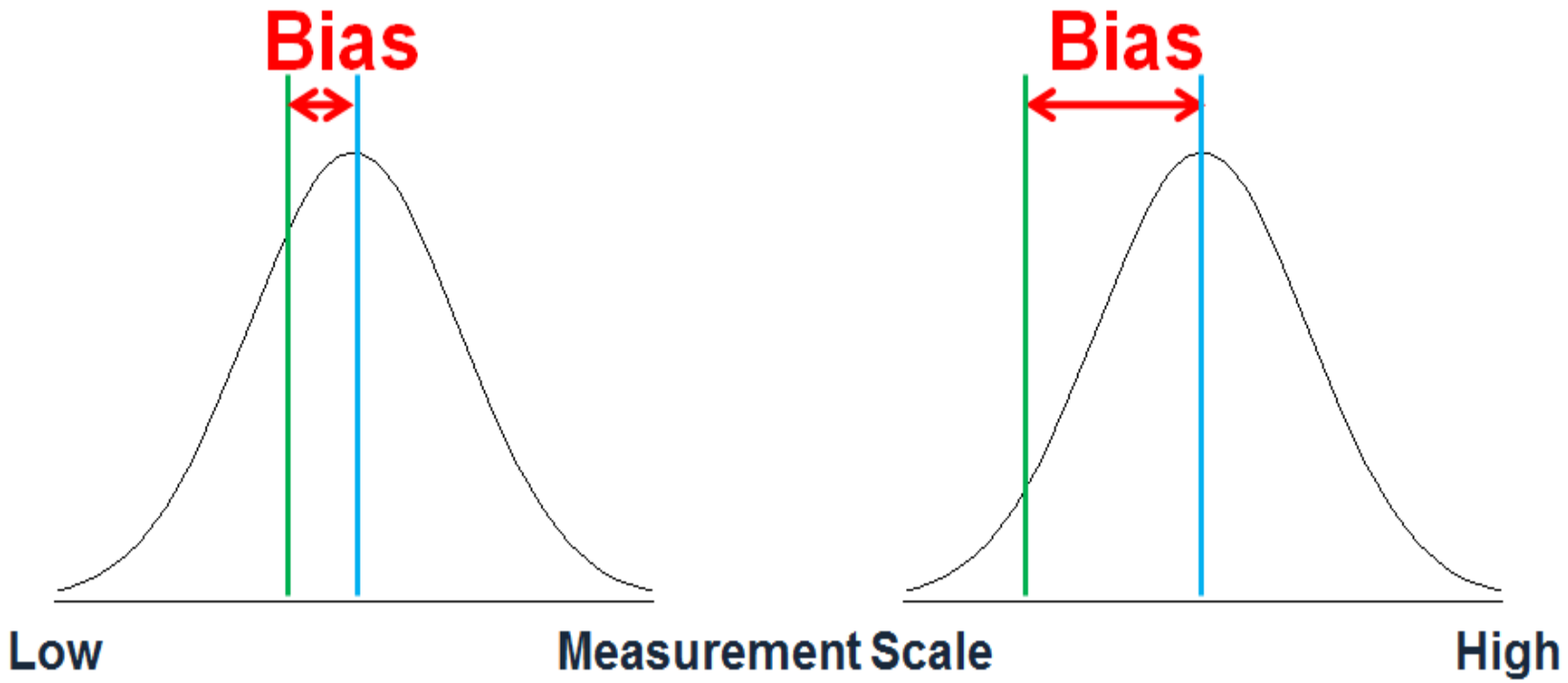
Linearity

- **Linearity** is the degree of the consistency of bias over the entire expected measurement range.

- It quantifies how the bias changes over the range of measurement.

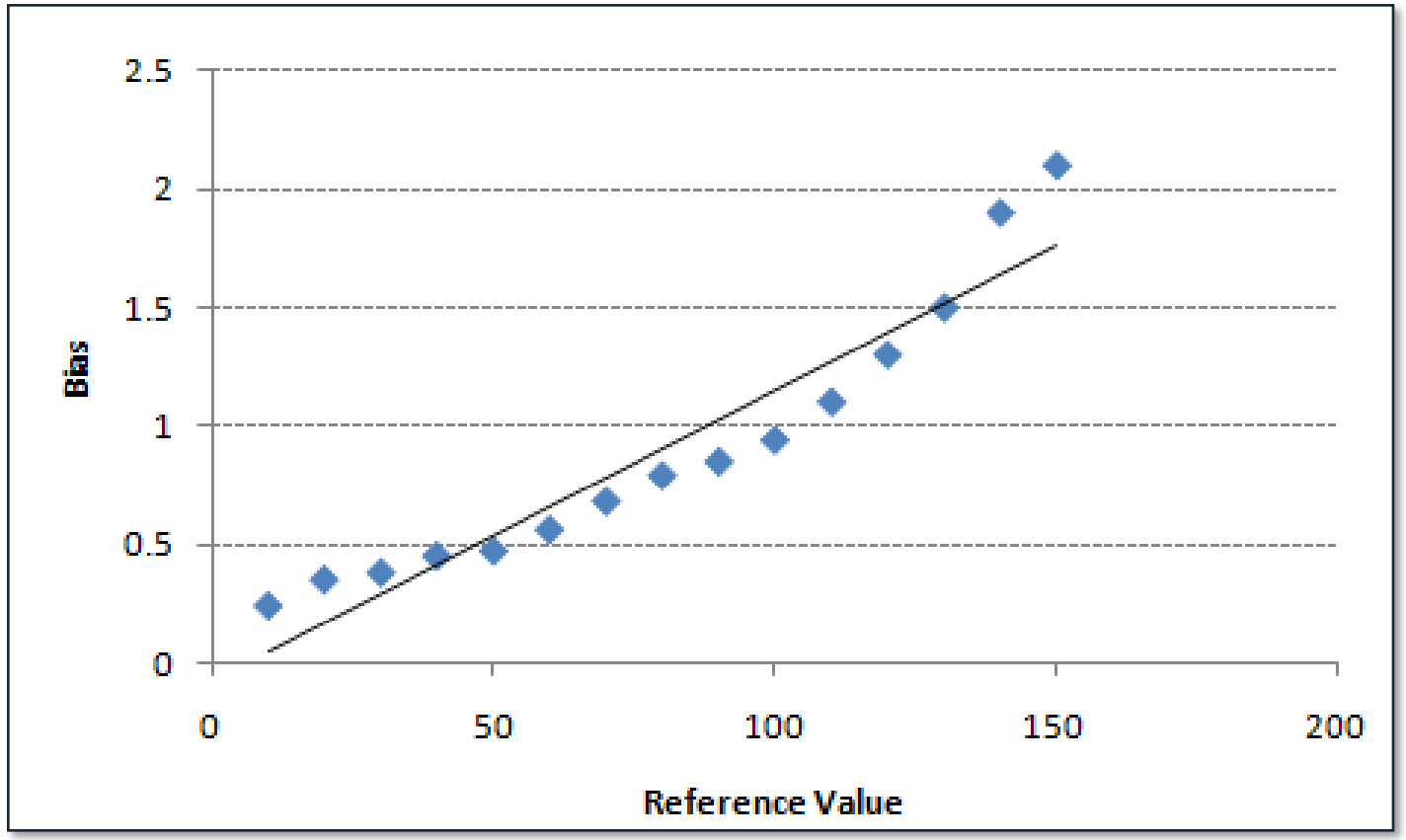
- For example, a scale is off by 0.01 pounds when measuring an object of 10 pounds. However, it is off by 10 pounds when measuring an object of 100 pounds. The scale's bias is not linear.

Linearity



Linearity

- Create a scatter plot for bias (Y-axis) and reference level (X-axis).
- Find a best fit linear regression line and compute the slope of the line.
- The closer the slope is to zero, the better the measurement system performs.



Linearity

- Formula of the linearity of a measurement system:

$$Linearity = |Slope| \times Process\ Variation$$

where

$$Slope = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \left(x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

x_i is the reference value;
y_i is the bias at each reference level;
n is the sample size.

Linearity

- Potential causes of linearity:
  - Errors in measuring the lower end or higher end of the reference value
  - Lack of proper training for appraisers
  - Damaged equipment or instrument
  - Measurement instrument not calibrated correctly at the lower or higher end of the measurement scale
  - Innate nature of the instrument.

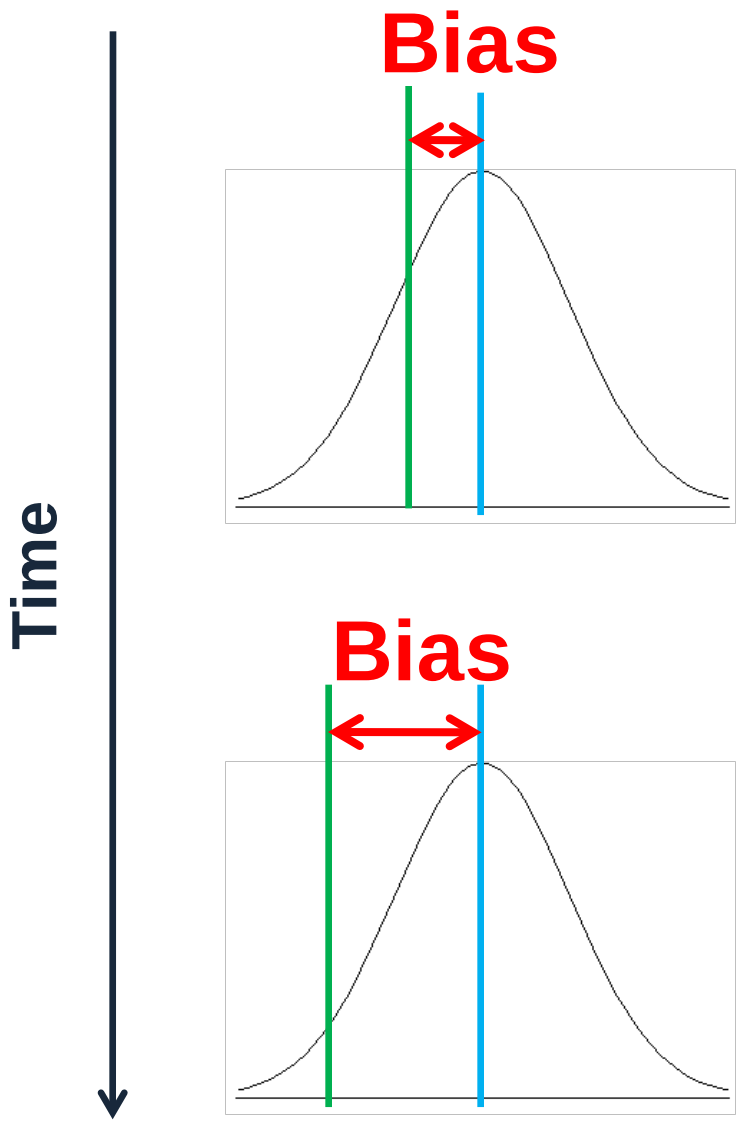
Stability

- **Stability** is the consistency level needed to obtain the same values when measuring the same objects over an extended period of time.

- A measurement system that has low bias and linearity close to zero but cannot consistently perform well would not deliver reliable data.

- Stability is evaluated using control charts.

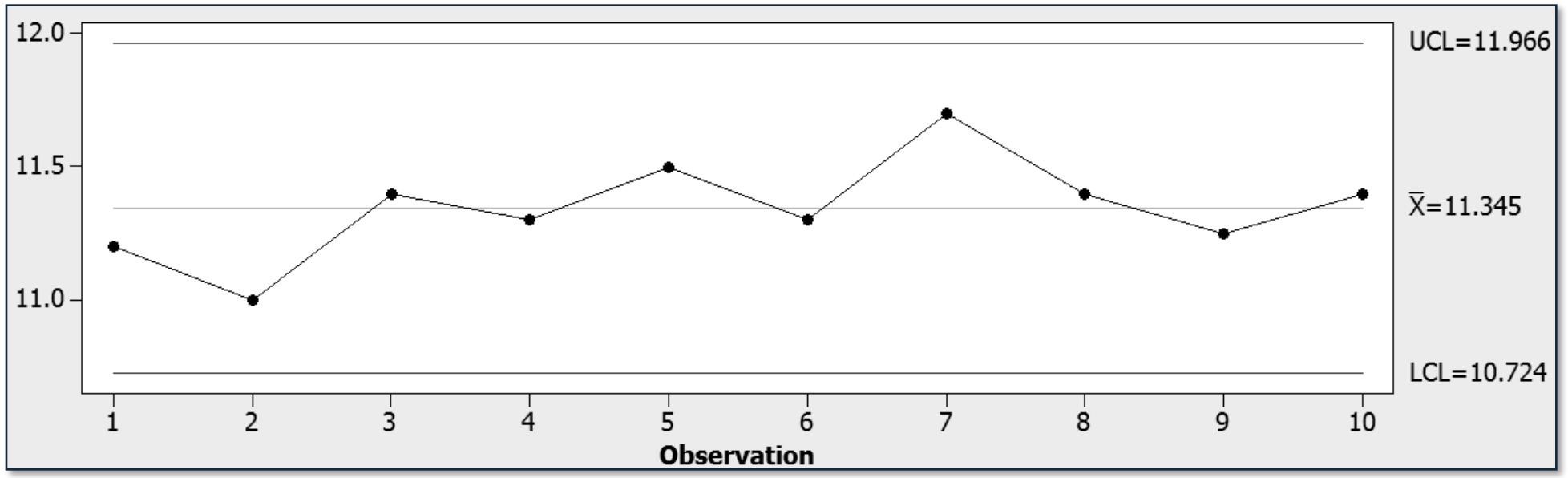
Stability



Stability

- Control charts are used to evaluate the stability of a measurement system.

- When there are no data points out of control, the measurement system is considered stable.



Stability

- Potential causes of instability:
  - Inconsistent training for appraisers
  - Damaged equipment or instrument
  - Worn equipment or instrument
  - Measurement instrument not calibrated
  - Appraisers do not follow the procedure consistently.

2.3.3 Gage R&R

Repeatability

- **Repeatability** evaluates whether the same appraiser can obtain the same value multiple times when measuring the same object using the same equipment under the same environment.

- It refers to the level of agreement between the repeated measurements of the same appraiser under the same condition.

- Repeatability measures the inherent variation of the measurement instrument.

Reproducibility

- **Reproducibility** evaluates whether different appraisers can obtain the same value when measuring the same object independently.
- It refers to the level of agreement between different appraisers.
- It is not caused by the inherent variation of the measurement instrument. It reflects the variability caused by different appraisers, locations, gauges, environments etc.

Gauge R&R

- Gauge R&R (i.e. Gauge Repeatability & Reproducibility) is a method to analyze the variability of a measurement system by partitioning the variation of the measurements using ANOVA (Analysis of Variance).

- Gauge R&R only addresses the precision of a measurement system.

Gauge R&R

- Data collection of a gauge R&R study:
  - let *k* appraisers measure *n* random samples independently and repeat the process *p* times.

- Different appraisers perform the measurement independently.

- The order of measurement (e.g. sequence of samples and sequence of appraisers) is randomized.

Gauge R&R

- The potential sources of variance in the measurement:
  - Appraisers: $\sigma^2_{appraisers}$
  - Parts: σ^2_{parts}
  - Appraisers × Parts: $\sigma^2_{appraisers \times parts}$
  - Repeatability: $\sigma^2_{repeatability}$

- Variance Components

$$\sigma^2_{total} = \sigma^2_{appraisers} + \sigma^2_{parts} + \sigma^2_{appraisers \times parts} + \sigma^2_{repeatability}$$

Gauge R&R

- A valid measurement system has low variability in both repeatability and reproducibility so that the total variability observed can reflect the true variability in the objects (parts) being measured.

$$\sigma^2_{total} = \sigma^2_{reproducibility} + \sigma^2_{repeatability} + \sigma^2_{parts}$$

where

$$\sigma^2_{reproducibility} = \sigma^2_{appraisers} + \sigma^2_{appraisers \times parts}$$

- Gauge R&R variance reflects the precision level of the measurement system.

$$\sigma^2_{R\&R} = \sigma^2_{repeatability} + \sigma^2_{reproducibility}$$

Gauge R&R

- Variation Components

$$Variation_{total} = Z_0 \times \sigma_{total}$$

$$Variation_{repeatability} = Z_0 \times \sigma_{repeatability}$$

$$Variation_{reproducibility} = Z_0 \times \sigma_{reproducibility}$$

$$Variation_{parts} = Z_0 \times \sigma_{parts}$$

where

$$\sigma^2_{total} = \sigma^2_{reproducibility} + \sigma^2_{repeatability} + \sigma^2_{parts}$$

Z_0 is a sigma multiplier that assumes a specific confidence level in the spread of the data.

Gauge R&R

- The percentage of variation R&R contributes to the total variation in the measurement:

$$Contribution\%_{R\&R} = \frac{Variation_{R\&R}}{Variation_{total}} \times 100\%$$

where $Variation_{R\&R} = Z_0 \times \sqrt{\sigma^2_{repeatability} + \sigma^2_{reproducibility}}$

Measurement System	% Study Var	% Contribution	Distinct Categories
Acceptable	10% or less	1% or Less	5 or Greater
Marginal	10% - 30%	1% - 9%	
Unacceptable	30% or Greater	9% or Greater	Less than 5

2.3.4 Variable and Attribute MSA

Variable Gage R&R

- Whenever something is measured repeatedly or by different people or processes, the results of the measurements will vary. Variation comes from two primary sources:
  1. Differences between the parts being measured
  2. The measurement system.
- We can use a gage R&R to conduct a measurement system analysis to determine what portion of the variability comes from the parts and what portion comes from the measurement system.
- There are key study results that help us determine the components of variation within our measurement system.

Key Measures of a Variable Gage R&R

- %Contribution: The percent of contribution for a source is 100 times the variance component for that source divided by the total variation.

- %Study Var (6*SD): The percent of study variation for a source is 100 times the study variation for that source divided by the total variation.

- %Tolerance (SV/Tolerance): The percent of spec range taken up by the total width of the distribution of the data based on variation from that source.

- Distinct Categories: The number of distinct categories of parts that the measurement system is able to distinguish. If a measurement system is not capable of distinguishing at least five types of parts, it is probably not adequate.

Variable Gage R&R Guidelines (AIAG)

- **Percent Tolerance and Percent Study Variation**
  - 10% or less – Acceptable
  - 10% to 30% – Marginal
  - 30% or greater – Unacceptable

- **Percent Contribution**
  - 1% or less – Acceptable
  - 1% to 9% – Marginal
  - 9% or greater – Unacceptable

- **Distinct Categories**
  - Look for five or more distinct categories to indicate that your measurement system is acceptable.

Guidelines for Distinct Categories

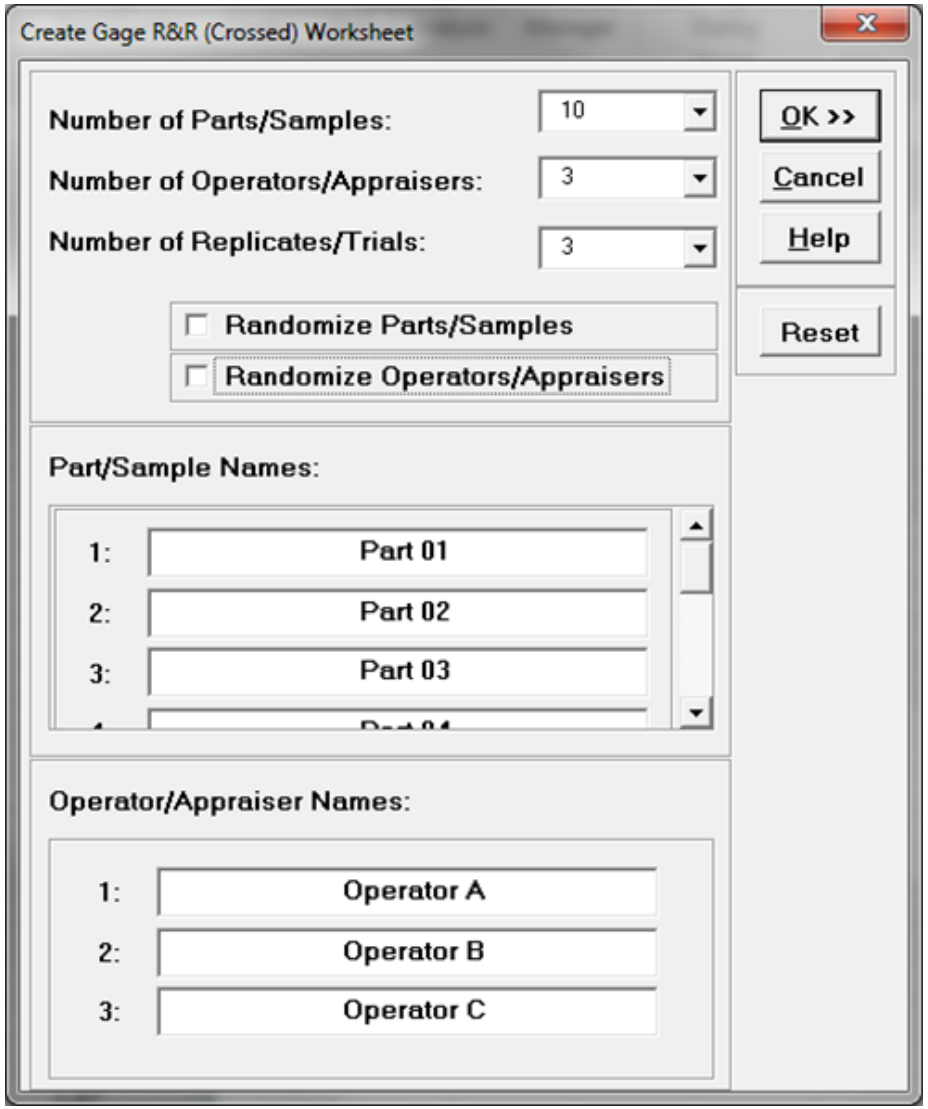
- Distinct categories is the number of categories of parts that your measurement system can distinguish. If it is below five, it is likely not able to distinguish between parts.

Number of Categories	Conclusion
Distinct Categories = 1	Measurement system cannot discriminate between parts
Distinct Categories = 2	Measurement system can only distinguish between high/low or big/small
Distinct Categories = 3 or 4	Measurement system is of little or no value
Distinct Categories = 5+	According to AIAG, the measurement system can acceptably discriminate parts

Use SigmaXL to Implement a Variable MSA

- Data File: “Variable MSA” tab in “Sample Data.xlsx” (an example in the AIAG MSA Reference Manual, 3rd Edition)
- Step 1: Initiate the MSA study
  - Click on SigmaXL -> Measurement Systems Analysis -> Create Gauge R&R (Crossed) Worksheet
  - A new window named “Create Gauge R&R (Crossed) Worksheet” appears
  - Enter 10 as the “Number of Parts/Samples”
  - Enter 3 as the “Number of Operators/Appraisers”
  - Enter 3 as the “Number of Replicates/Trials”
  - Uncheck the checkboxes for both “Randomize Parts/Sample” and “Randomize Operators/Appraisers”
  - Click “OK>>”
  - A new tab named “Gage R&R (Crossed) WKS” is generated.

Use SigmaXL to Implement a Variable MSA



Use SigmaXL to Implement a Variable MSA

Gage R&R Study (Crossed) Worksheet

Gage Name:	
Date of Study:	
Performed By:	
Notes:	

Run Order	Std. Order	Part	Operator	Measurement
1	1	Part 01	Operator A	
2	2	Part 01	Operator A	
3	3	Part 01	Operator A	
4	4	Part 02	Operator A	
5	5	Part 02	Operator A	
6	6	Part 02	Operator A	
7	7	Part 03	Operator A	
8	8	Part 03	Operator A	
9	9	Part 03	Operator A	
10	10	Part 04	Operator A	
11	11	Part 04	Operator A	
12	12	Part 04	Operator A	
13	13	Part 05	Operator A	
14	14	Part 05	Operator A	
15	15	Part 05	Operator A	
16	16	Part 06	Operator A	
17	17	Part 06	Operator A	
18	18	Part 06	Operator A	

Use SigmaXL to Implement a Variable MSA

- Step 2: Data collection
  - In the newly generated tab “Gage R&R (Crossed) WKS”, SigmaXL has provided the template which we organize the data
  - In the “Variable MSA” tab in “Sample Data.xlsx”, there are all the measurement data collected by three operators (i.e. operator A, B and C). The data are listed in the same standardized order as the tab “Gage R&R (Crossed) WKS”.

Run Order	Part	Operator	Measurement
1	Part 01	Operator A	0.29
2	Part 01	Operator A	0.41
3	Part 01	Operator A	0.64
4	Part 02	Operator A	-0.56
5	Part 02	Operator A	-0.68
6	Part 02	Operator A	-0.58
7	Part 03	Operator A	1.34
8	Part 03	Operator A	1.17
9	Part 03	Operator A	1.27
10	Part 04	Operator A	0.47
11	Part 04	Operator A	0.5
12	Part 04	Operator A	0.64
13	Part 05	Operator A	-0.8
14	Part 05	Operator A	-0.92
15	Part 05	Operator A	-0.84
16	Part 06	Operator A	0.02
17	Part 06	Operator A	-0.11
18	Part 06	Operator A	-0.21

Use SigmaXL to Implement a Variable MSA

- Step 3: Enter the data into the tab “Gage R&R (Crossed) WKS”
  - Transfer the data from the “Measurement” column in “Variable MSA” tab of “Sample Data.xlsx” to the “Measurement” column in “Gage R&R (Crossed) WKS” tab.

Gage R&R Study (Crossed) Worksheet

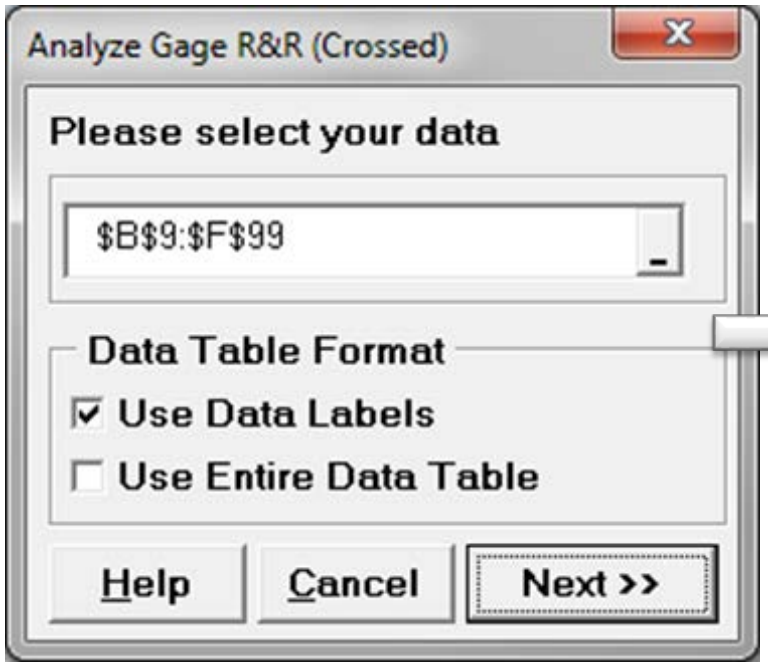
Gage Name:	
Date of Study:	
Performed By:	
Notes:	

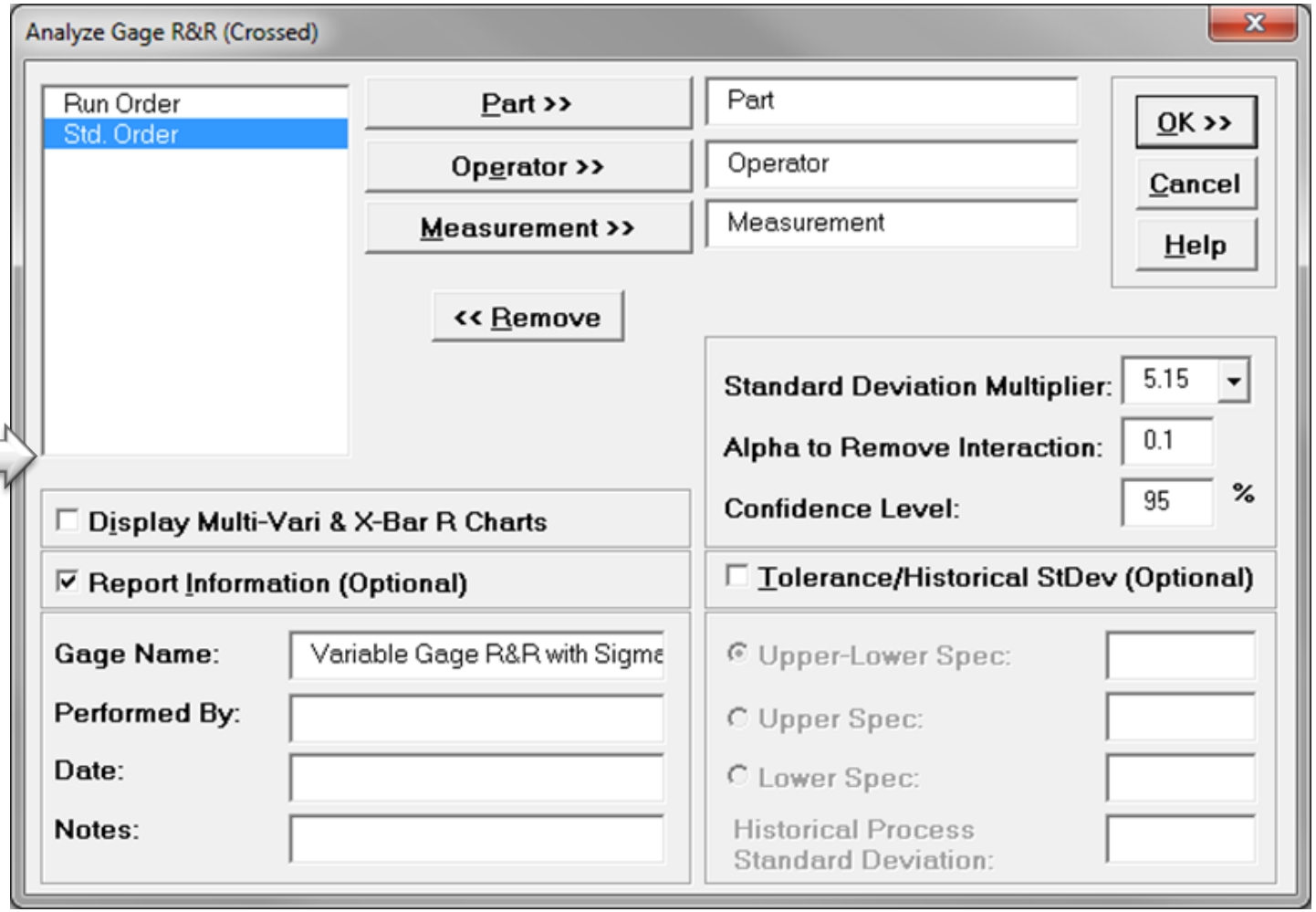
Run Order	Std. Order	Part	Operator	Measurement
1	1	Part 01	Operator A	0.29
2	2	Part 01	Operator A	0.41
3	3	Part 01	Operator A	0.64
4	4	Part 02	Operator A	-0.56
5	5	Part 02	Operator A	-0.68
6	6	Part 02	Operator A	-0.58
7	7	Part 03	Operator A	1.34
8	8	Part 03	Operator A	1.17
9	9	Part 03	Operator A	1.27
10	10	Part 04	Operator A	0.47
11	11	Part 04	Operator A	0.5
12	12	Part 04	Operator A	0.64
13	13	Part 05	Operator A	-0.8
14	14	Part 05	Operator A	-0.92
15	15	Part 05	Operator A	-0.84
16	16	Part 06	Operator A	0.02
17	17	Part 06	Operator A	-0.11
18	18	Part 06	Operator A	-0.21

Use SigmaXL to Implement a Variable MSA

- Step 4: Implement Gauge R&R
  - Click SigmaXL -> Measurement Systems Analysis -> Analyze Gage R&R (Crossed)
  - A new window named “Analyze Gage R&R (Crossed)” appears with the data range automatically selected in the box right below “Please select your data”
  - Click “Next>>”
  - A new window also named “Analyze Gauge R&R (Crossed)” pops up.
  - Select “Part” column as “Part”
  - Select “Operator” column as “Operator”
  - Select “Measurement” column as “Measurement”
  - Enter 5.15 as the “Standard Deviation Multiplier” and enter 95% as the “Confidence Level”.
  - Click “OK”
  - A new tab named “Analyze Gage R&R (1)” appears automatically.

Use SigmaXL to Implement a Variable MSA





Use SigmaXL to Implement a Variable MSA

- 5.15 is the recommended standard deviation multiplier by the Automotive Industry Action Group (AIAG). It corresponds to 99% of data in the normal distribution. If we use 6 as the standard deviation multiplier, it corresponds to 99.73% of the data in the normal distribution.

Confidence Level	Sigma Multiplier
90%	3.29
95%	3.92
99%	5.15
99.73%	6

Use SigmaXL to Implement a Variable MSA

- Step 4: Analyze the MSA results

The percentage of variation R&R contributes to the total variation is 27.86% and the precision level of this measurement system is not good. Actions are required to calibrate the measurement system.

Gage R&R Metrics	StDev	StDev Lower 95% CI	StDev Upper 95% CI	5.15 * StDev	% Total Variation (TV)	% TV Lower 95 % CI	% TV Upper 95 % CI
Gage R&R:	0.302372	0.227454	1.457	1.557	27.86	13.70	81.64
Operator (AV Appraiser Variation):	0.226838	0.113785	1.443	1.168	20.90		
Part * Operator (INT Interaction):	0	0	0	0	0.00		
Reproducibility (SQRT(AV^2 + INT^2)):	0.226838	0.113785	1.443	1.168	20.90		
Repeatability (EV Equipment Variation):	0.199933	0.172885	0.237094	1.029655876	18.42		
Part Variation (PV):	1.042327494	0.715272	1.906	5.368	96.04		
Total Variation (TV):	1.085299563	0.775728	2.106	5.589	100.00		

Note: The tab “Analyze Gage R&R (1)” in SigmaXL covers the detailed calculation of the sources of variation and also variance components.

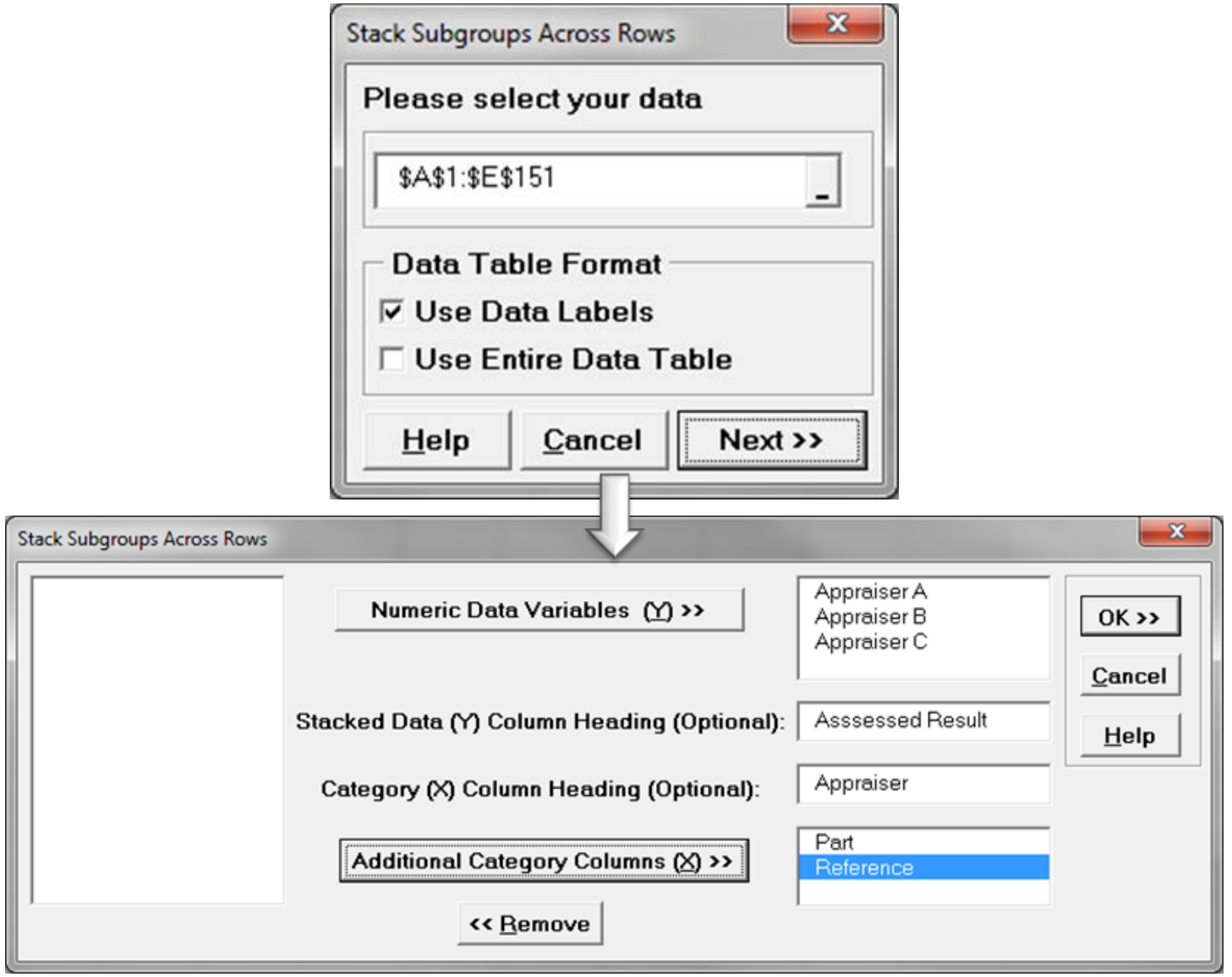
Use SigmaXL to Implement an Attribute MSA

- Data File: “Attribute MSA” tab in “Sample Data.xlsx” (an example in the AIAG MSA Reference Manual, 3rd Edition)
- Steps in SigmaXL to run an attribute MSA

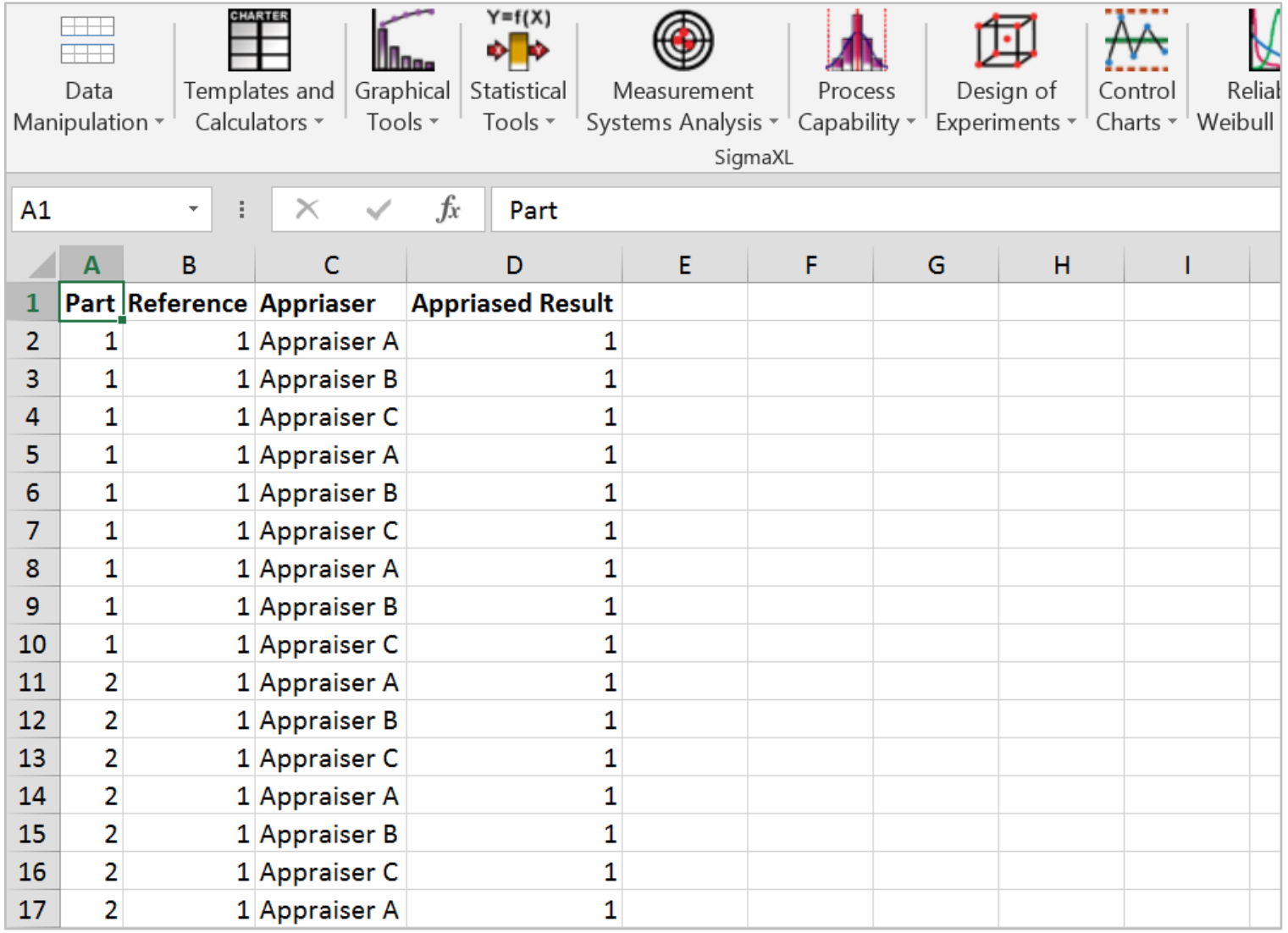
Use SigmaXL to Implement an Attribute MSA

- **Step 1: Organize the original data into four columns (“Part”, “Reference”, “Appraiser” and “Assessed Result”)**
  - Select the entire range of the original data (“Part”, “Reference”, “Appraiser A”, “Appraiser B” and “Appraiser C” columns)
  - Click SigmaXL -> Data Manipulation -> Stack Subgroups Across Rows
  - A new window named “Stack Subgroups” pops with the selected data range appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Stack Subgroups Across Rows” appears
  - Select “Appraiser A”, “Appraiser B” and “Appraiser C” as “Numeric Data Variables”
  - Select “Part” and “Reference” as the “Additional Category Columns”
  - Enter “Assessed Result” as the “Stacked Data (Y) Column Heading (Optional)
  - Enter “Appraiser” as the “Category (X) Column Heading (Optional)”
  - Click “OK>>”
  - The stacked data are created in a new worksheet.

Use SigmaXL to Implement an Attribute MSA



Use SigmaXL to Implement an Attribute MSA

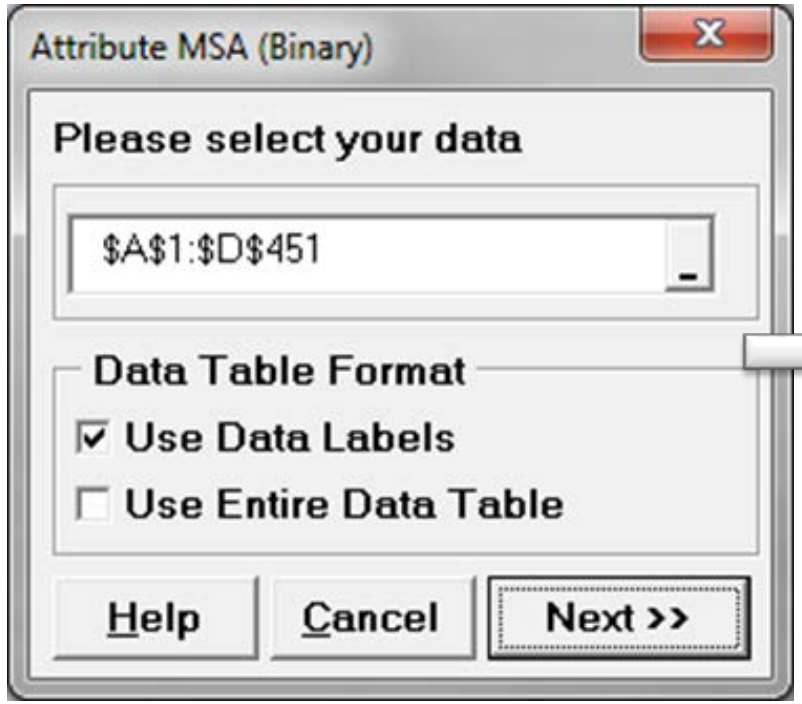


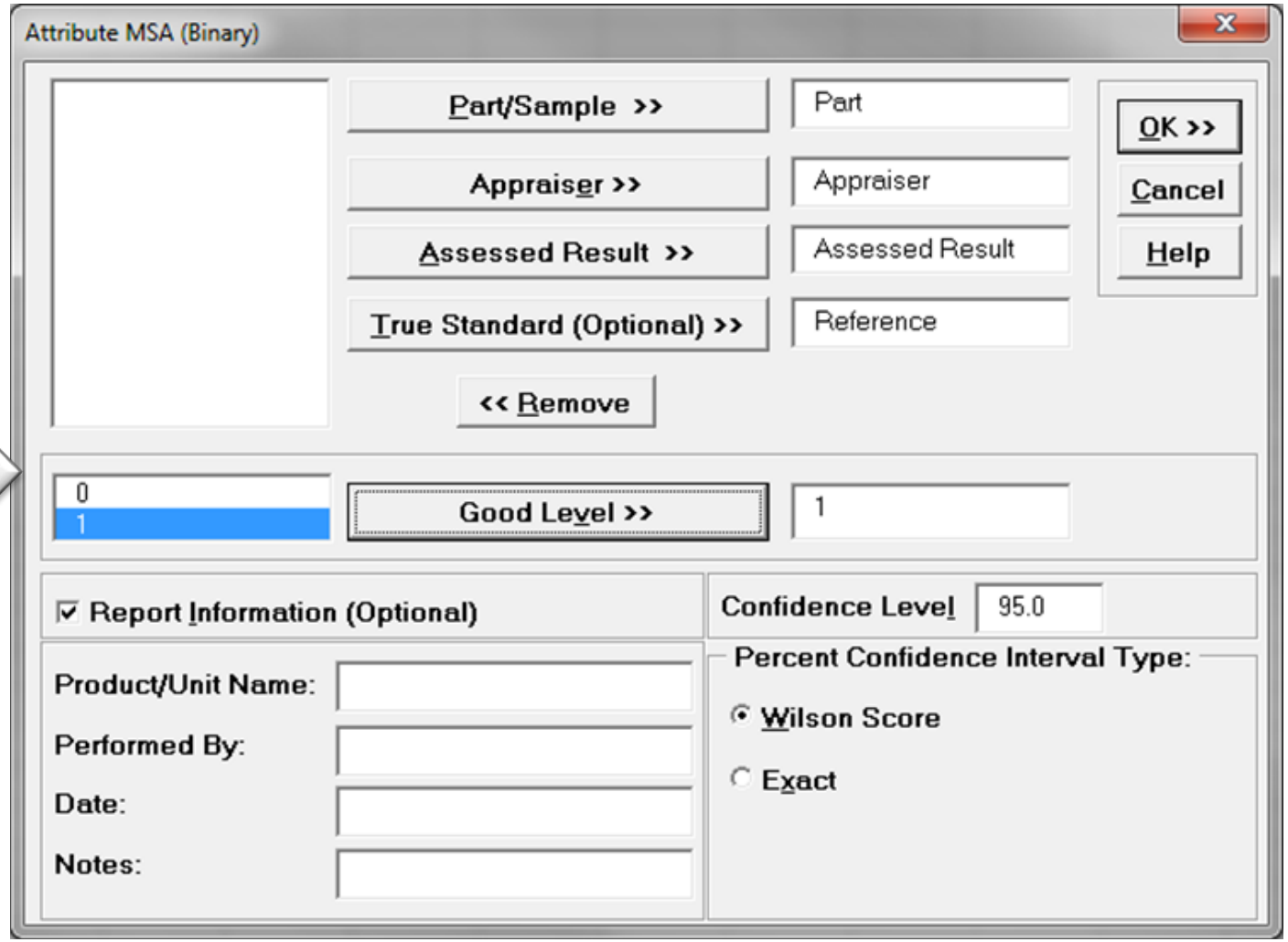
	A	B	C	D	E	F	G	H	I
1	Part	Reference	Appriaser	Appriased Result					
2	1	1	Appraiser A	1					
3	1	1	Appraiser B	1					
4	1	1	Appraiser C	1					
5	1	1	Appraiser A	1					
6	1	1	Appraiser B	1					
7	1	1	Appraiser C	1					
8	1	1	Appraiser A	1					
9	1	1	Appraiser B	1					
10	1	1	Appraiser C	1					
11	2	1	Appraiser A	1					
12	2	1	Appraiser B	1					
13	2	1	Appraiser C	1					
14	2	1	Appraiser A	1					
15	2	1	Appraiser B	1					
16	2	1	Appraiser C	1					
17	2	1	Appraiser A	1					

Use SigmaXL to Implement an Attribute MSA

- Step 2: Run MSA using SigmaXL
  - Select the entire range of the data (“Part”, “Reference”, “Appraiser” and “Assessment Result” columns)
  - Click SigmaXL -> Measurement Systems Analysis -> Attribute MSA (Binary)
  - A new window named “Attribute MSA (Binary)” pops with the selected data range appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Attribute MSA (Binary)” appears
  - Select “Part” column as “Part/Sample”
  - Select “Appraiser” column as “Appraiser”
  - Select “Assessed Result” column as “Assessed Result”
  - Select “1” as “Good Level”
  - Click “OK”
  - The MSA results appear in the newly generated tab “Att_MSA_Bin”.

Use SigmaXL to Implement an Attribute MSA





Use SigmaXL to Implement an Attribute MSA

Within Appraiser Agreement Percent: the agreement percentage within each individual appraiser.

Attribute Agreement Report:

Within Appraiser Agreement	# Inspected	# Matched	Percent	95.0% LC (Score)	95.0% UC (Score)	Fleiss' Kappa	Fleiss' Kappa P-Value	Fleiss' Kappa 95.0% LC	Fleiss' Kappa 95.0% UC
Appraiser A	50	42	84.00	71.49	91.66	0.7600	0.0000	0.6000	0.9200
Appraiser B	50	45	90.00	78.64	95.65	0.8451	0.0000	0.6850	1.0000
Appraiser C	50	40	80.00	66.96	88.76	0.7029	0.0000	0.5429	0.8629

Each Appraiser vs. Standard Agreement	# Inspected	# Matched	Percent	95.0% LC (Score)	95.0% UC (Score)	Fleiss' Kappa	Fleiss' Kappa P-Value	Fleiss' Kappa 95.0% LC	Fleiss' Kappa 95.0% UC
Appraiser A	50	42	84.00	71.49	91.66	0.8802	0.0000	0.7202	1.0000
Appraiser B	50	45	90.00	78.64	95.65	0.9226	0.0000	0.7626	1.0000
Appraiser C	50	40	80.00	66.96	88.76	0.7747	0.0000	0.6147	0.9347

Each Appraiser vs. Standard Agreement Percent: the agreement percentage between each appraiser and the standard. It reflects the accuracy of the measurement system.

Use SigmaXL to Implement an Attribute MSA

Between Appraiser Agreement Percent: the agreement percentage between different appraisers.

Between Appraiser Agreement	# Inspected	# Matched	Percent	95.0% LC (Score)	95.0% UC (Score)	Fleiss' Kappa	Fleiss' Kappa P-Value	Fleiss' Kappa 95.0% LC	Fleiss' Kappa 95.0% UC
Vertical (Value) Axis	50	39	78.00	64.76	87.25	0.7936	0.0000	0.7474	0.8398

All Appraisers vs. Standard Agreement	# Inspected	# Matched	Percent	95.0% LC (Score)	95.0% UC (Score)	Fleiss' Kappa	Fleiss' Kappa P-Value	Fleiss' Kappa 95.0% LC	Fleiss' Kappa 95.0% UC
	50	39	78.00	64.76	87.25	0.8592	0.0000	0.7668	0.9516

All Appraisers vs. Standard Agreement Percent: overall agreement percentage of both within and between appraisers. It reflects how precise the measurement system performs.

Use SigmaXL to Implement an Attribute MSA

- Kappa statistic is a coefficient indicating the agreement percentage above the expected agreement by chance.
- Kappa ranges from -1 (perfect disagreement) to 1 (perfect agreement).
- When the observed agreement is less than the chance agreement, Kappa is negative.
- When the observed agreement is greater than the chance agreement, kappa is positive.
- Rule of Thumb: If Kappa is greater than 0.7, the measurement system is acceptable. If Kappa is greater than 0.9, the measurement system is excellent.

Use SigmaXL to Implement an Attribute MSA

Kappa statistic of the agreement within each appraiser

Attribute Agreement Report:

Within Appraiser Agreement	# Inspected	# Matched	Percent	95.0% LC (Score)	95.0% UC (Score)	Fleiss' Kappa	Fleiss' Kappa P-Value	Fleiss' Kappa 95.0% LC	Fleiss' Kappa 95.0% UC
Appraiser A	50	42	84.00	71.49	91.66	0.7600	0.0000	0.6000	0.9200
Appraiser B	50	45	90.00	78.64	95.65	0.8451	0.0000	0.6850	1.0000
Appraiser C	50	40	80.00	66.96	88.76	0.7029	0.0000	0.5429	0.8629

Each Appraiser vs. Standard Agreement	# Inspected	# Matched	Percent	95.0% LC (Score)	95.0% UC (Score)	Fleiss' Kappa	Fleiss' Kappa P-Value	Fleiss' Kappa 95.0% LC	Fleiss' Kappa 95.0% UC
Appraiser A	50	42	84.00	71.49	91.66	0.8802	0.0000	0.7202	1.0000
Appraiser B	50	45	90.00	78.64	95.65	0.9226	0.0000	0.7626	1.0000
Appraiser C	50	40	80.00	66.96	88.76	0.7747	0.0000	0.6147	0.9347

Kappa statistic of the agreement between individual appraiser and the standard

Use SigmaXL to Implement an Attribute MSA

Kappa statistic of the agreement between appraisers

Between Appraiser Agreement	# Inspected	# Matched	Percent	95.0% LC (Score)	95.0% UC (Score)	Fleiss' Kappa	Fleiss' Kappa P-Value	Fleiss' Kappa 95.0% LC	Fleiss' Kappa 95.0% UC
Vertical (Value) Axis	50	39	78.00	64.76	87.25	0.7936	0.0000	0.7474	0.8398

All Appraisers vs. Standard Agreement	# Inspected	# Matched	Percent	95.0% LC (Score)	95.0% UC (Score)	Fleiss' Kappa	Fleiss' Kappa P-Value	Fleiss' Kappa 95.0% LC	Fleiss' Kappa 95.0% UC
	50	39	78.00	64.76	87.25	0.8592	0.0000	0.7668	0.9516

Kappa statistic of the overall agreement between appraisers and the standard

2.4 Process Capability

Black Belt Training: Measure Phase

2.1 Process Definition
- 2.1.1 Cause and Effect Diagrams
- 2.1.2 Cause and Effects Matrix
- 2.1.3 Process Mapping
- 2.1.4 FMEA: Failure Modes and Effects Analysis
- 2.1.5 Theory of Constraints

2.2 Six Sigma Statistics
- 2.2.1 Basic Statistics
- 2.2.2 Descriptive Statistics
- 2.2.3 Distributions and Normality
- 2.2.4 Graphical Analysis

2.3 Measurement System Analysis
- 2.3.1 Precision and Accuracy
- 2.3.2 Bias, Linearity, and Stability
- 2.3.3 Gage R&R
- 2.3.4 Variable and Attribute MSA

2.4 Process Capability
- 2.4.1 Capability Analysis
- 2.4.2 Concept of Stability
- 2.4.3 Attribute and Discrete Capability
- 2.4.4 Monitoring Techniques

2.4.1 Capability Analysis

What is Process Capability?

- The **process capability** measures how well the process performs to meet given specified outcome.

- It indicates the conformance of a process to meet given requirements or specifications.

- **Capability analysis** helps to better understand the performance of the process with respect to meeting customer's specifications and identify the process improvement opportunities.

Process Capability Analysis Steps

- Step 1: Determine the metric or parameter to measure and analyze.
- Step 2: Collect the historical data for the parameter of interest.
- Step 3: Prove the process is statistically stable (i.e., in control).
- Step 4: Calculate the process capability indices.
- Step 5: Monitor the process and ensure it remains in control over time. Update the process capability indices if needed.

Process Capability Indices

- Process capability can be presented using various indices depending on the nature of the process and the goal of the analysis.

- Popular process capability indices:
  - C_p
  - P_p
  - C_{pk}
  - P_{pk}
  - C_{pm}

C_p

- **C_p** stands for **capability of the process**.

$$C_p = \frac{USL - LSL}{6 \times \sigma_{within}}$$

where

$$\sigma_{within} = \frac{s_p}{c_4(d+1)} \qquad s_p = \sqrt{\frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)}{\sum_i (n_i - 1)}}$$

$$d = \sum_i (n_i - 1) \qquad c_4 = \frac{4(n-1)}{(4n-3)}$$

USL and ***LSL*** are the upper and lower specification limits.
n is the sample size.

C_p

- C_p measures the process' potential capability to meet the two-sided specifications.
- It does not take the process average into consideration.
- High C_p indicates the small spread of the process with respect to the spread of the customer specifications.
- C_p is recommended when the process is centered between the specification limits.
- C_p works when there are both upper and lower specification limits.

P_p

- **P_p** stands for **performance of the process**.

$$P_p = \frac{USL - LSL}{6 \times \sigma_{overall}}$$

where

$$\sigma_{overall} = \frac{s}{c_4(n)} \qquad s = \sqrt{\sum_i \sum_j \frac{(x_{ij} - \bar{x})^2}{n-1}}$$

$$c_4 = \frac{4(n-1)}{(4n-3)}$$

USL and ***LSL*** are the upper and lower specification limits.
n is the sample size.

P_p

- Similar to C_p, P_p measures the capability of the process to meet the two-sided specifications.
- It only focuses on the spread and does not take the process centralization into consideration.
- It is recommended when the process is centered between the specification limits.
- C_p considers the within-subgroup standard deviation and P_p considers the total standard deviation from the sample data.
- P_p works when there are both upper and lower specification limits.

C_{pk}

- **C_{pk}** stands for the capability of the process with a k factor adjustment.

$$C_{pk} = (1-k) \times C_p$$

where

$$k = \frac{|m-\mu|}{\frac{USL-LSL}{2}} \qquad m = \frac{USL+LSL}{2}$$

μ is the process mean; *n* is the sample size.

USL and ***LSL*** are the upper and lower specification limits.

C_{pk}

- The formulas to calculate C_{pk} can also be expressed as follows:

$$C_{pk} = \min\left(\frac{USL-\mu}{3\times\sigma_{within}}, \frac{\mu-LSL}{3\times\sigma_{within}}\right)$$

where

$$\sigma_{within} = \frac{s_p}{c_4(d+1)} \qquad s_p = \sqrt{\frac{\sum_i\sum_j (x_{ij}-\bar{x}_i)}{\sum_i (n_i-1)}}$$

$$d = \sum_i (n_i - 1) \qquad c_4 = \frac{4(n-1)}{(4n-3)}$$

USL and ***LSL*** are the upper and lower specification limits.

C_{pk}

- C_{pk} measures the process' actual capability by taking both the variation and average of the process into consideration.
- The process does not need to be centered between the specification limits to make the index meaningful.
- C_{pk} is recommended when the process is not in the center between the specification limits.
- When there is only a one-sided limit, C_{pk} is calculated using C_{pu} or C_{pl}.

C_{pk}

- C_{pk} for upper specification limit:

$$C_{pu} = \frac{USL - \mu}{3 \times \sigma_{within}}$$

- C_{pk} for lower specification limit:

$$C_{pl} = \frac{\mu - LSL}{3 \times \sigma_{within}}$$

USL and ***LSL*** are the upper and lower specification limits.
μ is the process mean.

P_{pk}

- **P_{pk}** stands for the performance of the process with a k factor adjustment.

$$P_{pk} = (1-k) \times P_p$$

where

$$k = \frac{|m - \mu|}{\frac{USL - LSL}{2}}$$

$$m = \frac{USL + LSL}{2}$$

USL and ***LSL*** are the upper and lower specification limits.
μ is the process mean.

P_{pk}

- The formulas to calculate P_{pk} can also be expressed as follows:

$$P_{pk} = \min\left(\frac{USL - \mu}{3 \times \sigma_{overall}}, \frac{\mu - LSL}{3 \times \sigma_{overall}}\right)$$

$$\sigma_{overall} = \frac{s}{c_4(n)} \qquad s = \sqrt{\sum_i \sum_j \frac{(x_{ij} - \bar{x})^2}{n-1}}$$

$$c_4 = \frac{4(n-1)}{(4n-3)}$$

USL and ***LSL*** are the upper and lower specification limits.
μ is the process mean. *n* is the sample size.

P_{pk}

- Similar to C_{pk}, P_{pk} measures the process capability by taking both the variation and the average of the process into consideration.
- P_{pk} solves the decentralization problem P_p cannot overcome.
- C_{pk} considers the within-subgroup standard deviation, while P_{pk} considers the total standard deviation from the sample data.
- When there is only a one-sided specification limit, P_{pk} is calculated using P_{pu} or P_{pl}.

P_{pk}

- P_{pk} for upper specification limit:

$$P_{pu} = \frac{USL - \mu}{3 \times \sigma_{overall}}$$

- P_{pk} for lower specification limit:

$$P_{pl} = \frac{\mu - LSL}{3 \times \sigma_{overall}}$$

USL and ***LSL*** are the upper and lower specification limits.

C_{pm}

- C_p, P_p, C_{pk}, and P_{pk} all consider the variation of the process. C_{pk} and P_{pk} take both the variation and the average of the process into consideration when measuring the process capability.

- It is possible that the process average fails to meet the target customers require while the process still remains between the specification limits. C_{pm} (Taguchi's capability index) helps to capture the variation from the specified target.

C_{pm}

- Formula to calculate C_{pm}

$$C_{pm} = \frac{\min(T - LSL, USL - T)}{3 \times \sqrt{s^2 + (\mu - T)^2}}$$

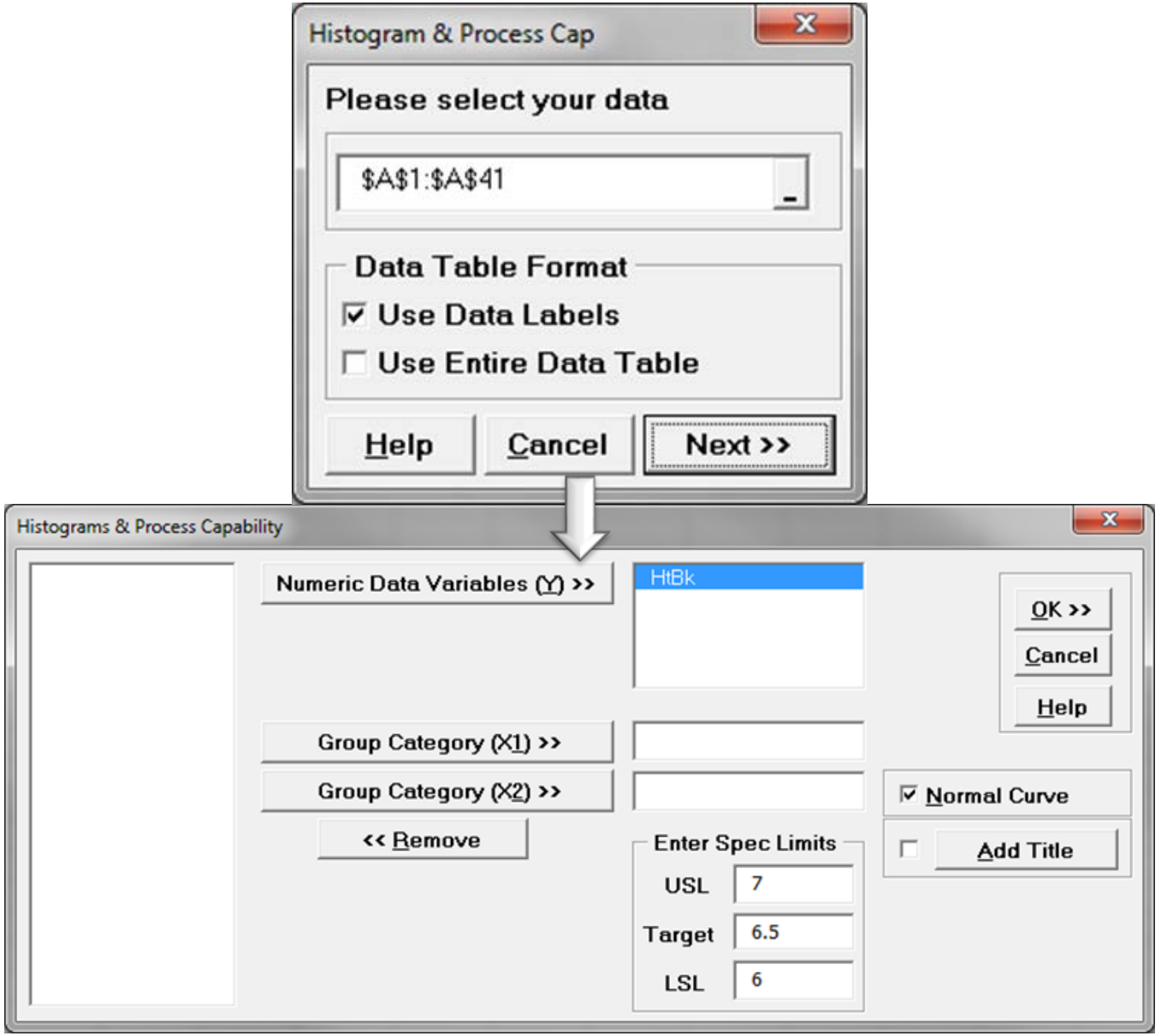
USL and ***LSL*** are the upper and lower specification limits.
T is the specified target.
μ is the process mean.

Note: Cpm can work only if there is a target value specified.

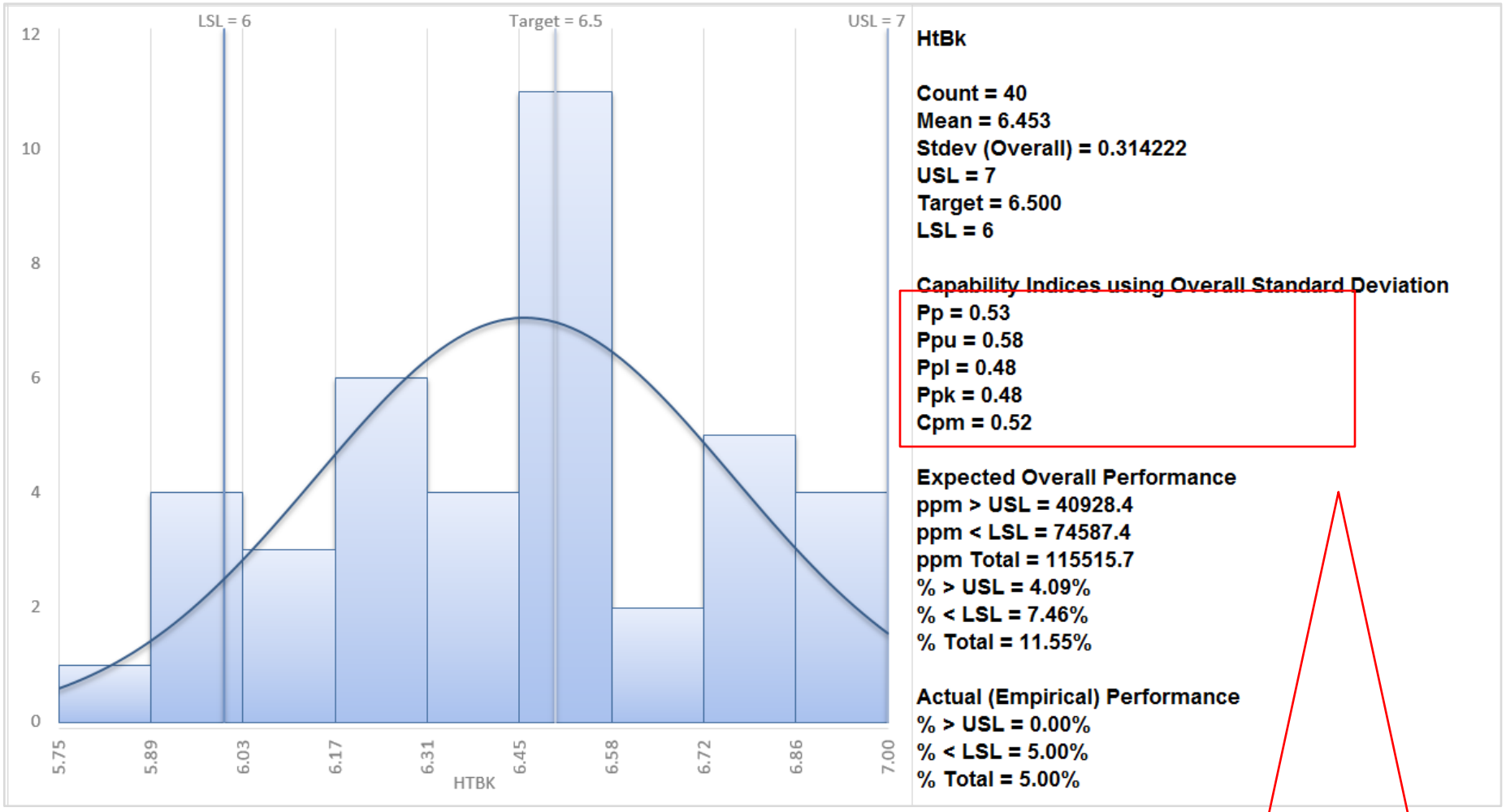
Use SigmaXL to Run a Process Capability Analysis

- Data File: “Capability Analysis” tab in “Sample Data.xlsx”
- Steps in SigmaXL to run a process capability analysis:
  - Select the entire range of data (i.e. the column “HtBk”)
  - Click SigmaXL -> Process Capability -> Histograms & Process Capability
  - A new window named “Histogram & Process Cap” pops up with the selected range of data appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Histograms & Process Capability” appears
  - Select “HtBk” as the “Numeric Data Variables”
  - Enter 6 in LSL, 6.5 in T and 7 in USL into the boxes for “Lower Spec Limit”, “Target” and “Upper Spec Limit” respectively
  - Click “OK”
  - The histogram and the process capability analysis results are in the newly generated tab “Hist Cap (1)”

Use SigmaXL to Run a Process Capability Analysis



Use SigmaXL to Run a Process Capability Analysis



Process capability indices

2.4.2 Concept of Stability

What is Process Stability?

- A process is said to be stable when:
  - the process is in control
  - the future behavior of the process is predictable at least between some limits
  - there is only random variation involved in the process.
  - the causes of variation in the process are only due to chance or common causes
  - there are not any trends, patterns, or outliers in the control chart of the process.

Root Causes of Variation in the Process

- **Common Cause:**
  - Chance
  - Random and anticipated
  - Natural noise
  - Inherent in the process
  - Unable to be eliminated from the process.

- **Special Cause:**
  - Assignable cause
  - Unanticipated
  - Unnatural pattern
  - Signal of changes in the process
  - Able to be eliminated from the process.

Control Charts

- **Control charts** are the graphical tools to analyze the stability of a process.
- A control chart is used to identify the presence of potential special causes in the process and to determine whether the process is statistically in control.
- If the samples or calculations of samples are all in control, the process is stable and the data from the process can be used to predict the future performance of the process.

Popular Control Charts

- I-MR Chart
- Xbar-R Chart
- Xbar-S Chart
- C Chart
- U Chart
- P Chart
- NP Chart
- EWMA Chart
- CUSUM Chart

Note: More details of the control charts will be introduced in the Control module.

Process Stability vs. Process Capability

- Process stability indicates how stable a process performed in the past.
- When the process is stable, we can use the data from the process to predict its future behavior.
- Process capability indicates how well a process performs with respect to meeting the customer's specifications.
- The process capability analysis is valid only if the process is statistically stable (i.e., in control, predictable).
- Being stable does *not* guarantee that the process is also capable. However, being stable is the prerequisite to determine whether a process is capable.

2.4.3 Attribute & Discrete Capability

Process Capability Analysis for Binomial Data

- If we are measuring the count of defectives in each sample set to assess the process performance of meeting the customer specifications, we use "%Defective" (percentage of items in the samples that are defective) as the process capability index.

$$\%\text{Defective} = \frac{N_{defectives}}{N_{overall}}$$

where $N_{defectives}$ is the total count of defectives in the samples and $N_{overall}$ is the sum of all the sample sizes.

Process Capability Analysis for Poisson Data

- If we are measuring the count of defects in each sample set to assess the process performance of meeting the customer specifications, we use Mean DPU (defects per unit of measurement) as the process capability index.

$$DPU = \frac{N_{defects}}{N_{overall}}$$

where $N_{defects}$ is the total count of defects in the samples and $N_{overall}$ is the sum of all the units in the samples.

2.4.4 Monitoring Techniques

Capability and Monitoring

- In the Measure phase of the project, process stability analysis and process capability analysis are used to baseline the performance of current process.

- In the Control phase of the project, process stability analysis and process capability analysis are combined to monitor whether the improved process is maintained consistently as expected.

3.0 Analyze Phase

Black Belt Training: Analyze Phase

3.1 Patterns of Variation
- 3.1.1 Multi-Vari Analysis
- 3.1.2 Classes of Distributions

3.2 Inferential Statistics
- 3.2.1 Understanding Inference
- 3.2.2 Sampling Techniques and Uses
- 3.2.3 Sample Size
- 3.2.4 Central Limit Theorem

3.3 Hypothesis Testing
- 3.3.1 Goals of Hypothesis Testing
- 3.3.2 Statistical Significance
- 3.3.3 Risk; Alpha and Beta
- 3.3.4 Types of Hypothesis Tests

3.4 Hypothesis Testing: Normal Data
- 3.4.1 One and Two Sample T-Tests
- 3.4.2 One sample variance
- 3.4.3 One Way ANOVA

3.5 Hyp Testing: Non-Normal Data
- 3.5.1 Mann-Whitney
- 3.5.2 Kruskal-Wallis
- 3.5.3 Moods Median
- 3.5.4 Friedman
- 3.5.5 One Sample Sign
- 3.5.6 One Sample Wilcoxon
- 3.5.7 One and Two Sample Proportion
- 3.5.8 Chi-Squared (Contingency Tables)
- 3.5.9 Test of Equal Variances

3.1 Patterns of Variation

Black Belt Training: Analyze Phase

3.1 Patterns of Variation
- 3.1.1 Multi-Vari Analysis
- 3.1.2 Classes of Distributions

3.2 Inferential Statistics
- 3.2.1 Understanding Inference
- 3.2.2 Sampling Techniques and Uses
- 3.2.3 Sample Size
- 3.2.4 Central Limit Theorem

3.3 Hypothesis Testing
- 3.3.1 Goals of Hypothesis Testing
- 3.3.2 Statistical Significance
- 3.3.3 Risk; Alpha and Beta
- 3.3.4 Types of Hypothesis Tests

3.4 Hypothesis Testing: Normal Data
- 3.4.1 One and Two Sample T-Tests
- 3.4.2 One sample variance
- 3.4.3 One Way ANOVA

3.5 Hyp Testing: Non-Normal Data
- 3.5.1 Mann-Whitney
- 3.5.2 Kruskal-Wallis
- 3.5.3 Moods Median
- 3.5.4 Friedman
- 3.5.5 One Sample Sign
- 3.5.6 One Sample Wilcoxon
- 3.5.7 One and Two Sample Proportion
- 3.5.8 Chi-Squared (Contingency Tables)
- 3.5.9 Test of Equal Variances

3.1.1 Multi-Vari Analysis

What is Multi-Vari Analysis?

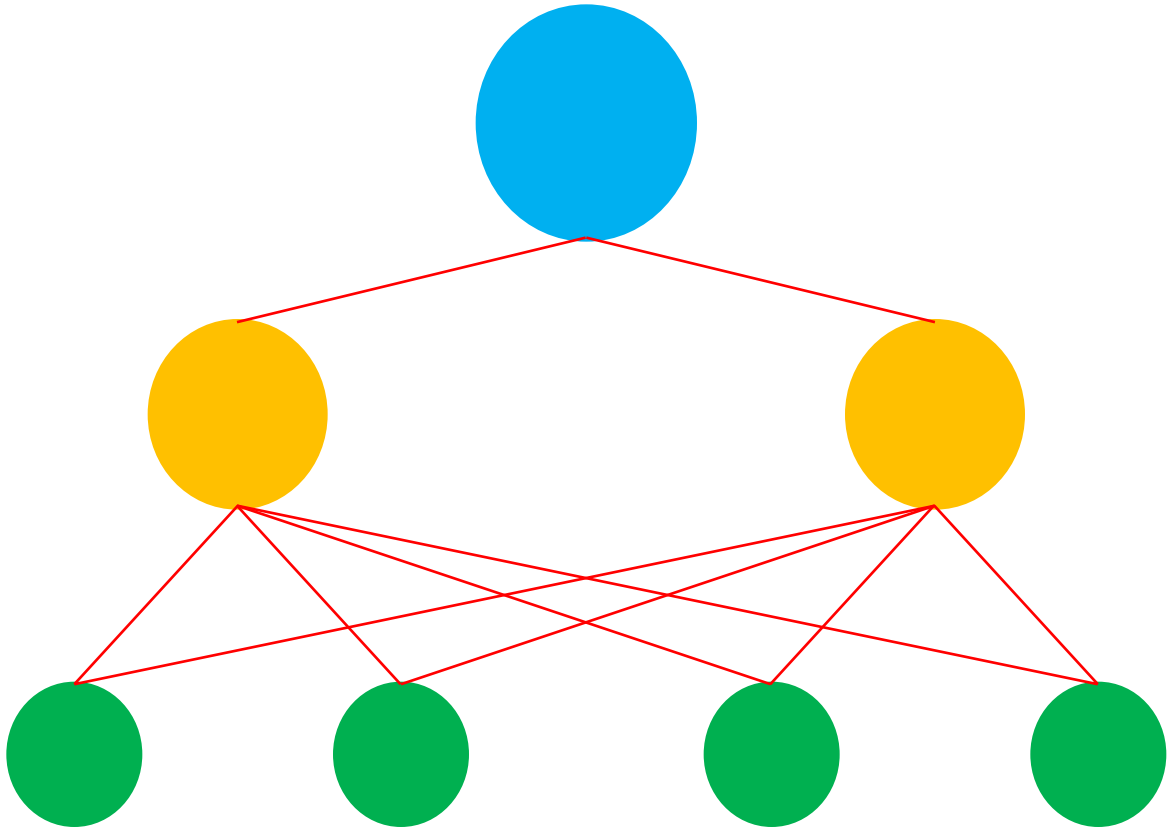
- **Multi-Vari analysis** is a graphic-driven method to analyze the effects of categorical inputs on a continuous output.
  - Y: continuous variable
  - X's: discrete categorical variables. One X may have multiple levels.
- It studies how the variation in the output changes across different inputs and helps us quantitatively determine the major source of variability in the output.
- Multi-Vari charts are used to visualize the source of variation. They work for both crossed and nested hierarchies.

Hierarchy

- **Hierarchy** is a structure of objects in which the relationship of objects can be expressed similar to an organization tree.
- Each object in the hierarchy is described as above, below, or at the same level as another one.
  - If object A is above object B and they are directly connected to each other in the hierarchy tree, A is B's parent and B is A's child.
- In Multi-Vari analysis, we use the hierarchy to present the relationship between categorical factors (inputs).
- Each object in the hierarchy tree indicates a specific level of a factor (input).
- There are generally two types of hierarchies: crossed and nested.

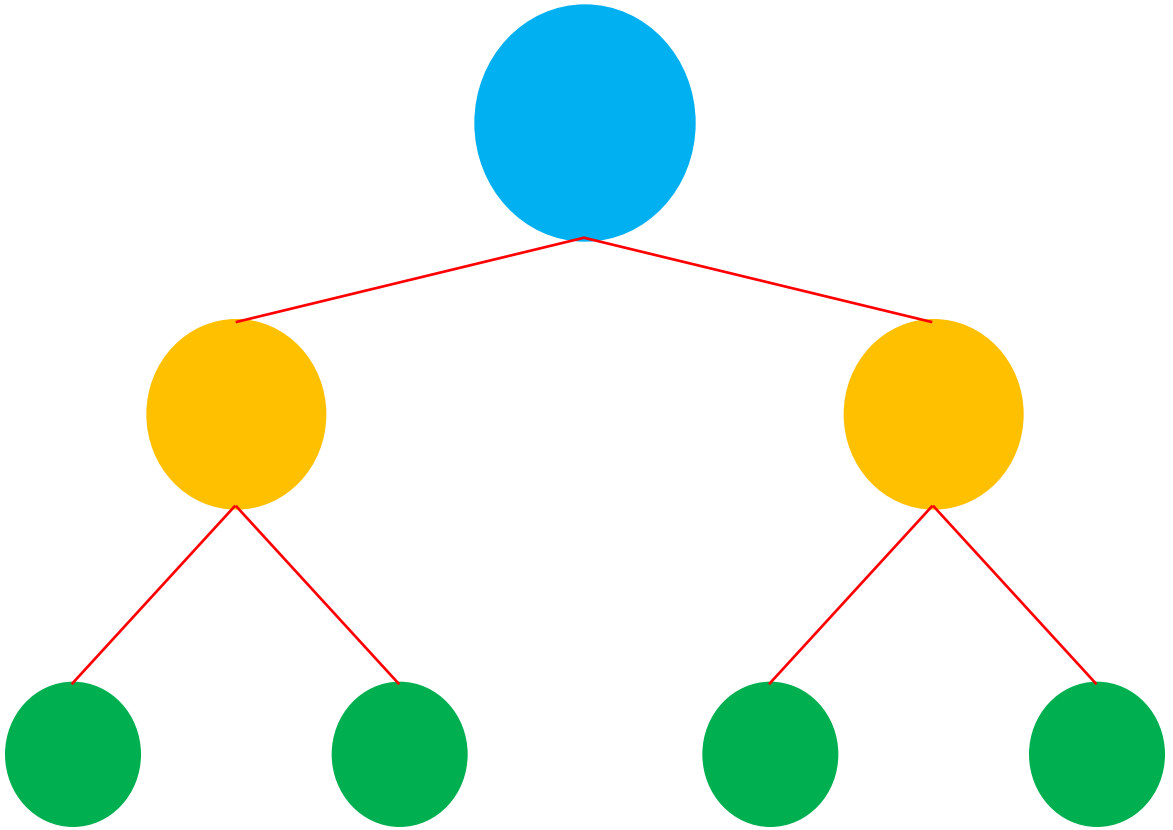
Crossed Hierarchy

- In the hierarchy tree, if one child item has more than one parent item at the higher level, it is a crossed relationship.



Nested Hierarchy

- In the hierarchy tree, if one child item only has one parent item at the higher level, it is a nested relationship.



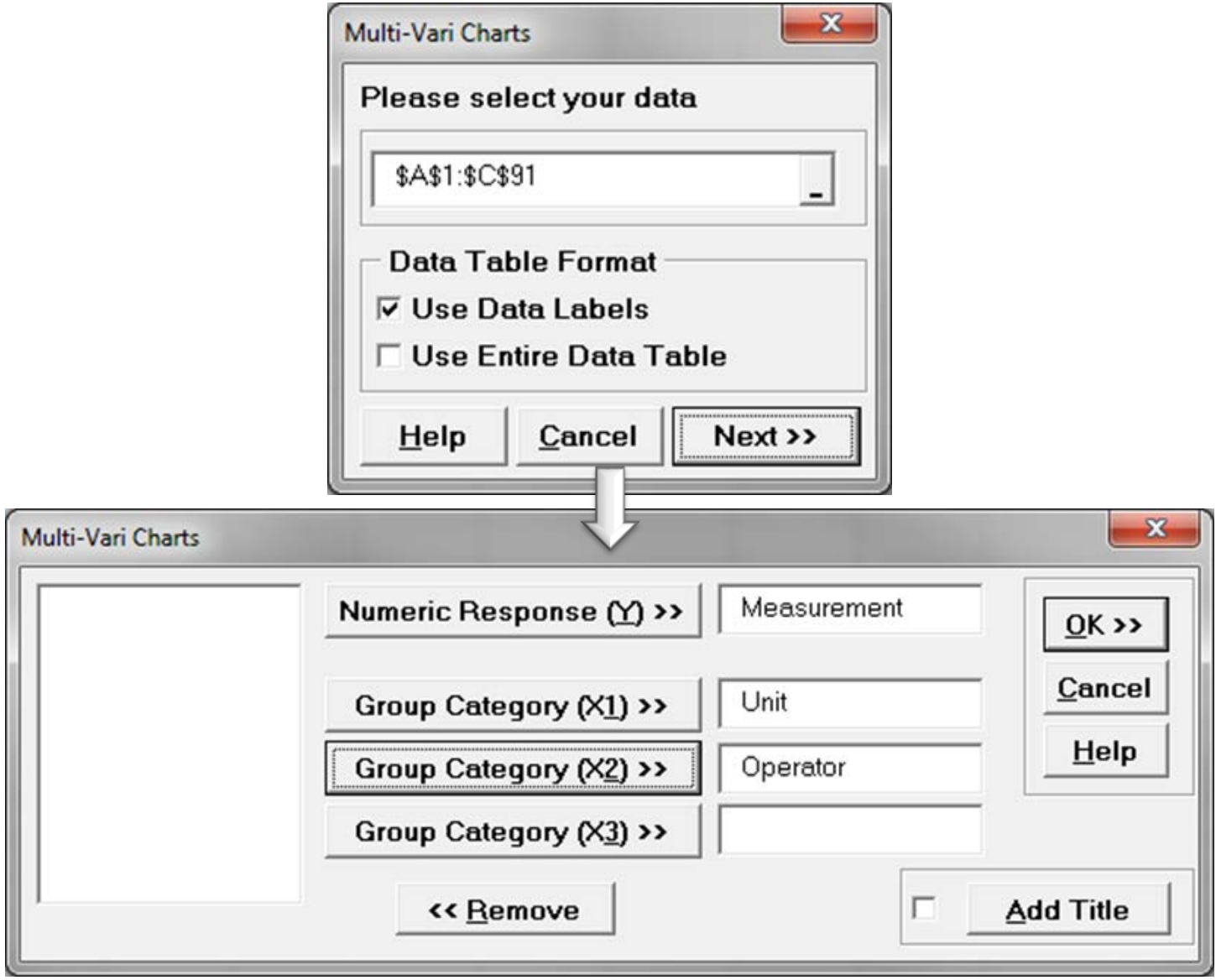
Use SigmaXL to Perform Multi-Vari Analysis

- Data File: “Multi-Vari’ tab in “Sample Data.xlsx”
  - Case study: ABC company produces 10 kinds of units with different weights. Operators measure the weights of the units before sending them out to customers.
  - Multiple factors would have impact on the weight measurement. ABC company wants to have a better understanding on the main source of the variability existing in the weight measurement.
  - ABC company randomly selected 3 operators (Joe, John and Jack) each of whom measures the weights of 10 kinds of units. For each kind of unit, there are 3 items sampled.

Use SigmaXL to Perform Multi-Vari Analysis

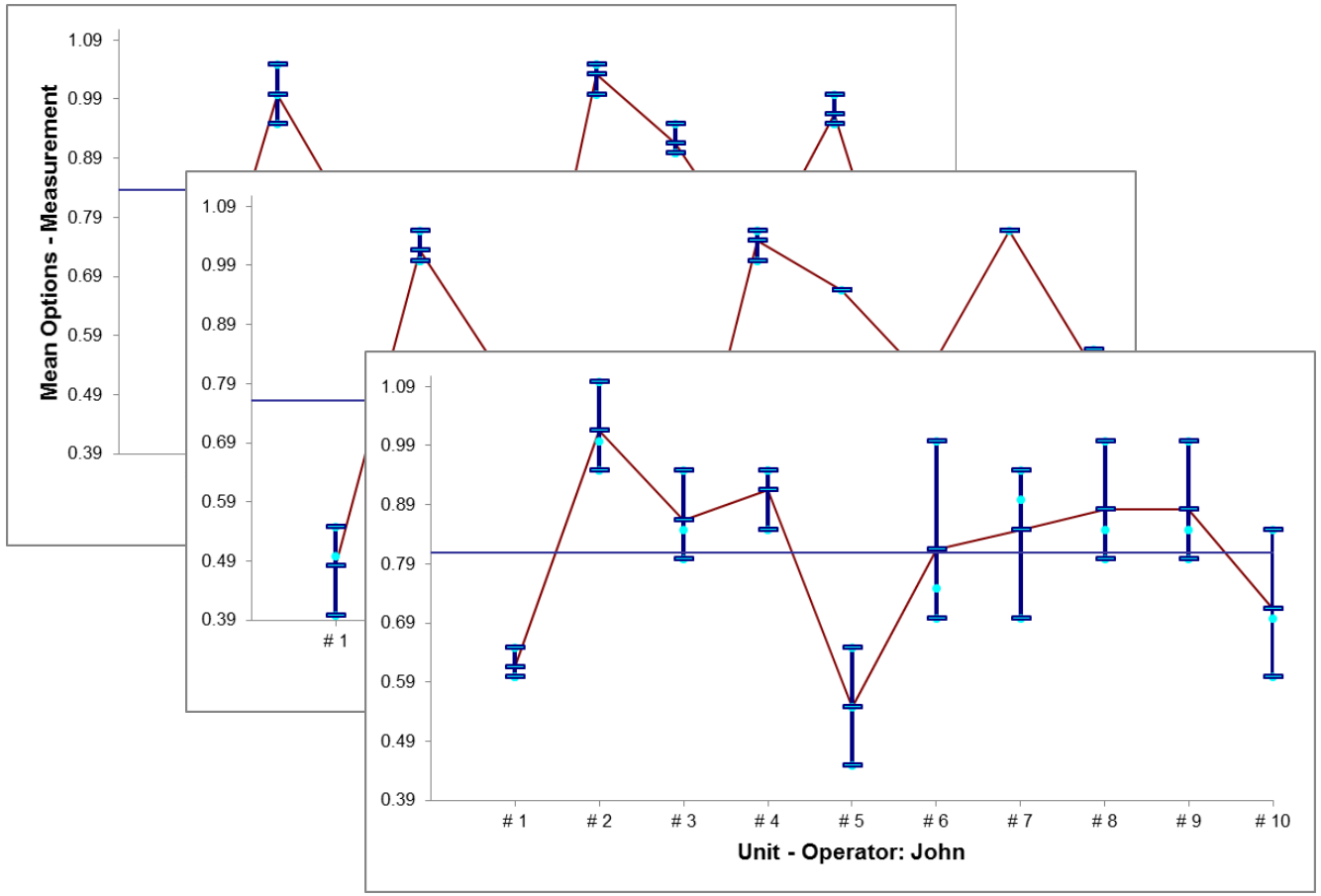
- Multi-Vari Analysis in SigmaXL:
  - Open the “Multi-Vari’ tab in “Sample Data.xlsx”
  - Select the range of data that we are interested in analyzing.
  - Click SigmaXL -> Graphical Tools -> Multi-Vari ChartsA window named “Multi-Vari Charts” pops up.
  - Click on “Next >>” button. A new window also named “Multi-Vari Charts” pops up.
  - Select “Measurement” as “Numeric Response (Y)”
  - Select “Unit” as “Group Category (X1)”
  - Select “Operator” as “Group Category (X2)”
  - Click “OK >>”

Use SigmaXL to Perform Multi-Vari Analysis



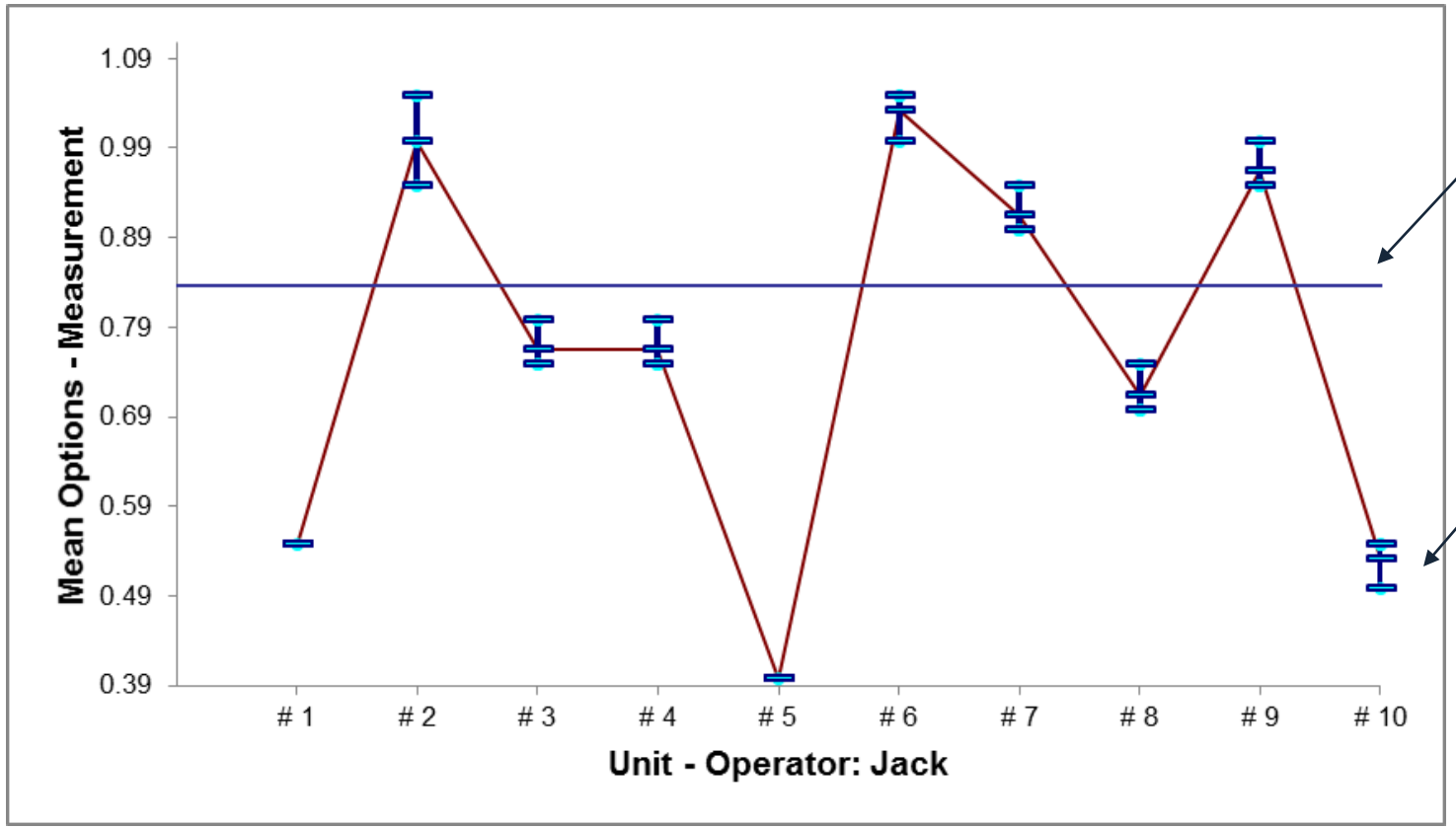
Use SigmaXL to Perform Multi-Vari Analysis

- A new spreadsheet with three charts is created.
- We call the newly generated charts “Multi-Vari Charts”.
- Analyze the charts to determine major sources of the variation



Interpreting SigmaXL's Multi-Vari Analysis

- Jack's Multi-Vari Chart



The horizontal line in the center of the chart indicates the group mean of Jack's measurement.

The maximum and minimum bars present the range of measurements for each unit. The center bar presents the mean measurement for each unit.

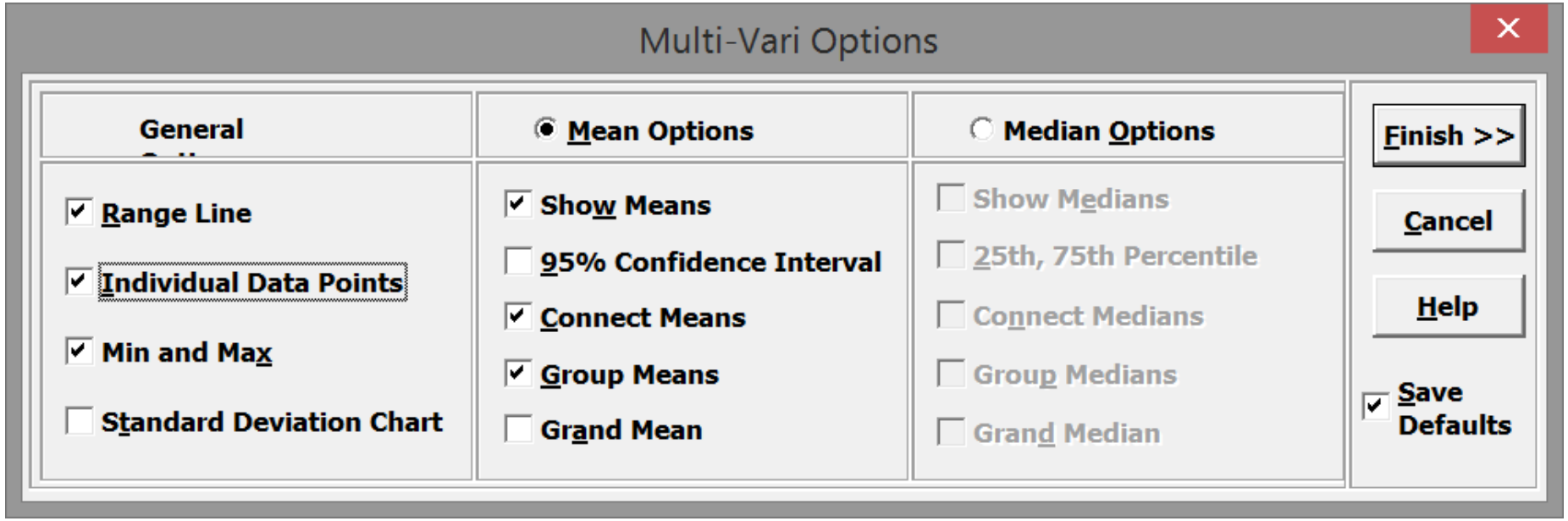
The red line connects all the center bars together.

Interpreting SigmaXL's Multi-Vari Analysis

- Based on the Multi-Vari Chart, the measurement of units ranges from 0.39 to 1.09.
- Jack's and John's average measurement stays between 0.79 and 0.89. Joe's average is slightly low than both Jack's and Johns.
- John has the worst variation when measuring the same kind of unit because John has the highest difference between maximum and minimum bars for any kind of unit.
- By observing the red lines of three operators, it seems like all three operators' measurements follow the same pattern. The operator-to-operator variability is not large.
- The unit-to-unit variability is large and it might be the main source of variation in measurements.

Use SigmaXL to Perform Multi-Vari Analysis

- To add more features to your Multi-Vari Chart
  - Click SigmaXL -> Graphical Tools -> Multi-Vari Options
  - A new window named “Multi-Vari Options” pops up.
  - In this new widow, you may add grand mean, confidence intervals, medians etc. to the Multi-Vari Chart to further your analysis of variation.



3.1.2 Classes of Distribution

What is Probability?

- **Probability** is the likelihood of an event occurring.
- The probability of event A occurring is written as P(A).
- Probability is a percentage between 0% and 100%. The higher the probability, the more likely the event will happen.
- When probability is equal to 0%, it indicates that the event will take place with no chance.
- When probability is equal to 100%, it indicates the event will definitely take place without any uncertainty.
- Example of probability: when tossing a coin, there is 50% chance that the head lands face up and 50% chance that the head lands face down.

Probability Property

- The probability of event A not occurring:

$$P(\overline{A}) = 1 - P(A)$$

To be, or not to be: that is the question.

William Shakespeare

Probability Property

- The probability of event A OR B OR both events occurring:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- If event A and B are mutually exclusive:

$$P(A \cap B) = 0$$

and

$$P(A \cup B) = P(A) + P(B)$$

Probability Property

- The probability of event A AND B occurring together:

$$P(A \cap B) = P(A \mid B)P(B)$$

- If event A and B are independent of each other:

$$P(A \mid B) = P(A)$$

and

$$P(A \cap B) = P(A)P(B)$$

Probability Property

- The probability of event A occurring given event B taking place:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{where} \quad P(B) \neq 0$$

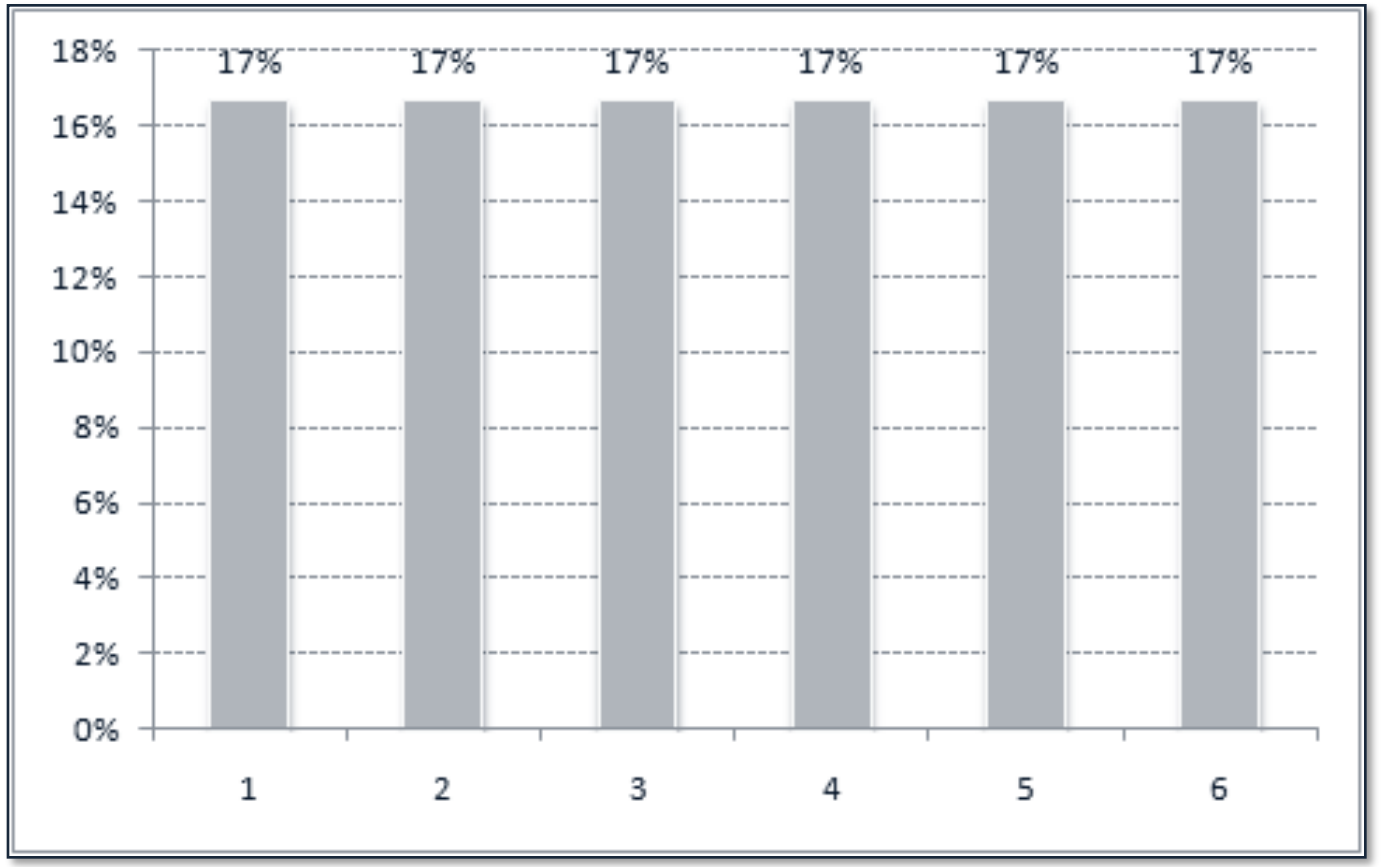
What is Distribution?

- **Distribution** (or probability distribution) describes the range of possible events and the possibility of each event occurring.

- In statistical terms, distribution is the probability of each possible value of a random variable when the variable is discrete, or the probability of a value falling in a specific interval when the variable is continuous.

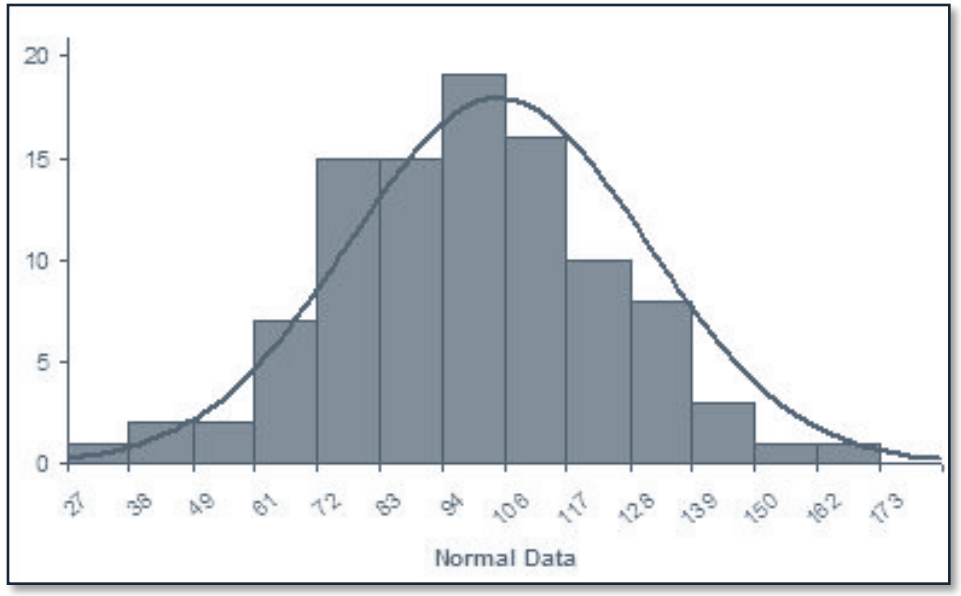
Probability Mass Function

- **PMF** (Probability Mass Function) is a function describing the probability of a discrete random variable equal to a particular value.
- Here is the probability mass function chart for a fair die.



Probability Density Function

- **PDF (Probability Density Function)** is a function used to retrieve the probability of a continuous random variable falling within a particular interval.
- The probability of a variable falling within a particular region is the integral of the probability density function over the region.
- Here is the probability density function chart of a standard normal distribution.



Advantages of Distribution

- Distributions capture the most basic features of data
  - Shape
    - Which distribution family may the data belong to?
    - Is the data symmetric?
    - Are the tails of the data flat?
  - Center
    - Where is the midpoint of the data?
  - Scale
    - What is the range of data?

Measures of Skewness and Kurtosis

- The shape, center (i.e. location) and scale (i.e. variability) are three basic data characteristics that probability distributions capture.

- Skewness and kurtosis are the two more advanced characteristics of the probability distribution.

Skewness

- **Skewness** is a measure of the asymmetry degree of the probability distribution.
- The mathematical definition of the skewness is

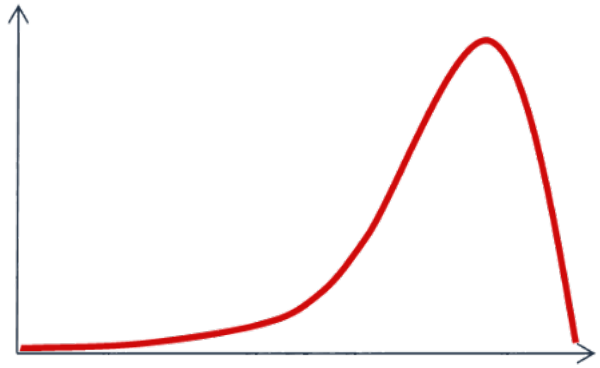
$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{(\mathrm{E}\left[(X-\mu)^2\right])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

where μ_3 is the third moment about the mean and σ is the standard deviation.

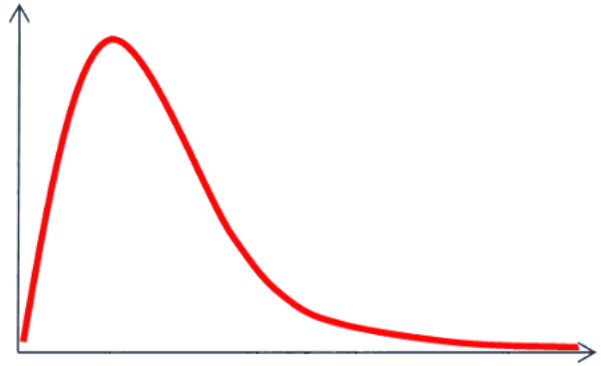
- The skewness of a normal distribution is zero.
- When the distribution looks symmetric to the left side and the right side of the center point, the skewness is close to zero.
- Potential causes of skewness:
  - Extreme values exist in the data
  - Data have a lower or higher bound.

Skewness

- Left Skew
  - Skewness < 0
  - The left tail of the distribution is longer than the right tail.



- Right Skew
  - Skewness > 0
  - The right tail of the distribution is longer than the left tail.



Kurtosis

- **Kurtosis** is a measure of the peakedness of the probability distribution.
- The mathematical definition of the excess kurtosis is

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3$$

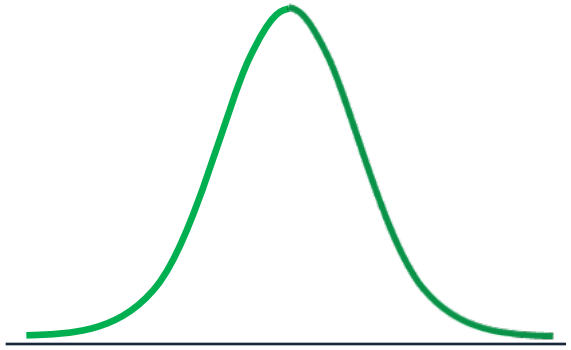
where μ_4 is the fourth moment about the mean and σ is the standard deviation.

- The kurtosis of a normal distribution is zero.
- Distributions with a zero excess kurtosis are called mesokurtic.

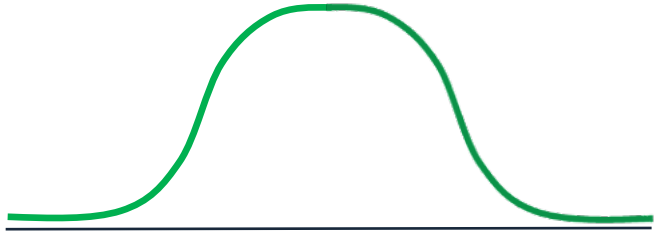
- The distribution with a high kurtosis has a more distinct peak and fatter, longer tails.
- The distribution with a low kurtosis has a more rounded peak and thinner, shorter tails.

Kurtosis

- Leptokurtic
  - Excess Kurtosis > 0
  - A sharper peak near the mean, declining rapidly, and fatter tails.



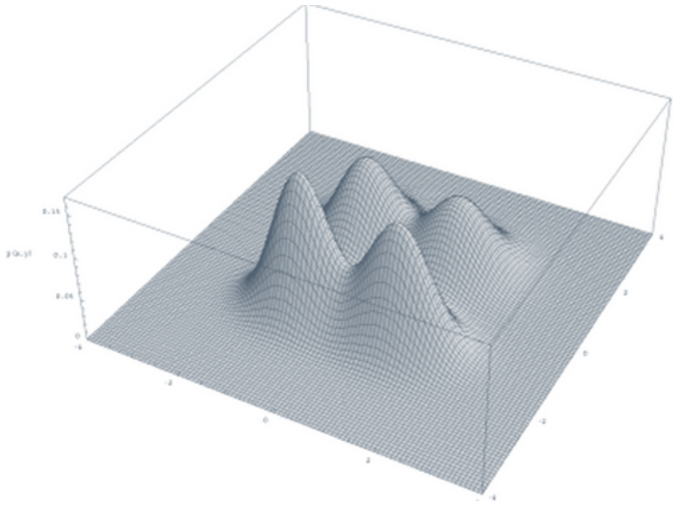
- Platykurtic
  - Excess Kurtosis < 0
  - A flatter top near the mean, declining slowly, and thinner tails.



Multimodal Distribution

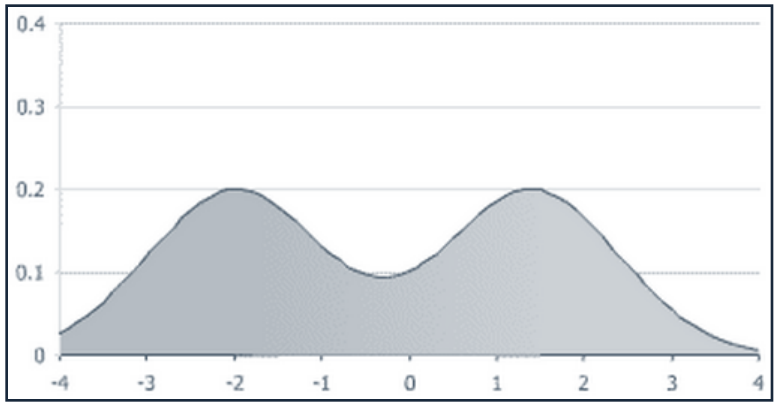
- In statistics, a continuous probability distribution with two or more different unimodal distributions is called a **multimodal distribution**.
- A **unimodal distribution** has only one local maximum which is equal to its global maximum.
- A multimodal probability distribution has more than one local maximum.
  - Example: a bivariate, multimodal distribution

(picture source: http://en.wikipedia.org/wiki/Multimodal_distribution)



Bimodal Distribution

- In statistics, a continuous probability distribution with two different unimodal distributions is called a **bimodal distribution**.
- A bimodal probability distribution has two local maxima.
- Example (picture source: http://en.wikipedia.org/wiki/Multimodal_distribution)



- By mixing two unimodal distributions with different means, the mixture distribution is not necessarily a bimodal distribution.
- The mixture of two normal distributions with the same standard deviation is a bimodal distribution only if the difference between their means is at least twice their common standard deviation.

Examples of Discrete Distribution

- Binomial Distribution
- Poisson Distribution

Binomial Distribution

- Assume there is an experiment with *n* independent trials, each of which only has two potential outcomes (i.e. yes/no, fail/pass). *p* is the probability of one outcome and (1 – *p*) is the probability of the other.

- **Binomial distribution** is a discrete probability distribution describing the probability of any outcome of the experiment.

- Excel formula for binomial distribution:
  - **BINOMDIST**(**number_s**, **trials**, **probability_s**, **cumulative**)

Binomial Distribution

- PMF of Binomial Distribution
  the probability of getting *k* successes in *n* trials:

$$f(k;n,p)=\binom{n}{k}p^{k}(1-p)^{n-k} \quad \text{where} \quad \binom{n}{k}=\frac{n!}{k!(n-k)!}$$

- Mean of Binomial Distribution

$$n\times p$$

- Variance of Binomial Distribution

$$n\times p\times(1-p)$$

Poisson Distribution

- **Poisson distribution** is a discrete probability distribution describing the probability of a number of events occurring at a known average rate and in a fixed period of time. The expected number of occurrences in this fixed period of time is λ.

- Excel formula for Poisson distribution
  - **POISSON(x, mean, cumulative)**

Poisson Distribution

- PMF of Poisson Distribution
  the probability of getting k occurrences in *n* trials

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Mean of Poisson Distribution

$$\lambda$$

- Variance of Poisson Distribution

$$\lambda$$

Examples of Continuous Distribution

- Normal Distribution
- Exponential Distribution
- Weibull Distribution
- Student’s t distribution
- Chi-square distribution
- F distribution

Normal Distribution

- **Normal distribution** is a continuous probability distribution describing random variables which cluster around the mean.
- Its probability density function is a bell-shaped curve. But not all the bell-shaped PDFs are from normal distributions.
- It is the most extensively used distribution.
- Excel formula for normal distribution
  - **NORMDIST**(**x**, **mean**, **standard_dev**, **cumulative**)
- The 68-95-99.7 rule for normal distribution:
  - about 68% of the data stay within σ from the mean
  - about 95% of the data stay within 2σ from the mean
  - about 99.7% of the data stay within 3σ from the mean.

Normal Distribution

- PDF of Normal Distribution

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Mean of Normal Distribution

$$\mu$$

- Variance of Normal Distribution

$$\sigma^2$$

Exponential Distribution

- **Exponential distribution** is a continuous probability distribution describing the probability of events occurring at a known constant average rate between points of time.

- The exponential distribution is very similar to the Poisson distribution, except that the former is built on continuous variables and the latter is built on discrete variables.

- Excel formula for exponential distribution
  - **EXPONDIST**(**x**, **lambda**, **cumulative**)

Exponential Distribution

- PDF of Exponential Distribution

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq o \\ 0 & x < o \end{cases}$$

- Mean of Exponential Distribution

$$\frac{1}{\lambda}$$

- Variance of Exponential Distribution

$$\frac{1}{\lambda^2}$$

Weibull Distribution

- **Weibull distribution** is a continuous probability distribution which is widely used to model a great variety of data due to its flexibility.

- *K* is the shape parameter and *λ* is the scale parameter. Both are positive numbers.

- Excel formula for Weibull distribution
  - **WEIBULL**(**x**, **alpha**, **beta**, **cumulative**)

Weibull Distribution

- PDF of Weibull Distribution

$$f(x)=\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)k} & x\geq o\\ 0 & x<o\end{cases}$$

- Mean of Weibull Distribution

$$\lambda\Gamma\left(1+\frac{1}{k}\right)$$

- Variance of Weibull Distribution

$$\lambda^2\Gamma\left(1+\frac{2}{k}\right)-\mu^2$$

Student's T Distribution

- **Student's t distribution** is a continuous probability distribution that resembles a normal distribution.

- When *n* random samples were drawn from a normally-distributed population with the mean μ and the standard deviation σ, then

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

follows a t distribution with (n – 1) degrees of freedom.

- Its probability density function is a bell-shaped curve but with heavier tails than those of the normal distribution.

Student's T Distribution

- PDF of Student's T Distribution

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

- Mean of Student's T Distribution

0 for degrees of freedom greater than 1, otherwise undefined

- Variance of Student's T Distribution

$$\begin{cases} \nu/(\nu-2) & \nu > 2 \\ \infty & 1 < \nu \le 2 \\ \text{undefined} & \text{otherwise} \end{cases}$$

Chi-Square Distribution

- **Chi-square distribution** is a continuous probability distribution of the sum of squares of multiple independent standard normal random variables.

- If Y_1, Y_2, …, Y_k are a group of independent normal distributed variables, each with mean 0 and standard deviation 1, then

$$\chi^2 = \sum_{i=1}^{k} Y_i^2$$

follows a chi-square distribution with *k* degrees of freedom.

Chi-Square Distribution

- PDF of Chi-Square Distribution

$$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}$$

- Mean of Chi-Square Distribution

$$k$$

- Variance of Chi-Square Distribution

$$2k$$

F Distribution

- **F distribution** is a continuous probability distribution which arises in the analysis of variance or test of equality between two variances.

- If $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$ are independent variables with chi-square distributions with ν_1 and ν_2 degrees of freedom,

$$F = \frac{\chi^2_{\nu_1} / \nu_1}{\chi^2_{\nu_2} / \nu_2}$$

follows a F distribution.

F Distribution

- PDF of F Distribution

$$\frac{\sqrt{\frac{(\nu_1 x)^{\nu_1}\,\nu_2^{\nu_2}}{(\nu_1 x+\nu_2)^{\nu_1+\nu_2}}}}{xB\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}$$

- Mean of F Distribution

$$\frac{\nu_2}{\nu_2-2}$$

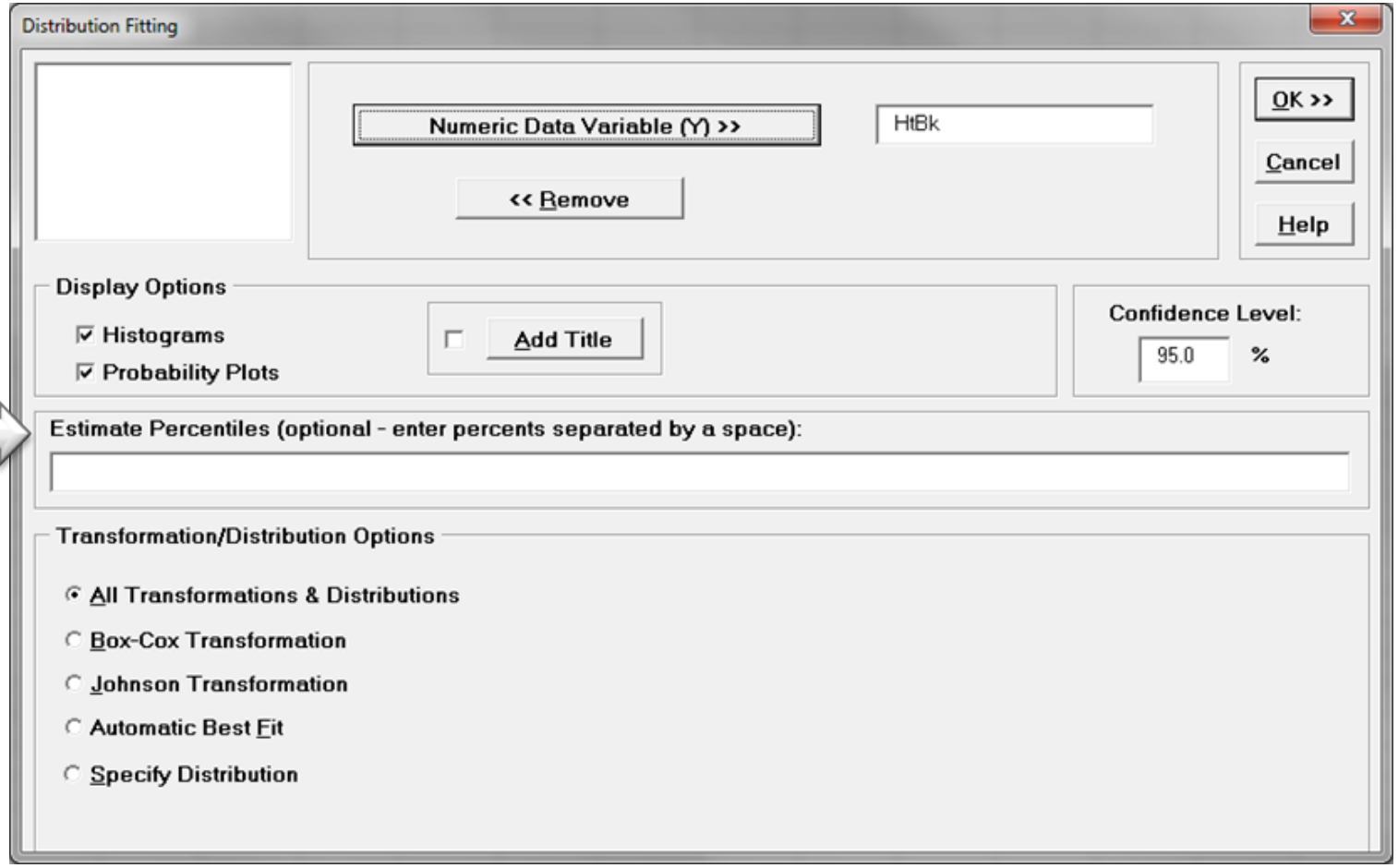
- Variance of F Distribution

$$\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$$

Use SigmaXL to Fit a Distribution

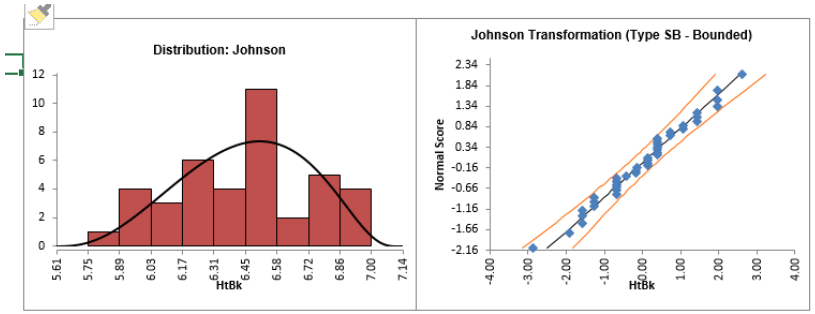
- Case Study: we are interested to fit the height data of basketball players into a distribution.
  - Data File: “One Sample T-Test” tab in “Sample Data.xlsx”
- Steps to fit a distribution in SigmaXL
  - Select the entire range of the variable “HtBk”
  - Click SigmaXL -> Process Capability -> Nonnormal -> Distribution Fitting
  - A new window named “Distribution Fitting” pops up and the selected range appears automatically in the box under “Please select your data”.
  - Click “Next>>”
  - Another window also named “Distribution Fitting” appears
  - Select ‘HtBk” as the “Numeric Data Variable (Y)”
  - Click “OK>>”

Use SigmaXL to Fit a Distribution



Use SigmaXL to Fit a Distribution

- The distribution fitting report is in the newly generated tab “Distribution Fitting (1)”.
- The distributions and transformations are sorted in the descending order of AD p-values.

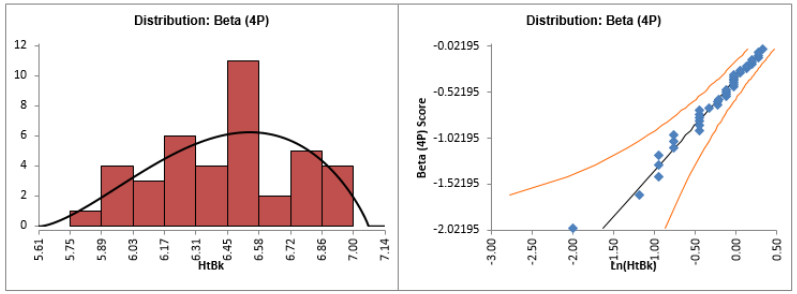


Distribution Fitting Report: HtBk	
Johnson Transformation (Type SB - Bounded)	
Z = Shape1 + Shape2 * Ln((Y - Location) / (Scale + Location - Y))	
Sample Count	40
Sample Mean	6.453

Model Summary and Goodness-of-Fit	
Log-Likelihood	-63.448
AD Johnson	0.306980
AD Johnson P-Value	0.548000

Parameter Estimates				
Parameter	Estimate	SE Estimate	Lower 95.0 CI	Upper 95.0 CI
Shape1	-0.356777			
Shape2	1.266			
Location	5.561			
Scale	1.578			
Mean (Transformed Data)	0.037621358	0.185617	-0.326182	0.401424
StDev (Transformed Data)	1.197	0.133865	0.961526	1.490

Percentile Report				
Percentage	HtBk	SE Percentile	Lower 95.0 CI	Upper 95.0 CI
0.135	5.678		5.622	5.777
50	6.472		6.359	6.579
99.865	7.074		7.015	7.105



Distribution Fitting Report: HtBk	
Distribution: Beta (4P)	
Sample Count	40
Sample Mean	6.453

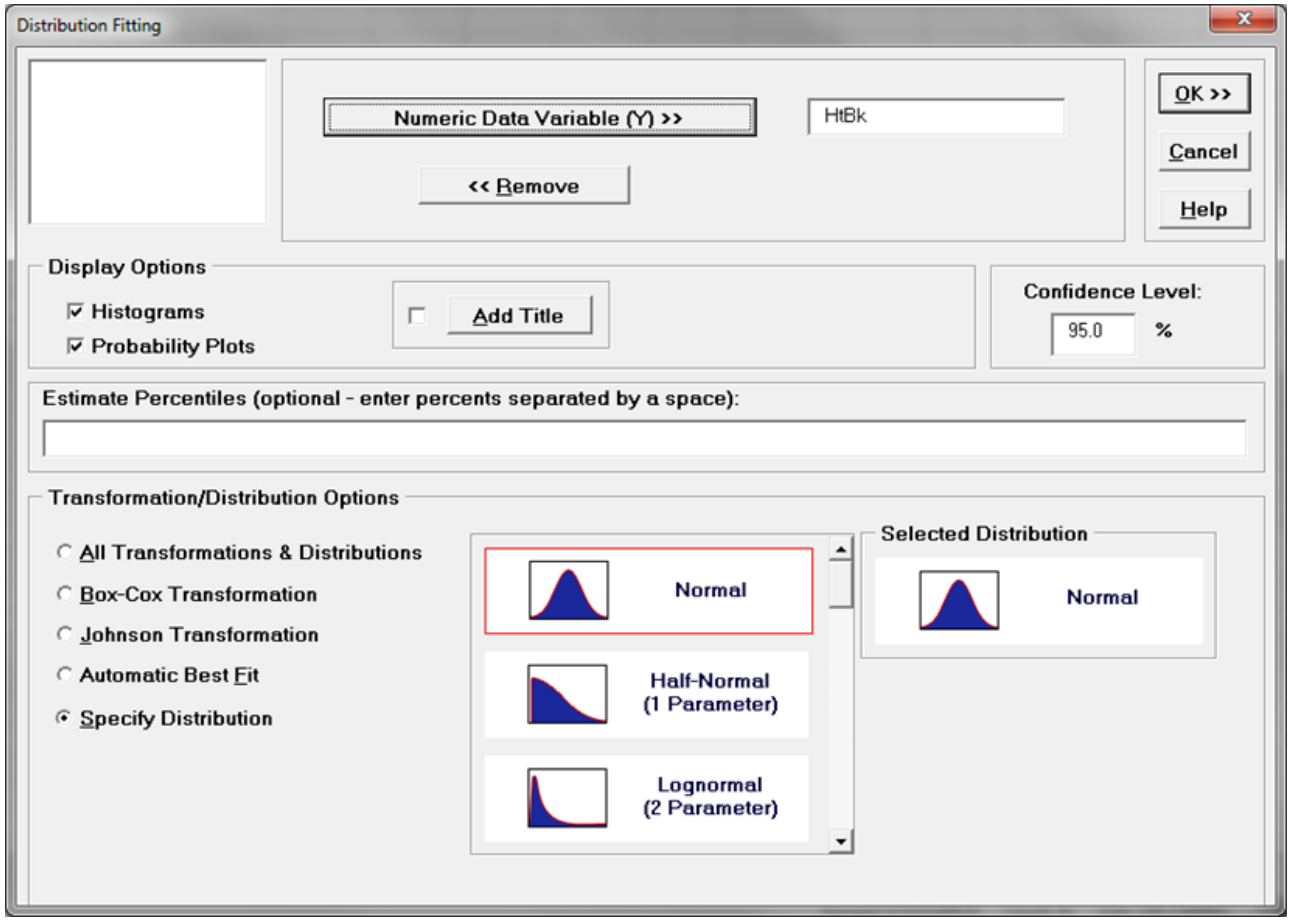
Model Summary and Goodness-of-Fit	
Log-Likelihood	-7.818
AD Beta (4P)	0.310210
AD Beta (4P) P-Value	0.540 (Z-Score Est.)

Parameter Estimates				
Parameter	Estimate	SE Estimate	Lower 95.0 CI	Upper 95.0 CI
Shape1	2.550	0.554064	1.666	3.904
Shape2	1.873	0.393329	1.241	2.827
Upper Threshold (Optimal)	7.068	0.111803	6.849	7.288
Lower Threshold (Optimal)	5.614	0.111803	5.395	5.833

Percentile Report				
Percentage	HtBk	SE Percentile	Lower 95.0 CI	Upper 95.0 CI
0.135	5.685	0.112119	5.465	5.904
50	6.471	0.097709009	6.279	6.662
99.865	7.049	0.111069472	6.831	7.266

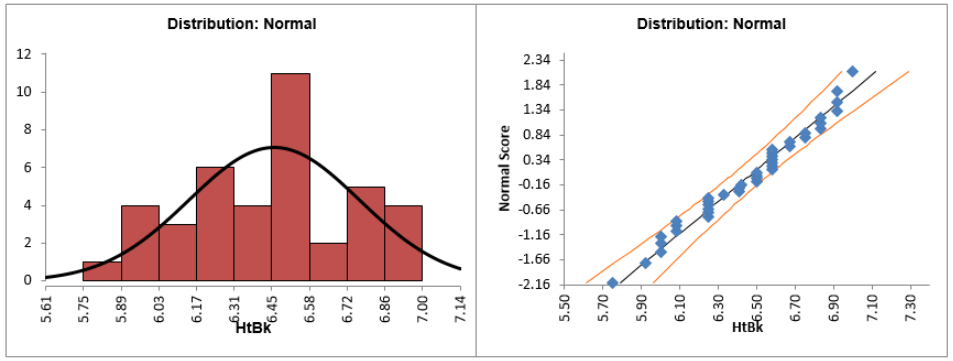
Use SigmaXL to Fit a Distribution

- To fit a specific distribution, in the window “Distribution Fitting”, click on the radio button “Specify Distribution”, select the distribution of interest and click on “OK>>”.
- In this example, we select “Normal” (i.e. normal distribution).



Use SigmaXL to Fit a Distribution

- The result of fitting the specified distribution, normal distribution in this case, is generated in the tab “Dist Fit Normal (1)”.
- Since the p-value is 0.2748 higher than the alpha level, the variable HtBk is normally distributed.



Distribution Fitting Report: HtBk	
Distribution: Normal	
Sample Count	40
Sample Mean	6.453

Model Summary and Goodness-of-Fit	
Log-Likelihood	-9.945
AD Normal	0.441866
AD Normal P-Value	0.2748

Parameter Estimates				
Parameter	Estimate	SE Estimate	Lower 95.0 CI	Upper 95.0 CI
Mean	6.453	0.04905786	6.357098341	6.549401617
StDev	0.314222	0.03513082	0.252388584	0.391203602

Percentile Report				
Percentage	HtBk	SE Percentile	Lower 95.0 CI	Upper 95.0 CI
0.135	5.511	0.116246098	5.282753721	5.738430053
50	6.453	0.04905786	6.357098341	6.549401617
99.865	7.395908071	0.116254004	7.16805441	7.623761732

3.2 Inferential Statistics

Black Belt Training: Analyze Phase

3.1 Patterns of Variation
- 3.1.1 Multi-Vari Analysis
- 3.1.2 Classes of Distributions

3.2 Inferential Statistics
- 3.2.1 Understanding Inference
- 3.2.2 Sampling Techniques and Uses
- 3.2.3 Sample Size
- 3.2.4 Central Limit Theorem

3.3 Hypothesis Testing
- 3.3.1 Goals of Hypothesis Testing
- 3.3.2 Statistical Significance
- 3.3.3 Risk; Alpha and Beta
- 3.3.4 Types of Hypothesis Tests

3.4 Hypothesis Testing: Normal Data
- 3.4.1 One and Two Sample T-Tests
- 3.4.2 One sample variance
- 3.4.3 One Way ANOVA

3.5 Hyp Testing: Non-Normal Data
- 3.5.1 Mann-Whitney
- 3.5.2 Kruskal-Wallis
- 3.5.3 Moods Median
- 3.5.4 Friedman
- 3.5.5 One Sample Sign
- 3.5.6 One Sample Wilcoxon
- 3.5.7 One and Two Sample Proportion
- 3.5.8 Chi-Squared (Contingency Tables)
- 3.5.9 Test of Equal Variances

3.2.1 Understanding Inference

What is Statistical Inference?

- **Statistical inference** is the process of making inferences regarding the characteristics of an unobservable population based on the characteristics of an observed sample.

- We rely on sample data to draw conclusions about the population from which the sample is drawn.

- Statistical inference is widely used since it is difficult or sometimes impossible to collect the entire population data.

Outcome of Statistical Inference

- The outcome or conclusion of statistical inference is a **statistical proposition** about the population.

- Examples of statistical propositions:
  - Estimating a population parameter
  - Identifying an interval or a region where the true population parameter would fall with some certainty
  - Deciding whether to reject a hypothesis made on characteristics of the population of interest
  - Making predictions
  - Clustering or partitioning data into different groups.

Population and Sample

- A **statistical population** is an entire set of objects or observations about which statistical inferences are to be drawn based on its sample.

- It is usually impractical or impossible to obtain the data for the entire population. For example, if we are interested in analyzing the population of all the trees, it is extremely difficult to collect the data for all the trees that existed in the past, exist now, and will exist in the future.

- A **sample** is a subset of the population (like a piece of the pie above). It is necessary for samples to be *representative* of the population.

- The process of selecting a subset of observations within a population is referred to as **sampling**.

Population and Sample

- Population Parameters (Greek letters)
  - Mean: μ
  - Standard deviation: σ
  - Variance: σ^2
  - Median: η

- Sample Statistics (Roman letters)
  - Sample Mean: $\bar{X}$
  - Standard deviation: S
  - Variance: S^2
  - Median: $\tilde{X}$

- The **population parameter** is the numerical summary of a population.
- The **sample statistic** is the numerical measurement calculated based on a sample of that population. It is used to estimate the true population parameter.

Descriptive Statistics vs. Statistical Inference

- **Descriptive Statistics**
  - Descriptive statistics summarize the characteristics of a collection of data.
  - Descriptive statistics are descriptive only and they do not make any generalizations beyond the data at hand.
  - Data used for descriptive statistics are for the purpose of representing or reporting.
- **Statistical Inference**
  - Statistical inference makes generalizations from a sample at hand of a population.
  - Data used for statistical inference are for the purpose of making inferences on the entire population of interest.
- A complete statistical analysis includes *both* descriptive statistics and statistical inference.

Error Sources of Statistical Inference

- Statistical inference uses sample data to best approximate the true features of the population.
- A valid sample must be unbiased and representative of the population.
- Two sources of error in statistical inference:
  - Random sampling error
  - Selection bias.

Error Sources of Statistical Inference

- Random Sampling Error
  - Random variation due to observations being selected randomly
  - It is inherent to the sampling process and beyond one's control.

- Selection Bias
  - Non-random variation due to inadequate design of sampling
  - It can be improved by adjusting the sampling size and sampling strategy.

3.2.2 Sampling Techniques

What is Sampling?

- **Sampling** is the process of selecting objects or observations from a population of interest about which we wish to make a statistical inference.

- It is extensively used to collect information about a population, which can represent physical or intangible objects.

Advantages of Sampling

- It is usually impractical or impossible to collect the data of an entire population:
  - High cost
  - Time consuming
  - Unavailability of historical records
  - Dynamic nature of the population.

- Advantages of sampling a representative subset of the population:
  - Lower cost
  - Faster data collection
  - Easier to manipulate.

Uses of Sample

- With valid samples collected, we can draw statistical conclusions about the population of our interest.

- There are two major uses of samples in making statistical inference:
  - *Estimation*: estimating the population parameters using the sample statistics.
  - *Hypothesis testing*: testing a statement about the population characteristics using sample data.

- This module covers sample size calculation for estimation purposes only. Sample size calculations for hypothesis testing purposes will be covered in the Hypothesis Testing module.

Basic Sampling Steps

1) Determine the population of interest
2) Determine the sampling frame
3) Determine the sampling strategy
4) Calculate the sample size
5) Conduct sampling

Population

- **Population** in statistics is the entire set of objects or observations about which we are interested to draw conclusions or make generalizations based on some representative sample data.
- A population can be either physical or intangible.
  - Physical: trees, customers, monitors etc.
  - Intangible: credit score, pass/fail decisions etc.
- A population can be static or dynamic
  - Characteristics of individuals are relatively static over time
  - Items making up the population continue to change or be generated over time
- The population covers all the items with characteristics we are interested to analyze.

Sampling Frame

- In the ideal situation, the scope of a population could be defined.
  - Example: Mary is interested to know how the soup she is cooking tastes. The population is simply the pot of soup.
- However, in some other situations, the population cannot be identified or defined precisely.
  - Example: to collect the information for an opinion poll, we do not have a list of all the people in the world at hand.

- A **sampling frame** is a list of items of the population (preferably the entire population; if not, approximate population).
  - Example: The telephone directory would be a sampling frame for opinion poll data collection.

Basic Sampling Strategies

- Simple Random Sampling
- Matched Random Sampling
- Stratified Sampling
- Systematic Sampling
- Proportional to Size Sampling
- Cluster Sampling

Simple Random Sampling

- **Simple random samples** are selected in such a way that each item in the population has an equal chance of being selected.

- There is no bias involved in the sample selection. Such selection minimizes the variation between the characteristics of samples and the population.

- It is the basic sampling strategy.

Matched Random Sampling

- **Matched random samples** are samples randomly selected in pairs, each of which has the same attribute.

- Example:
  - Researchers are interested in understanding the weight of twins.
  - Researchers are interested in understanding the patients' blood pressure before and after taking some medicine.

Stratified Sampling

- A population can be grouped or “stratified” into distinct and independent categories. An individual category can be considered as a sub-population. **Stratified samples** are randomly selected in each category of the population.
- The categories can be gender, region, income level etc.
- Stratified sampling requires advanced knowledge of the population characteristics.
- Example: A fruit store wants to measure the quality of all their oranges. They decide to use stratified sampling by region to collect sample data. Since about 40% of their oranges are from California, 40% of the sample is selected from the California oranges sub-population.

Systematic Sampling

- **Systematic samples** are selected at regular intervals based on an ordered list where items in the population are arranged according to a certain criterion.

- Systematic sampling has a random start and then every i^{th} item is selected going forward.

- For example, we are sampling the every 5^{th} unit produced on the production line.

Cluster Sampling

- **Cluster sampling** is a sampling method in which samples are only selected from certain clusters or groups of the population.

- It reduces the cost and time spent on the sampling but bears the risk that the selected clusters are biased.

- For example, selecting samples from the region where researchers are located so that the cost and time spent on travelling is reduced.

Sampling Strategy Decision Factors

- When determining the sampling strategy, we need to consider the following factors:
  - Cost and time constraints
  - Nature of the population of interest
  - Availability of advanced knowledge of the population
  - Accuracy requirement.

Sample Size

- The **sample size** is a critical element that can influence the results of statistical inference.
- The smaller the sample size, the higher the risk that the sample statistic will not reflect the true population parameter.
- The greater the sample size, the more time and money we will spend on collecting the samples.

Sample Size Factors

- Is the variable of interest continuous or discrete?
- How large is the population size?
- How much risk do you want to take regarding missing the true population parameters?
- What is the acceptable margin of error you want to detect?
- How much is the variation in the population?

Sample Size Calculation for Continuous Data

- Sample size equation for continuous data

$$n_0 = \left(\frac{Z_{\alpha/2} \times s}{d} \right)^2$$

where

n_0 is the number of samples.

$Z_{\alpha/2}$ is the Z score when risk level is $\alpha/2$.

- When α is 0.05, it is 1.96.
- When α is 0.10, it is 1.65.

s is the estimation of standard deviation in the population

d is the acceptable margin of error.

Sample Size Calculation for Continuous Data

- When the sample size calculated using the formula

$$n_0 = \left(\frac{Z_{\alpha/2} \times s}{d} \right)^2$$

exceeds 5% of the population size, we use a correction formula to calculate the final sample size.

$$n = \frac{n_0}{\left(1 + \frac{n_0}{N}\right)}$$

where
n_0 is the sample size calculated using equation $n_0 = \left(\frac{Z_{\alpha/2} \times s}{d} \right)^2$
N is the population size.

Sample Size Calculation for Discrete Data

- Sample size equation for discrete data

$$n_0 = \left(\frac{Z_{\alpha/2}}{d} \right)^2 \times p \times (1 - p)$$

where

n_0 is the number of samples.

$Z_{\alpha/2}$ is the Z score when risk level is $\alpha/2$.

- When α is 0.05, it is 1.96.
- When α is 0.10, it is 1.65.

s is the estimation of standard deviation in the population.

d is the acceptable margin of error.

p is the proportion of one type of event occurring (e.g., proportion of passes).

$p \times (1 - p)$ is the estimate of variance.

Sample Size Calculation for Discrete Data

- When the sample size calculated using the formula

$$n_0 = \left(\frac{Z_{\alpha/2}}{d} \right)^2 \times p \times (1-p)$$

exceeds 5% of the population size, we use a correction formula to calculate the final sample size.

$$n = \frac{n_0}{\left(1 + \frac{n_0}{N}\right)}$$

where
n_0 is the sample size calculated using equation $n_0 = \left(\frac{Z_{\alpha/2}}{d} \right)^2 \times p \times (1-p)$
N is the population size.

Sampling Errors

- Random Sampling Error
  - Random variation due to observations being selected randomly
  - It is inherent in the sampling process and beyond one's control.

- Selection Bias
  - Non-random variation due to inadequate design of sampling
  - It can be improved by adjusting the sampling size and sampling strategy.

3.2.3 Sample Size

Sample Size

- The sample size is a critical element that can influence the results of hypothesis testing.
- The smaller the sample size, the higher the risk that the statistical conclusions will not reflect the population relationship.
- The greater the sample size, the more time and money we will spend on collecting the samples.

Sample Size Calculation

- General sample size formula for *continuous data*

$$n = \left(\frac{\left(Z_{\alpha/2} + Z_{\beta} \right) \times s}{d} \right)^2$$

- General sample size formula for *discrete data*

$$n = \left(\frac{Z_{\alpha/2} + Z_{\beta}}{d} \right)^2 \times p \times (1 - p)$$

Sample Size Calculation

- *n* is the number of observations in the sample.
- *α* is the risk of committing a false positive error.
- *β* is the risk of committing a false negative error.
- *s* is the estimation of standard deviation in the population
- *d* is the size of effect you want to be able to detect.
- *p* is the proportion of one type of event occurring (e.g., proportion of passes).

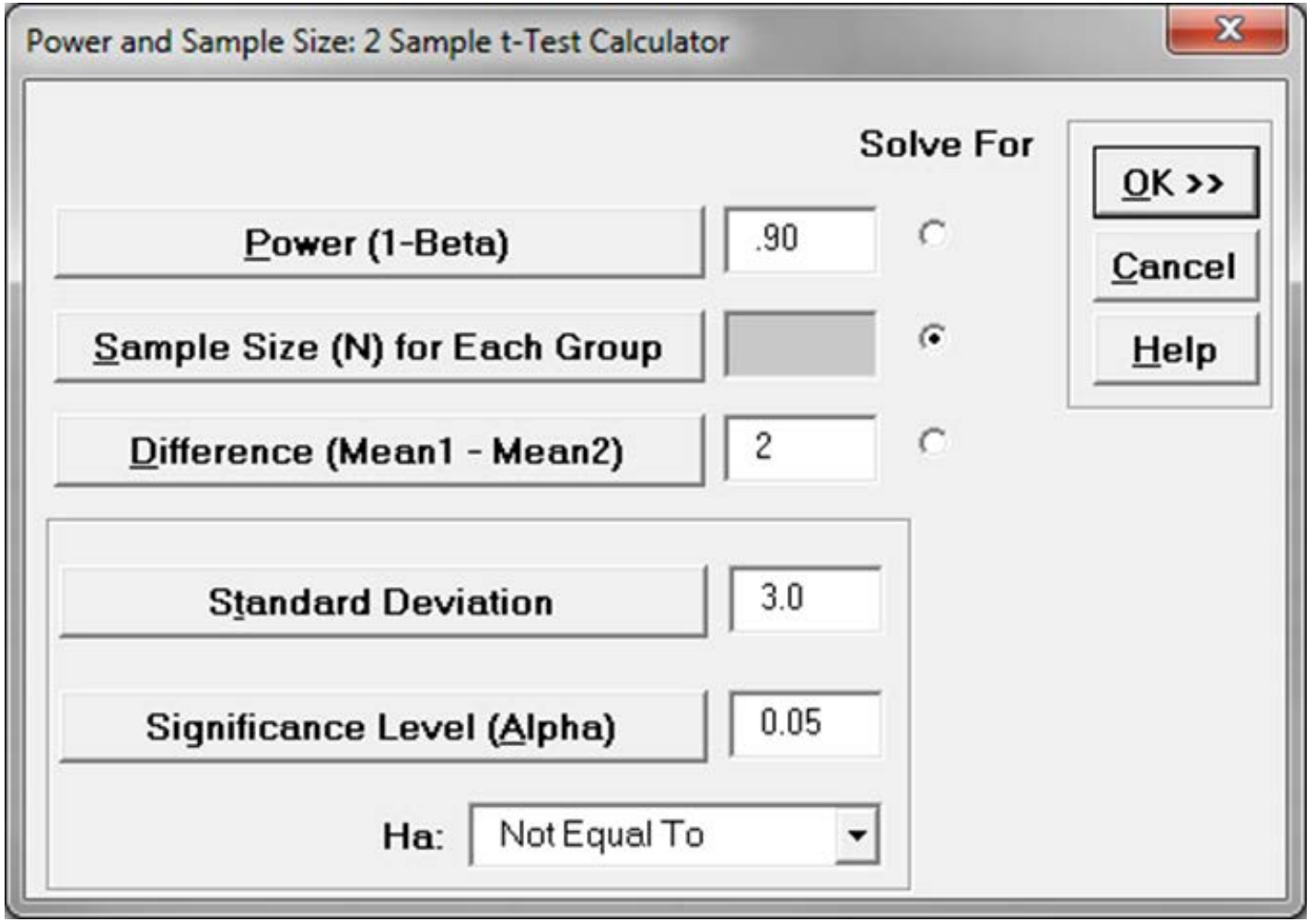
Use SigmaXL to Calculate the Sample Size

- Case Study:
  - We are interested in comparing the average retail price of a product between two states.
  - We will run a hypothesis test on the two sample means to determine whether there is a statistically significant difference between the retail price in the two states.
  - The average retail price of the product is 23 based on our estimation and the standard deviation is 3. We want to detect at least 2 dollars difference with 90% chance when it is true and we can tolerate the alpha risk at 5%
  - **What should the sample size be?**

Use SigmaXL to Calculate the Sample Size

- Steps to calculate the sample size in SigmaXL
  - Click SigmaXL -> Statistical Tools -> Power & Sample Size Calculators -> 2 Sample t-Test Calculator
  - A window named “Power & Sample Size: 2 Sample t-Test Calculator” appears
  - Click the radio button “Solve for Sample Size (N) for Each Group”
  - Enter “0.9” as “Power (1-Beta)”
  - Enter “2” as “Difference (Mean1 – Mean2)”
  - Enter “3” as “Standard Deviation”
  - Enter “0.05” as “Significance Level (Alpha)”
  - Click “OK>>”
  - The sample size calculation results appear in the new tab “PS - 2 Sample t-Test (1)”

Use SigmaXL to Calculate the Sample Size



Use SigmaXL to Calculate the Sample Size

- The sample size for each group is 49 based on the sample size calculator.
- When the difference to detect decreases, the required sample size would increase.

Power and Sample Size: 2 Sample t Test

H0: Mean 1 = Mean 2

Ha: Mean 1 ≠ Mean 2

Solve For: Sample Size (N) for Each Group

Power (1 - Beta)	Difference	Standard Deviation	Significance Level (Alpha)	Sample Size (N)	Actual Power
0.9	2	3	0.05	49	0.904339441

3.2.4 Central Limit Theorem

What is Central Limit Theorem?

- The **Central Limit Theorem** is one of the fundamental theorems of probability theory.

- It states a condition under which the mean of a large number of independent and identically-distributed random variables, each of which has a finite mean and variance, would be approximately normally distributed.

What is Central Limit Theorem?

- Let us assume Y_1, Y_2, …, Y_n is a sequence of *n* i.i.d. random variables, each of which has finite mean μ and variance σ^2, where $\sigma^2 > 0$.

- When *n* increases, the sample average of the *n* random variables is approximately normally distributed, with the mean equal to μ and variance equal to σ^2/n, regardless of the common distribution Y_i follows where $i = 1, 2, \ldots, n$.

Independent and Identically Distributed

- A sequence of random variables is **independent and identically distributed** (i.i.d.) if each random variable is independent of others and has the same probability distribution as others.

- It is one of the basic assumptions in Central Limit Theorem.

Central Limit Theorem Example

- Let us assume we have 10 fair die at hand.
- Each time we roll all 10 die together we record the average of the 10 die.
- We repeat rolling the die 50 times until we will have 50 data points.
- Upon doing so, we will discover that the probability distribution of the sample average approximates the normal distribution even though a single roll of a fair die follows a discrete uniform distribution.

Central Limit Theorem Explained in Formulas

- Let us assume $Y_1, Y_2, \ldots, Y_n$ are i.i.d. random variables with

$$E(Y_i) = \mu_Y \quad \text{where} \quad -\infty < \mu_Y < \infty$$

$$\text{var}(Y_i) = \sigma_Y^2 \quad \text{where} \quad 0 < \sigma_Y^2 < \infty$$

- As $n \to \infty$, the distribution of $\overline{Y}$ becomes approximately normally distributed with

$$E(\overline{Y}) = \mu_Y$$

$$\text{var}(\overline{Y}) = \sigma_{\overline{Y}}^2 = \frac{\sigma_Y^2}{n}$$

Central Limit Theorem Application

- Use the sample mean to estimate the population mean.

- If the assumptions of Central Limit Theorem are met,

$$E(Y_i) = E(\overline{Y}) \quad \text{where} \quad i = 1,2,\ldots,n$$

Central Limit Theorem Application

- Use standard error of the mean to measure the standard deviation of the sample mean estimate of a population mean.

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

where s is the standard deviation of the sample and n is the sample size.

- Standard deviation of the population mean

$$SD_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and n is the sample size.

Central Limit Theorem Application

- Use a larger sample size, if economically feasible, to decrease the variance of the sampling distribution.
- The larger the sample size, the more precise the estimation of the population parameter.
- Use a confidence interval to describe the region which the population parameter would fall in.
- The sample distribution approximates the normal distribution in which 95% of the data stays within two standard deviations from the center.
- Population mean would fall in the interval of two standard errors of the mean away from the sample mean, 95% of the time.

Confidence Interval

- The **confidence interval** is an interval where the true population parameter would fall within a certain confidence level.
- A 95% confidence interval indicates that the population parameter would fall in that region 95% of the time or we are 95% confident that the population parameter would fall in that region.
- 95% is the most commonly used confidence level.
- Confidence interval is used to describe the reliability of a statistical estimate of a population parameter.

Confidence Interval

- The width of a confidence interval depends on:
  - Confidence level
  - Sample size
  - Variability in the data.
- The higher the confidence level, the wider the confidence interval.
- The smaller the sample size, the wider the confidence interval.
- The more variability, the wider the confidence interval.

Confidence Interval of the Mean

- Confidence interval of the population mean μ_Y of a continuous variable Y is

$$\left[\overline{Y} - Z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right), \overline{Y} + Z_{\alpha/2} \times \left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

where

$\overline{Y}$ is the sample mean

σ is the standard deviation of the population

n is the sample size

$Z_{\alpha/2}$ is the Z score when risk level is *α/2*.

- When *α* is 0.05, it is 1.96.
- When *α* is 0.05, it is 1.65.

α is (1 – confidence level).

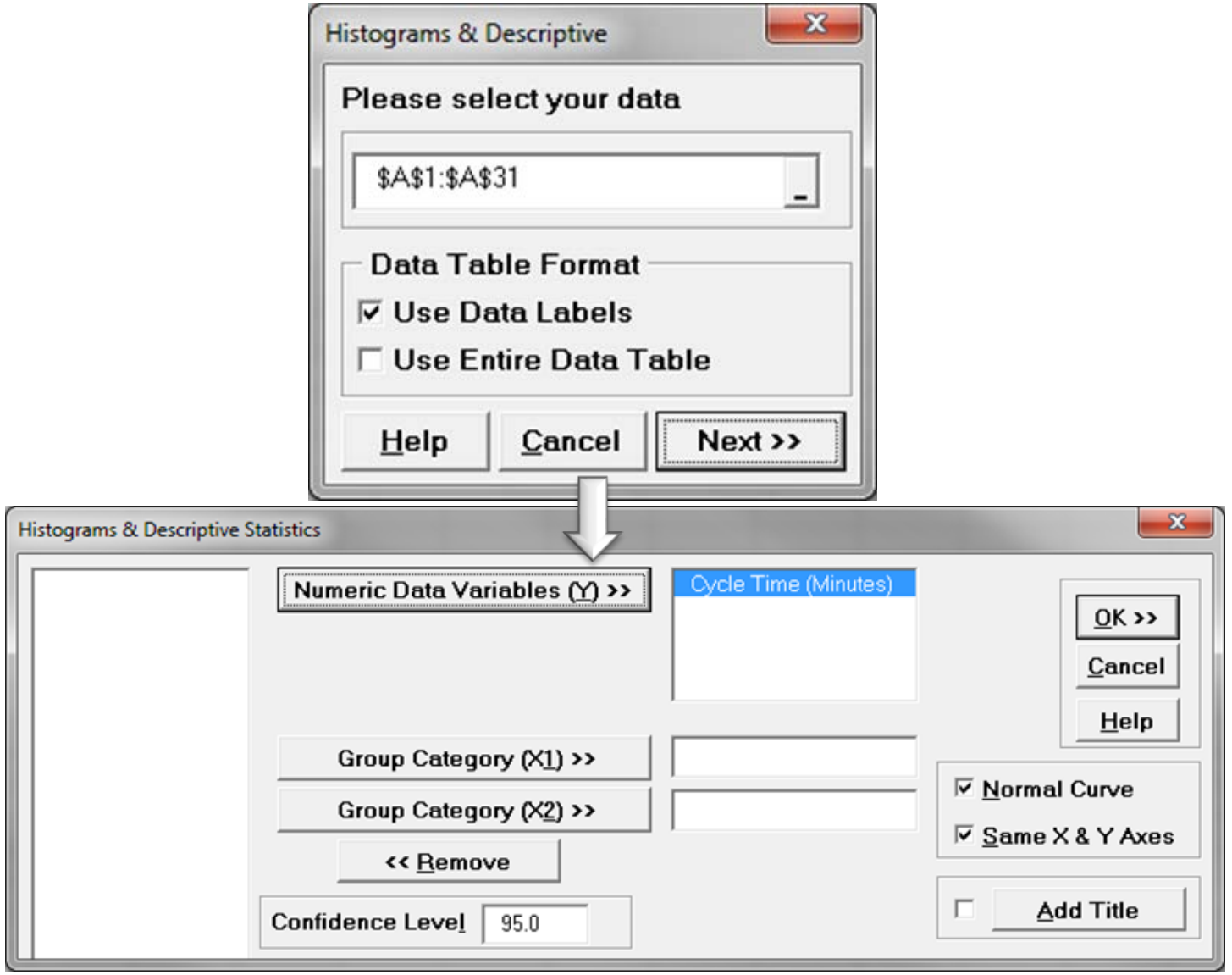
When confidence level is 95%, *α* is 5%. When the confidence level is 90%, *α* is 10%.

Note: Since 95% is the most commonly used confidence level, 0.05 is the most commonly used α (also called alpha level).

SigmaXL: Calculate the Confidence Interval of the Mean

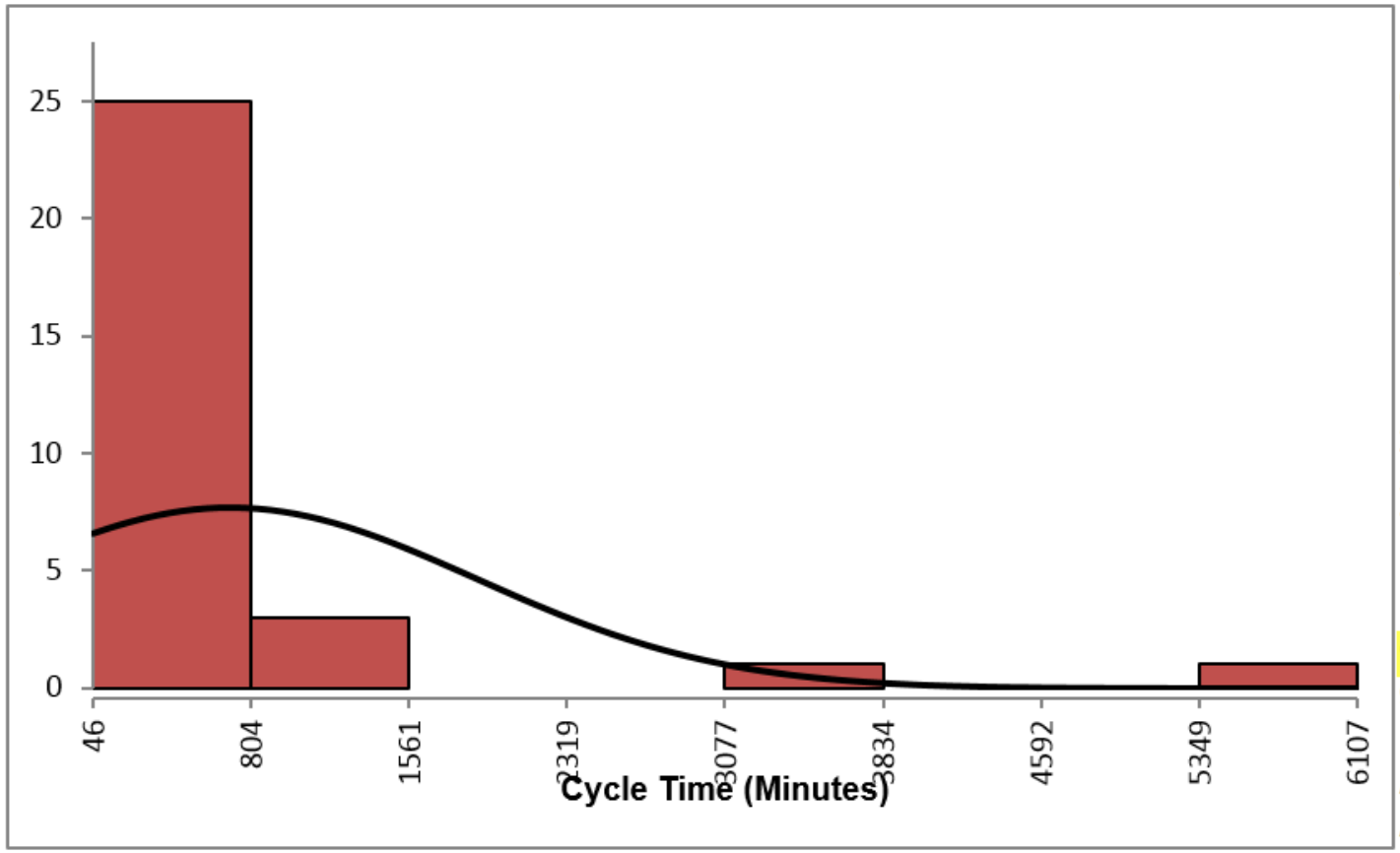
- Data File:
  - “Central Limit Theorem” tab in “Sample Data.xlsx”
- Step 1:
  - Select the entire range of “Cycle Time (Minutes)”
  - Click SigmaXL -> Graphical Tools -> Histogram & Descriptive Statistics
  - A new window named “Histogram & Descriptive” pops up with the selected range automatically appearing in the box under “Please select your data”
  - Click “Next”
  - Another window named “Histogram & Descriptive Statistics” pops up.
  - Select “Cycle Time (Minutes)” as the “Numeric Data Variable (Y)”
  - Click “OK>>”

SigmaXL: Calculate the Confidence Interval of the Mean



SigmaXL: Calculate the Confidence Interval of the Mean

- The 95% confidence interval of the mean is shown in the newly generated ta “Hist Descript (1)”



Cycle Time (Minutes)

Count = 30
Mean = 703.17
Stdev = 1181.0
Range = 6055.00

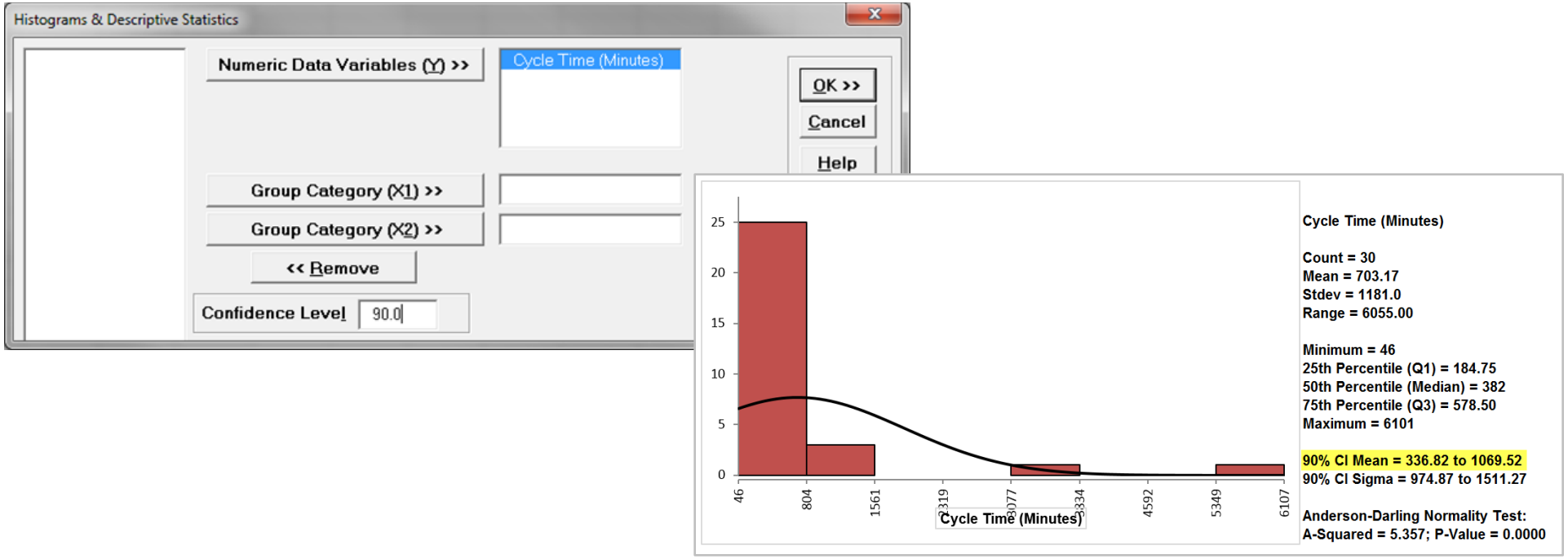
Minimum = 46
25th Percentile (Q1) = 184.75
50th Percentile (Median) = 382
75th Percentile (Q3) = 578.50
Maximum = 6101

95% CI Mean = 262.19 to 1144.14
95% CI Sigma = 940.52 to 1587.57

Anderson-Darling Normality Test:
A-Squared = 5.357; P-Value = 0.0000

SigmaXL: Calculate the Confidence Interval of the Mean

- In SigmaXL, the confidence level is 95% by default.
- In order to see the confidence interval of “Cycle Time (Minutes)” at other confidence levels, we need to enter the confidence level of our interest in the window “Histogram and Descriptive Statistics” and click “OK>>”
- Below shows how to generate 90% confidence interval of the mean.



3.3 Hypothesis Testing

Black Belt Training: Analyze Phase

3.1 Patterns of Variation
- 3.1.1 Multi-Vari Analysis
- 3.1.2 Classes of Distributions

3.2 Inferential Statistics
- 3.2.1 Understanding Inference
- 3.2.2 Sampling Techniques and Uses
- 3.2.3 Sample Size
- 3.2.4 Central Limit Theorem

3.3 Hypothesis Testing
- 3.3.1 Goals of Hypothesis Testing
- 3.3.2 Statistical Significance
- 3.3.3 Risk; Alpha and Beta
- 3.3.4 Types of Hypothesis Tests

3.4 Hypothesis Testing: Normal Data
- 3.4.1 One and Two Sample T-Tests
- 3.4.2 One sample variance
- 3.4.3 One Way ANOVA

3.5 Hyp Testing: Non-Normal Data
- 3.5.1 Mann-Whitney
- 3.5.2 Kruskal-Wallis
- 3.5.3 Moods Median
- 3.5.4 Friedman
- 3.5.5 One Sample Sign
- 3.5.6 One Sample Wilcoxon
- 3.5.7 One and Two Sample Proportion
- 3.5.8 Chi-Squared (Contingency Tables)
- 3.5.9 Test of Equal Variances

3.3.1 Goals of Hypothesis Testing

What is Hypothesis Testing?

- A **hypothesis test** is a statistical method in which a specific hypothesis is formulated about a population, and the decision of whether to reject the hypothesis is made based on sample data.

- Hypothesis tests help to determine whether a hypothesis about a population or multiple populations is true with certain confidence level based on sample data.

Hypothesis Testing Examples

- Hypothesis testing tries to answer whether there is a difference between different groups or there is some change occurring.
  - Are the average SAT scores of graduates from high school A and B the same?
  - Is the error rate of one group of operators higher than that of another group?
  - Are there any non-random causes influencing the height of kids in one state?

What is Statistical Hypothesis?

- A **statistical hypothesis** is an assumption about one or multiple population.
- It is a statement about whether there is any difference between different groups.
- It can be a conjecture about the population parameters or the nature of the population distributions.
- A statistical hypothesis is formulated in pairs:
  - Null Hypothesis
  - Alternative Hypothesis.

Null and Alternative Hypotheses

- *Null Hypothesis* (H_0) states that:
  - there is no difference in the measurement of different groups
  - no changes occurred
  - sample observations result from random chance.

- *Alternative Hypothesis* (H_1 or H_a) states that:
  - there is a difference in the measurement of different groups
  - some changes occurred
  - sample observations are affected by non-random causes.

Null and Alternative Hypotheses

- A statistical hypothesis can be expressed in mathematical language by using population parameters (Greek letters) and mathematical symbols.

- Population Parameters (Greek letters)
  - Mean: μ
  - Standard deviation: σ
  - Variance: σ^2
  - Median: η

- Mathematical Symbols
  - Equal: =
  - Not equal: ≠
  - Greater than: >
  - Smaller than: <

Null and Alternative Hypotheses

- Examples of null and alternative hypotheses written in mathematical language.

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_1: \mu_1 \neq \mu_2 \end{cases}$$

$$\begin{cases} H_0: \sigma_1 = 0 \\ H_1: \sigma_1 \neq 0 \end{cases}$$

$$\begin{cases} H_0: \eta_1 = 10 \\ H_1: \eta_1 > 10 \end{cases}$$

$$\begin{cases} H_0: \mu_1 = 10 \\ H_1: \mu_1 < 10 \end{cases}$$

Hypothesis Testing Conclusion

- There are two possible conclusions of hypothesis testing:
  - Reject the null
  - Fail to reject the null.

- When there is enough evidence based on the sample information to prove the alternative hypothesis, we reject the null.

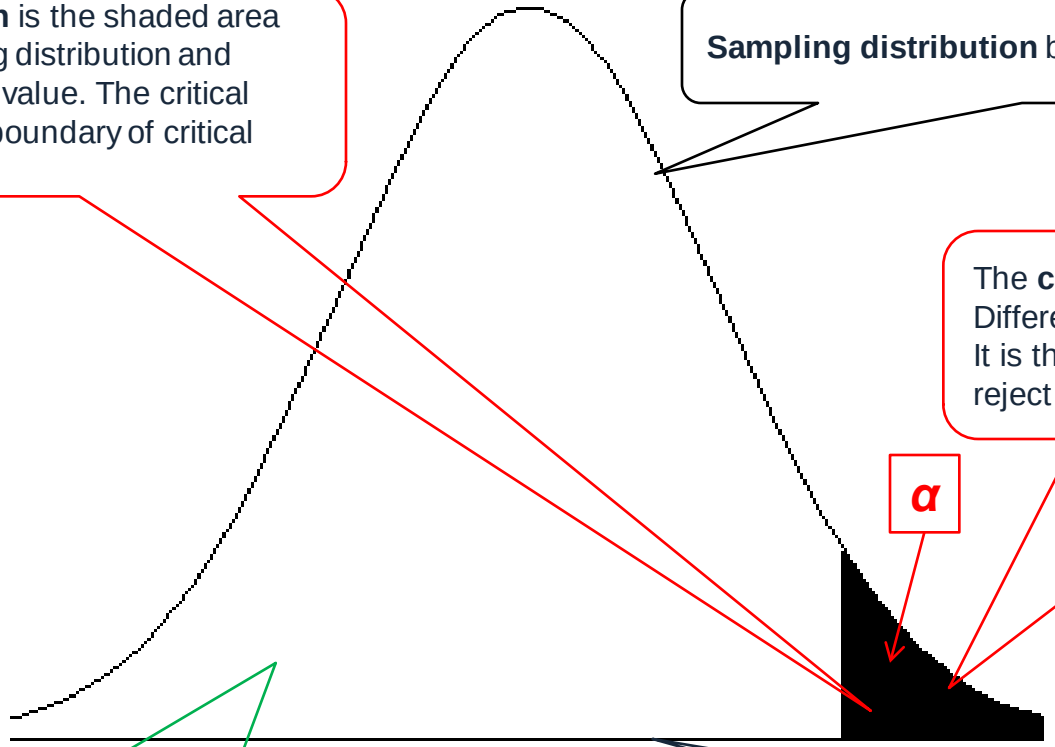
- When there is *not* enough evidence or the sample information is *not* sufficiently persuasive, we fail to reject the null.

Decision Rules in Hypothesis Testing

The **critical region** is the shaded area under the sampling distribution and beyond the critical value. The critical value defines the boundary of critical region.

Sampling distribution based on the null hypothesis

The **critical value** is derived from statistical tables. Different hypotheses use different statistical tables. It is the threshold value to determine whether to reject the null.



The **acceptance region** is the non-shaded area under the sampling distribution and not beyond the critical value. It is the opposite area of critical region.

The **test statistic** is calculated based on a function of sample data. Different hypotheses use different formulas to calculate test statistic. It can be any value along the x axis.

Decision Rules in Hypothesis Testing

- The **test statistic** in hypothesis testing is a value calculated using a function of the sample.
- Test statistics are considered the sample data's numerical summary that can be used in hypothesis testing.
- Different hypothesis tests have different formulas to calculate the test statistic.
- The **critical value** in hypothesis testing is a threshold value to which the test statistic is compared in order to determine whether the null hypothesis is rejected.
- The critical value is obtained from statistical tables.
- Different hypothesis tests need different statistical tables for critical values.

Decision Rules in Hypothesis Testing

- When the test statistic falls into the acceptance region, we fail to reject the null and claim that there is no statistically significant difference between the groups.

- When the test statistic falls into the critical region, we reject the null and claim that there is a statistically significant difference between the groups.

Decision Rules in Hypothesis Testing

- The proportion of the area under the sampling distribution and beyond the critical value indicates α risk (also called *α level*). The most commonly selected α level is 5%.

- The proportion of the area under the sampling distribution and beyond the test statistic is the *p-value*. It is the probability of getting a test statistic at least as extreme as the observed one, given the null is true.

Decision Rules in Hypothesis Testing

- When the p-value is smaller than the α level, we reject the null and claim that there is a statistically significant difference between different groups.

- When the p-value is higher than the α level, we fail to reject the null and claim that there is no statistically significant difference between different groups.

Steps in Hypothesis Testing

- Step 1: State the null and alternative hypothesis.
- Step 2: Determine α level.
- Step 3: Collect sample data.
- Step 4: Select a proper hypothesis test.
- Step 5: Run the hypothesis test.
- Step 6: Determine whether to reject the null.

3.3.2 Statistical Significance

Statistical Significance

- In statistics, an observed difference is *statistically significant* if it is unlikely that the difference occurred by pure chance, given a predetermined probability threshold.

- Statistical significance indicates that there are some non-random factors causing the result to take place.

- The statistical significance level in hypothesis testing indicates the amount of evidence which is sufficiently persuasive to prove that a difference between groups exists not due to random chance alone.

Practical Significance

- An observed difference is *practically significant* when it is large enough to make a practical difference.
- A difference between groups that is *statistically significant* might not be large enough to be practically significant.
- In some business situations, statistical differences can have little to no meaning because the difference is not large enough to be practical for a business to act upon.

Example

- You started to use the premium gas recently, which was supposed to make your car run better.
- After running a controlled experiment to measure the performance of the car before and after using the premium gas, you performed a statistical hypothesis test and found that the difference before and after was statistically significant.
- Using premium gas did improve the performance.
- However, due to the high cost of the premium gas, you decided that the difference was not large enough to make you pay extra money for it. In other words, the difference is not practically significant.

3.3.3 Risk; Alpha & Beta

Errors in Hypothesis Testing

- In statistical hypothesis testing, there are two types of errors:
  - *Type I Error*
    - a null hypothesis is rejected when it is true in fact.
  - *Type II Error*
    - a null hypothesis is not rejected when it is not true in fact.

	Null hypothesis is true	Alternative hypothesis is true
Fail to reject null hypothesis	Correct	Incorrect (Type II Error)
Reject null hypothesis	Incorrect (Type I Error)	Correct

Type I Error

- Type I error is also called false positive, false alarm, or alpha (*α)* error.
- Type I error is associated with the risk of accepting false positives.
- It occurs when we think there is a difference between groups but in fact there is none.
- Example: telling a patient he is sick and in fact he is not.

Alpha (α)

- *α* indicates the probability of making a type I error. It ranges from 0 to 1.
- *α* risk is the risk of making a type I error.
- *5%* is the most commonly used *α.*
- (*100% – α*) is the confidence level which is used to calculate the confidence intervals.
- When making a decision on whether to reject the null, we compare the p-value against *α:*
  - If p-value is smaller than *α*, we reject the null
  - If p-value is greater than *α*, we fail to reject the null.
- To reduce the *α* risk, we decrease the *α* value to which the p-value is compared.

Type II Error

- Type II error is also called false negative, oversight, or beta (β) error.
- Type II error is associated with the risk of accepting false negatives.
- It occurs when we think there is not any difference between groups but in fact there is.
- Example: telling a patient he is not sick and in fact he is.

Beta (β)

- β indicates the probability of making a type II error. It ranges from 0 to 1.
- β risk is the risk of making a type II error.
- *10%* is the most commonly used β.
- (*100%* – β) is called *power,* which denotes the probability of detecting a difference between groups when in fact the difference truly exists.
- To reduce the β risk, we increase the sample size.
- When holding other factors constant, β is inversely related to *α.*

3.3.4 Types of Hypothesis Tests

Two Types of Hypothesis Tests

- **Two-tailed hypothesis test**
  - It is also called two-sided hypothesis test
  - It is a statistical hypothesis test in which the critical region is split into two equal areas, each of which stays on one side of the sampling distribution.

- **One-tailed hypothesis test**
  - It is also called one-sided hypothesis test
  - It is a statistical hypothesis test in which the critical region is only on one side of the sampling distribution.

Two-Tailed Hypothesis Test

- A two-tailed hypothesis test is used when we care about whether there is a difference between groups and we do not care about the direction of the difference.
- Examples of two-tailed hypothesis tests:

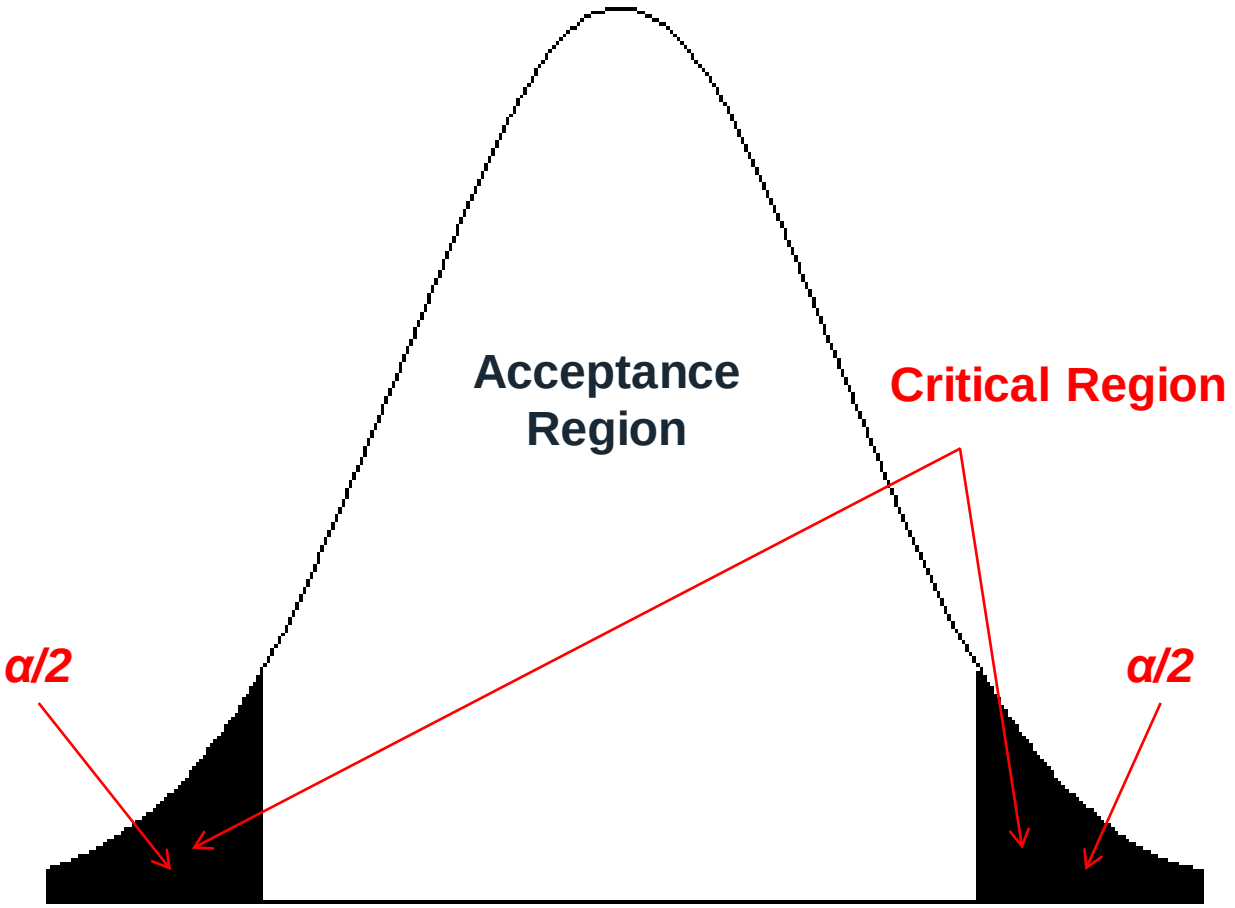
$$\begin{cases} H_0: \mu_1 = 10 \\ H_a: \mu_1 \neq 10 \end{cases}$$

The null hypothesis (H_0) is rejected when:

  - the test statistic falls into either half of the critical region
  - the test statistic is either sufficiently small or sufficiently large
  - the absolute value of the test statistic is greater than the absolute value of the critical value.

- A two-tailed hypothesis test is the most commonly used hypothesis test. In the next modules we will cover more details about it.

Two-Tailed Hypothesis Test



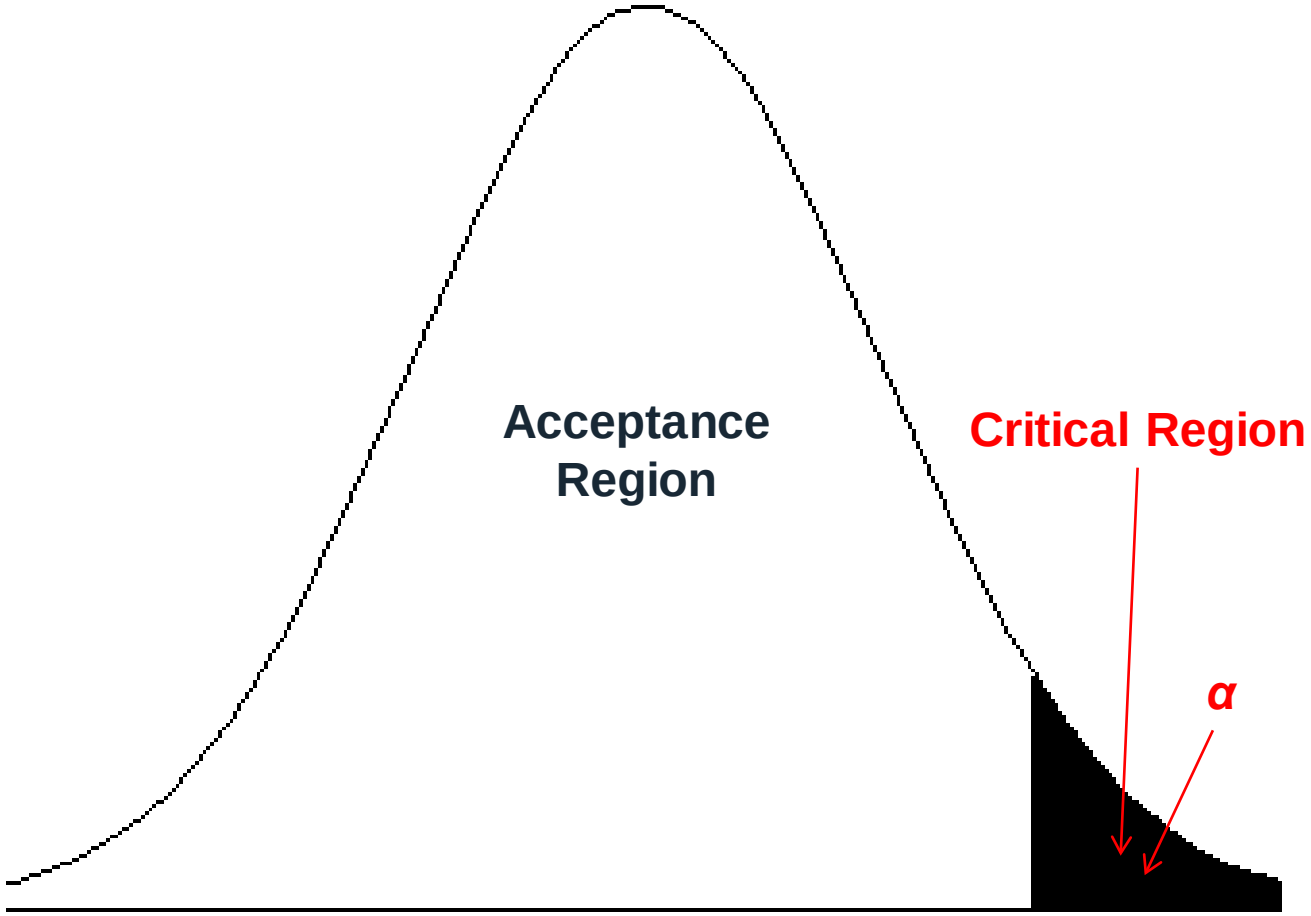
One-Tailed Hypothesis Test

- A one-tailed hypothesis test is used when we care about one direction of the difference between groups.
- Examples of one-tailed hypothesis tests:

$$\begin{cases} H_0: \mu_1 = 10 \\ H_a: \mu_1 > 10 \end{cases}$$

- The null hypothesis is rejected when the test statistic:
  - falls into the critical region which only exits on the right side of the sampling distribution
  - is sufficiently large
  - is greater than the critical value.

One-Tailed Hypothesis Test



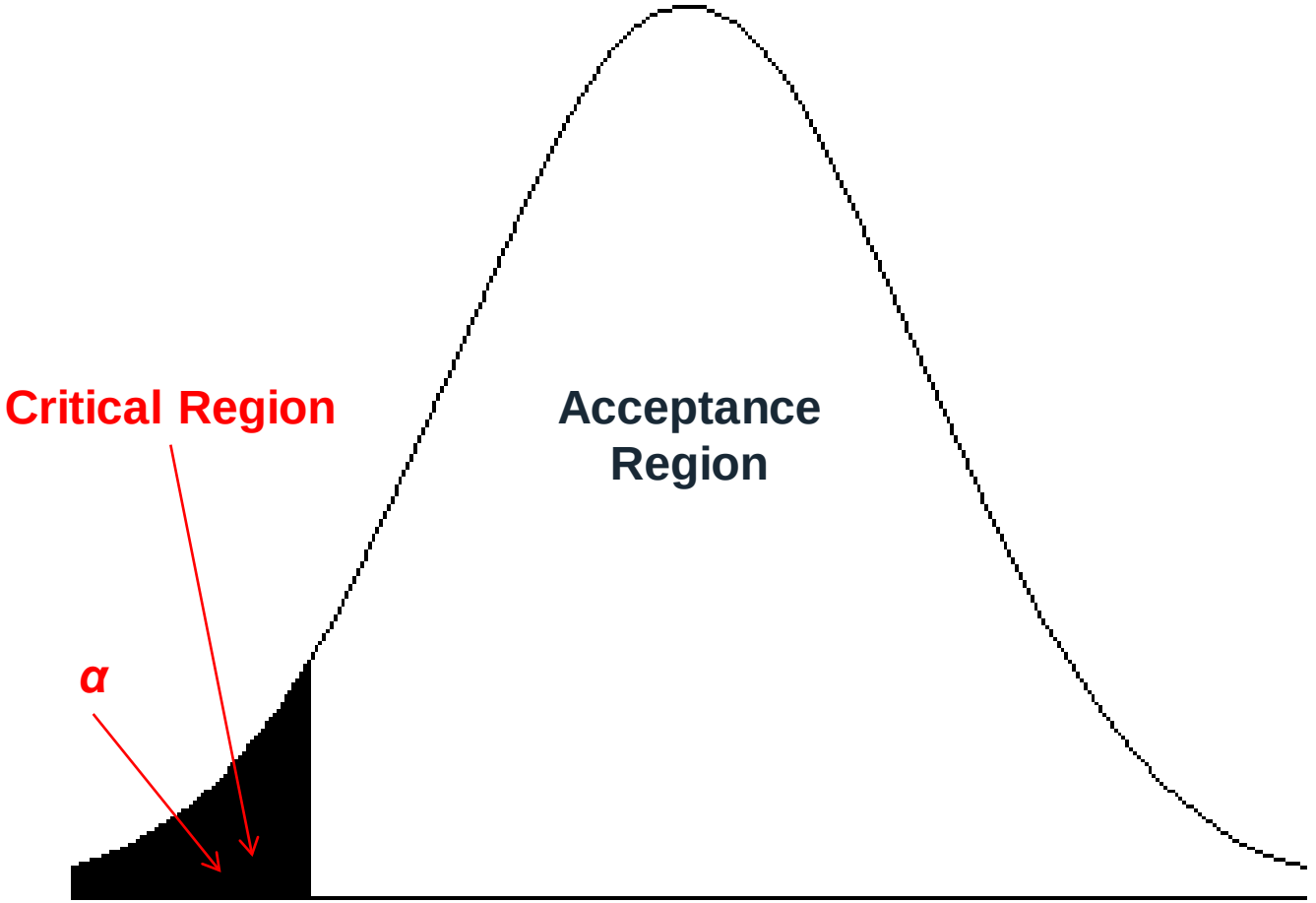
One-Tailed Hypothesis Test

- Examples of one-tailed hypothesis test

$$\begin{cases} H_0: \mu_1 = 10 \\ H_a: \mu_1 < 10 \end{cases}$$

- The null hypothesis is rejected when the test statistic:
  - falls into the critical region which only exits on the left side of the sampling distribution
  - is sufficiently small
  - is smaller than the critical value.

One-Tailed Hypothesis Test



3.4 Hypothesis Tests: Normal Data

Black Belt Training: Analyze Phase

3.1 Patterns of Variation
- 3.1.1 Multi-Vari Analysis
- 3.1.2 Classes of Distributions

3.2 Inferential Statistics
- 3.2.1 Understanding Inference
- 3.2.2 Sampling Techniques and Uses
- 3.2.3 Sample Size
- 3.2.4 Central Limit Theorem

3.3 Hypothesis Testing
- 3.3.1 Goals of Hypothesis Testing
- 3.3.2 Statistical Significance
- 3.3.3 Risk; Alpha and Beta
- 3.3.4 Types of Hypothesis Tests

3.4 Hypothesis Testing: Normal Data
- 3.4.1 One and Two Sample T-Tests
- 3.4.2 One sample variance
- 3.4.3 One Way ANOVA

3.5 Hyp Testing: Non-Normal Data
- 3.5.1 Mann-Whitney
- 3.5.2 Kruskal-Wallis
- 3.5.3 Moods Median
- 3.5.4 Friedman
- 3.5.5 One Sample Sign
- 3.5.6 One Sample Wilcoxon
- 3.5.7 One and Two Sample Proportion
- 3.5.8 Chi-Squared (Contingency Tables)
- 3.5.9 Test of Equal Variances

3.4.1 1 & 2 Sample T-Tests

What is a T-Test?

- In statistics, a **t-test** is a hypothesis test in which the test statistic follows a *Student t* distribution if the null hypothesis is true.

- We apply a t-test when the population variance (σ) is unknown and we use the sample standard deviation (*s*) instead.

What is One Sample T-Test?

- **One sample t-test** is a hypothesis test to study whether there is a statistically significant difference between a population mean and a specified value.

  - Null Hypothesis (H_0): $\mu = \mu_0$
  - Alternative Hypothesis (H_a): $\mu \neq \mu_0$

  where μ is the mean of a population of our interest and μ_0 is the specific value we want to compare against.

Assumptions of One Sample T-Test

- The sample data drawn from the population of interest are unbiased and representative.
- The data of the population are continuous.
- The data of the population are normally distributed.
- The variance of the population of our interest is unknown.
- One sample t-test is more robust than the z-test when the sample size is small (< 30).

Normality Test

- To check whether the population of our interest is normally distributed, we need to run normality test.
  - Null Hypothesis (H_0): The data are normally distributed.
  - Alternative Hypothesis (H_a): The data are not normally distributed.
- There are a lot of normality tests available:
  - Anderson-Darling
  - Sharpiro-Wilk
  - Jarque-Bera etc.

Test Statistic and Critical Value of One Sample T-Test

- Test Statistic

$$t_{calc} = \frac{\overline{Y}}{s/\sqrt{n}}$$, where

$\overline{Y}$ is the sample mean, *n* is the sample size, and s is the sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}$$

- Critical Value
  - t_{crit} is the t-value in a Student t distribution with the predetermined significance level α and degrees of freedom (*n* –1).
  - t_{crit} values for a two-sided and a one-sided hypothesis test with the same significance level α and degrees of freedom (*n* – 1) are different.

Decision Rules of One Sample T-Test

- Based on the sample data, we calculated the test statistic t_{calc}, which is compared against t_{crit} to make a decision of whether to reject the null.
  - Null Hypothesis (H_0): $\mu = \mu_0$
  - Alternative Hypothesis (H_a): $\mu \neq \mu_0$

- If $|t_{calc}| > t_{crit}$, we reject the null and claim there is a statistically significant difference between the population mean μ and the specified value μ_0.

- If $|t_{calc}| < t_{crit}$, we fail to reject the null and claim there is not any statistically significant difference between the population mean μ and the specified value μ_0.

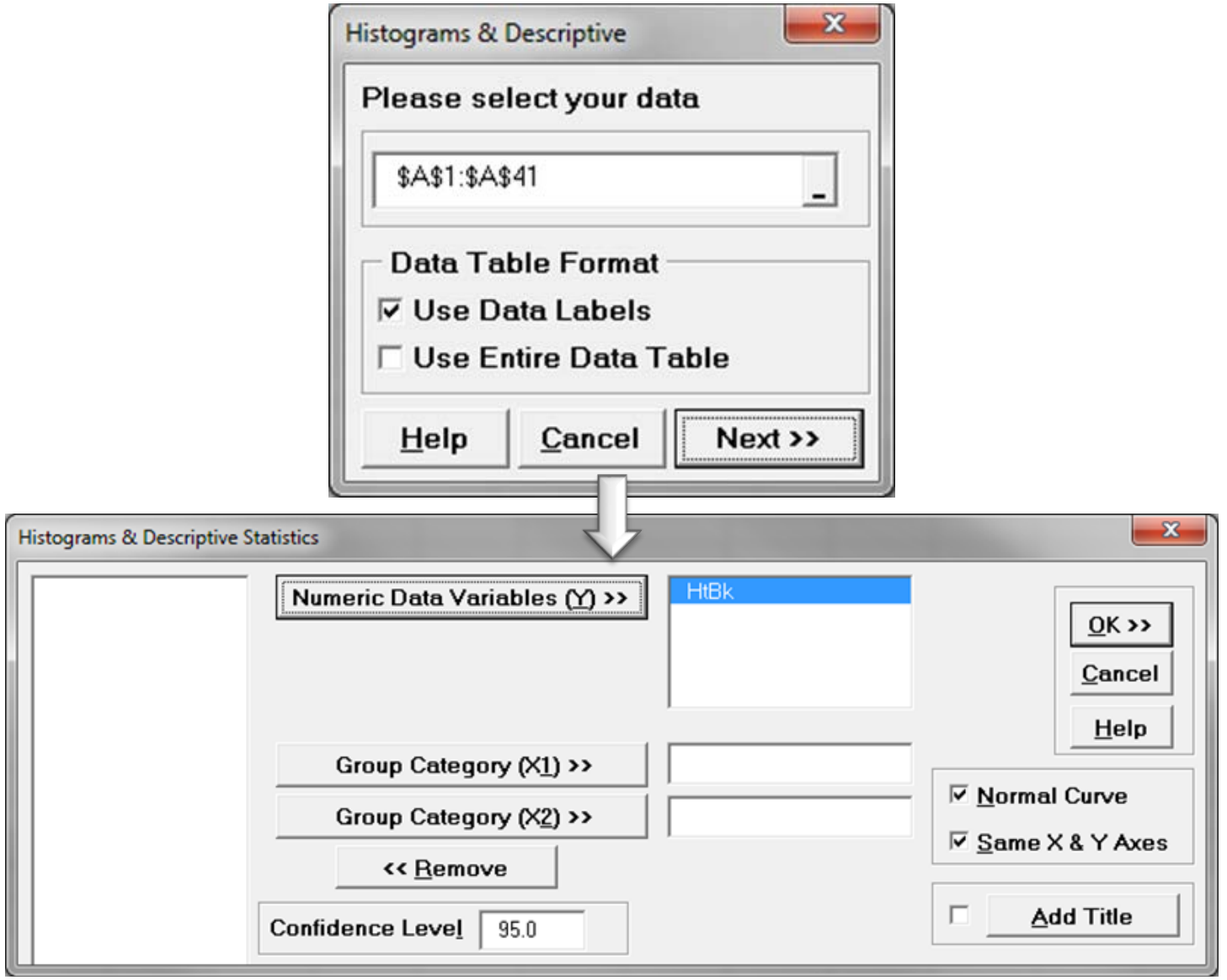
Use SigmaXL to Run a One-Sample T-Test

- Case Study: we are trying to compare the average height of basketball players against 7 feet.
  - Data File: “One Sample T-Test” tab in “Sample Data.xlsx”
- Null Hypothesis (H_0): $\mu = 7$
- Alternative Hypothesis (H_a): $\mu \neq 7$

Use SigmaXL to Run a One-Sample T-Test

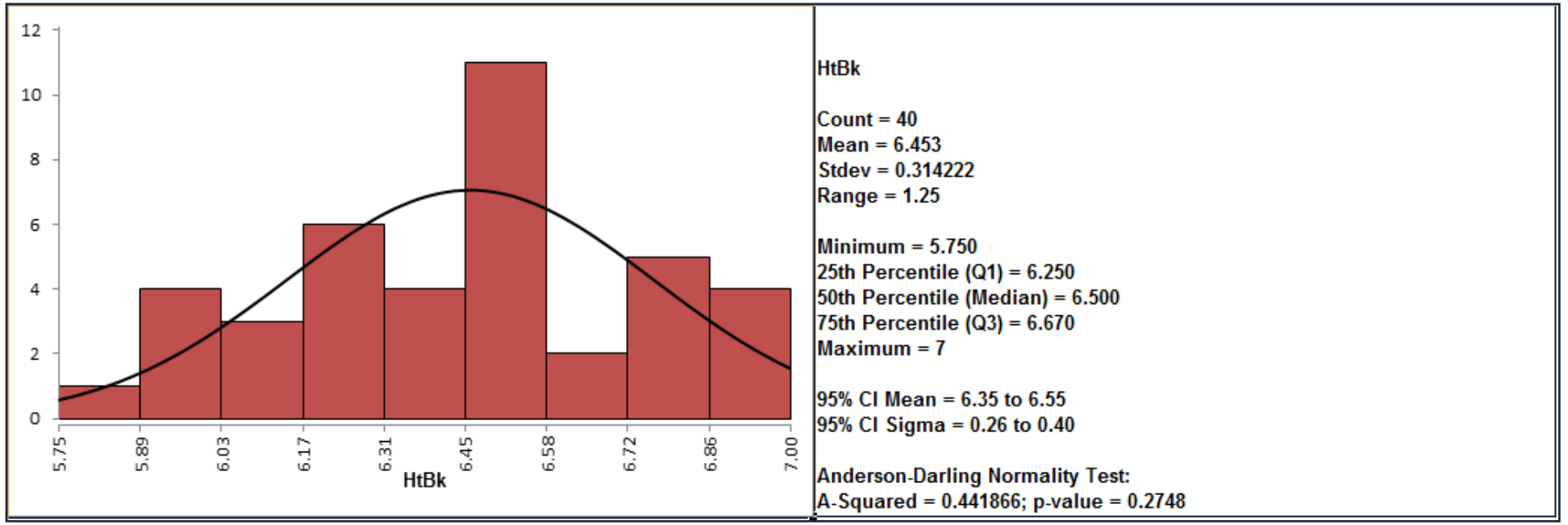
- Step 1: Test whether the data are normally distributed
  - Select the entire range of “HtBk”
  - Click SigmaXL -> Graphical Tools -> Histogram & Descriptive Statistics
  - A new window named “Histogram & Descriptive” pops up with the selected range automatically appearing in the box under “Please select your data”
  - Click “Next >>”
  - Another window named “Histogram & Descriptive Statistics” pops up.
  - Select “HtBk” as the “Numerical Data Variable (Y)”
  - Click “OK”

Use SigmaXL to Run a One-Sample T-Test



Use SigmaXL to Run a One-Sample T-Test

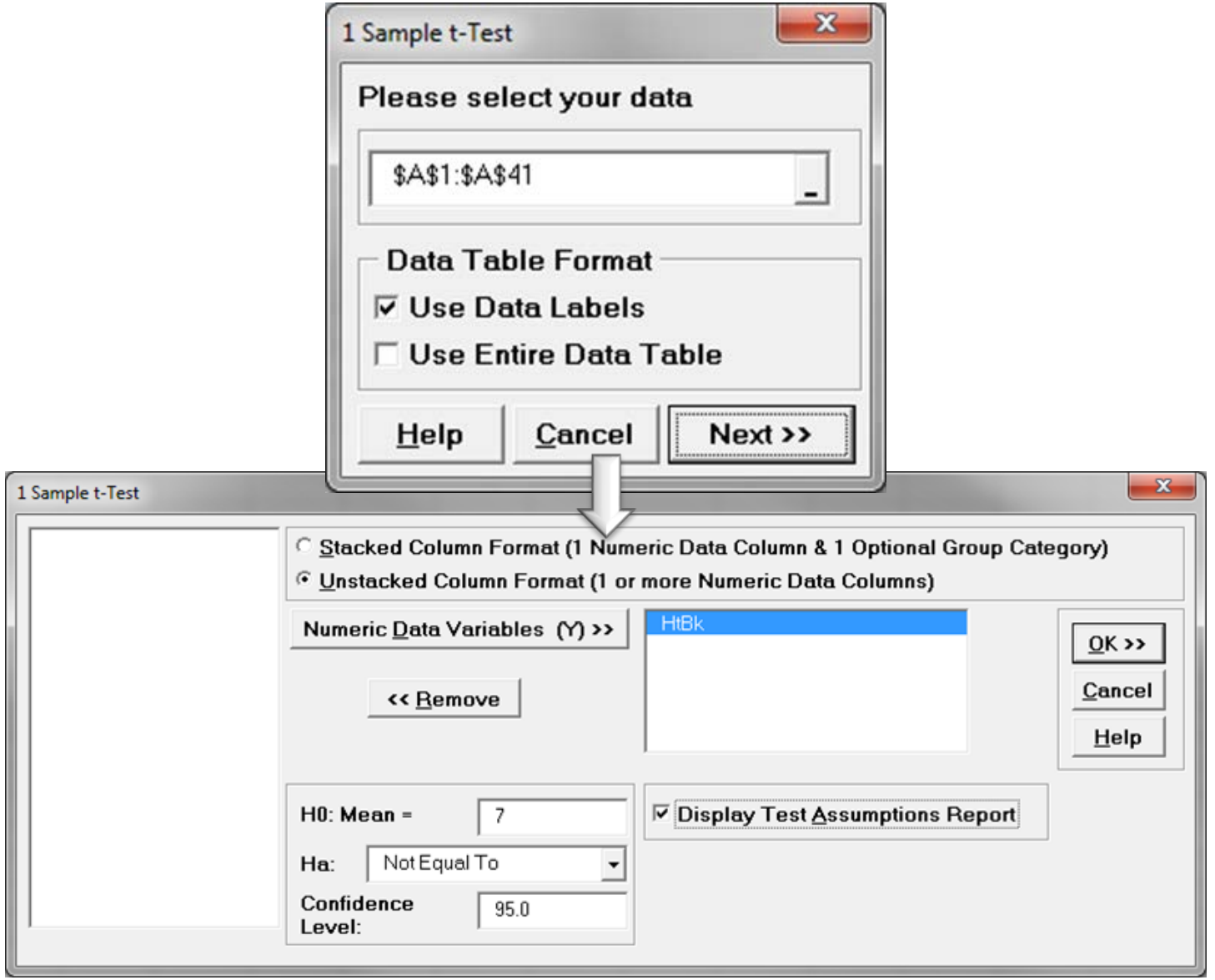
- Null Hypothesis (H_0): The data are normally distributed.
- Alternative Hypothesis (H_a): The data are not normally distributed.
- Since the p-value of the normality is 0.2748 greater than alpha level (0.05), we fail to reject the null and claim that the data are normally distributed.
- If the data are not normally distributed, you need to use other hypothesis tests other than one sample t-test.



Use SigmaXL to Run a One-Sample T-Test

- Step 2: Run the one-sample t-test
  - Select the entire range of “HtBk”
  - Click SigmaXL -> Statistical Tools -> 1 Sample t-Test & Confidence Intervals
  - A new window named “1 Sample t-Test” pops up with the selected range pre-populated in the box under “Please select your data”
  - Click “Next>>”
  - Another window also named “1 Sample t-Test” appears
  - Select “HtBk” as the “Numerical Data Variable (Y)”
  - Enter the hypothesized value “7” into the box next to “H0: Mean =”
  - Select “Not Equal To” in the box next to “Ha:”
  - Click “OK>>”
  - The one-sample t-test result appears automatically in the tab “1 Sample t-Test (1)”.

Use SigmaXL to Run a One-Sample T-Test



Use SigmaXL to Run a One-Sample T-Test

1 Sample t-Test

Test Information

H_0: Mean (Mu) = 7

H_a: Mean (Mu) Not Equal To 7

Results:	**HtBk**
Count	40
Mean	6.453
StDev	0.314222
SE Mean	0.049682825
t	-11.005
P-Value (2-sided)	**0.0000**
UC (2-sided, 95%)	6.554
LC (2-sided, 95%)	6.353

1 Sample t-Test Assumptions Report	
Normality:	**Anderson Darling P-Value = 0.275. Fail to reject null hypothesis: "data are sampled from a normal distribution," so conclude that the assumption of normality is not violated.**
Robustness:	**Not applicable for normal data.**
Outliers (Boxplot Rules):	**No outliers found.**
Randomness (Independence):	**Nonparametric Runs Test (Exact) P-Value = 0.334. Fail to reject null hypothesis: "data are random," so conclude that the assumption of randomness (independence) is not violated.**

Null Hypothesis: (H_0): $\mu = 7$
Alternative Hypothesis: (H_a): $\mu \neq 7$

Since the p-value is smaller than alpha level (0.05), we reject the null hypothesis and claim that the average height of our basketball players is statistically different from 7 feet.

What is Two Sample T-Test?

- **Two sample t-test** is a hypothesis test to study whether there is a statistically significant difference between the means of two populations.

  - Null Hypothesis (H_0): $\mu_1 = \mu_2$
  - Alternative Hypothesis (H_a): $\mu_1 \neq \mu_2$

  where μ_1 is the mean of one population and μ_2 is the mean of the other population of our interest.

Assumptions of Two Sample T-Tests

- The sample data drawn from both populations are unbiased and representative.
- The data of both populations are continuous.
- The data of both populations are normally distributed.
- The variances of both populations are unknown.
- Two sample t-test is more robust than a z-test when the sample size is small (< 30).

Three Types of Two Sample T-Tests

1. Two sample t-test when the variances of two populations are unknown but equal
   - Two sample t-test (when $\sigma_1 = \sigma_2$)

2. Two sample t-test when the variances of the two population are unknown and unequal
   - Two sample t-test (when $\sigma_1 \neq \sigma_2$)

3. Paired t-test when the two populations are dependent of each other

Test of Equal Variance

- To check whether the variances of two populations of interest are statistically significant different, we use the test of equal variance.
  - Null Hypothesis (H_0): $\sigma_1^2 = \sigma_2^2$
  - Alternative Hypothesis (H_1): $\sigma_1^2 \neq \sigma_2^2$
- An F-test is used to test the equality of variances between two normally distributed populations.

Test of Equal Variance

- An **F-test** is a statistic hypothesis test in which the test statistic follows an F-distribution when the null hypothesis is true.

- The most known F-test is the test of equal variance for two normally distributed populations.

- The F-test is very sensitive to non-normality. When any one of the two populations is not normal, we use the Brown-Forsythe test for checking the equality of variances.

Test of Equal Variance

- Test Statistic

$$F_{calc} = \frac{s_1^2}{s_2^2}$$

where

s_1 and s_2 are the sample standard deviations.

- Critical Value
  - F_{crit} is the F value in a F distribution with the predetermined significance level α and degrees of freedom ($n_1 - 1$) and ($n_2 - 1$).
  - F_{crit} values for a two-sided and a one-sided F-test with the same significance level α and degrees of freedom ($n_1 - 1$) and ($n_2 - 1$) are different.

Test of Equal Variance

- Based on the sample data, we calculated the test statistic F_{calc}, which is compared against F_{crit} to make a decision of whether to reject the null.
  - Null Hypothesis (H_0): $\sigma_1^2 = \sigma_2^2$
  - Alternative Hypothesis (H_a): $\sigma_1^2 \neq \sigma_2^2$

- If $F_{calc} > F_{crit}$, we reject the null and claim there is a statistically significant difference between the variances of the two populations.

- If $F_{calc} < F_{crit}$, we fail to reject the null and claim there is not any statistically significant difference between the variances of the two populations.

Test Statistic & Critical Value of a Two Sample T-Test

when $\sigma_1 = \sigma_2$

- Test Statistic

$$t_{calc} = \frac{\overline{Y}_1 - \overline{Y}_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{(n_1 + n_2 - 2)}}$$

where

$\overline{Y}_1$ and $\overline{Y}_2$ are the sample means of the two populations of our interest.

n_1 and n_2 are the sample sizes. n_1 is not necessarily equal to n_2.
s_{pooled} is a pooled estimate of variance. s_1 and s_2 are the sample standard deviations.

- Critical Value

t_{crit} is the t value in a Student t distribution with the predetermined significance level α and degrees of freedom ($n_1 + n_2 - 2$).

t_{crit} values for a two-sided and a one-sided t-test with the same significance level α and different degrees of freedom ($n_1 + n_2 - 2$).

Test Statistic & Critical Value of a Two Sample T-Test

when $\sigma_1 \neq \sigma_2$

- Test Statistic

$$t_{calc} = \frac{\overline{Y}_1 - \overline{Y}_2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2/n_2\right)^2}{n_2 - 1}}$$

where

$\overline{Y}_1$ and $\overline{Y}_2$ are the sample means of the two populations of our interest.

n_1 and n_2 are the sample sizes. n_1 is not necessarily equal to n_2.

s_1 and s_2 are the sample standard deviations.

- Critical Value

t_{crit} is the t value in a Student t distribution with the predetermined significance level α and degrees of freedom df calculated using the formula above.

t_{crit} values for a two-sided and a one-sided t-test with the same significance level α and different degrees of freedom df.

Test Statistic & Critical Value of a Paired T-Test

- Test Statistic

$$t_{calc} = \frac{\bar{d}}{s_d / \sqrt{n}}$$

where

d is the difference between each pair of data.

$\bar{d}$ is the average of d.

n is the sample size of either population of interest.

s_d is standard deviation of *d*.

- Critical Value

t_{crit} is the t value in a Student t distribution with the predetermined significance level *α* and degrees of freedom (*n* – 1).

t_{crit} values for a two-sided and a one-sided t-test with the same significance level α and different degrees of freedom (*n* – 1).

Decision Rules of a Two Sample T-Test

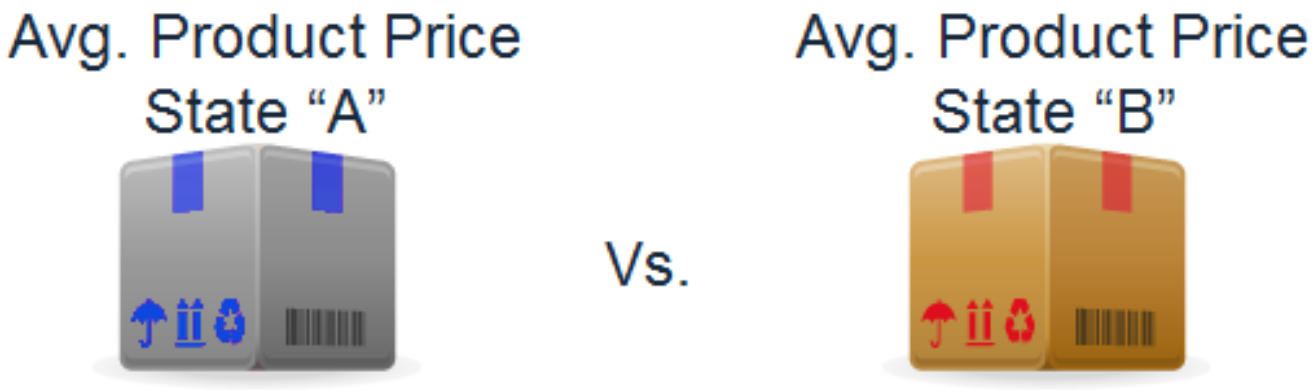
- Based on the sample data, we calculated the test statistic t_{calc}, which is compared against t_{crit} to make a decision of whether to reject the null.
  - Null Hypothesis (H_0): $\mu_1 = \mu_2$
  - Alternative Hypothesis (H_a): $\mu_1 \neq \mu_2$

- If $|t_{calc}| > t_{crit}$, we reject the null and claim there is a statistically significant difference between the means of the two populations.

- If $|t_{calc}| < t_{crit}$, we fail to reject the null and claim there is not any statistically significant difference between the means of the two populations.

Use SigmaXL to Run a Two-Sample T-Test

- Case Study: We are trying to compare the average retail price of a product in state A and state B.
  - Data File: “Two-Sample T-Test” tab in “Sample Data.xlsx”

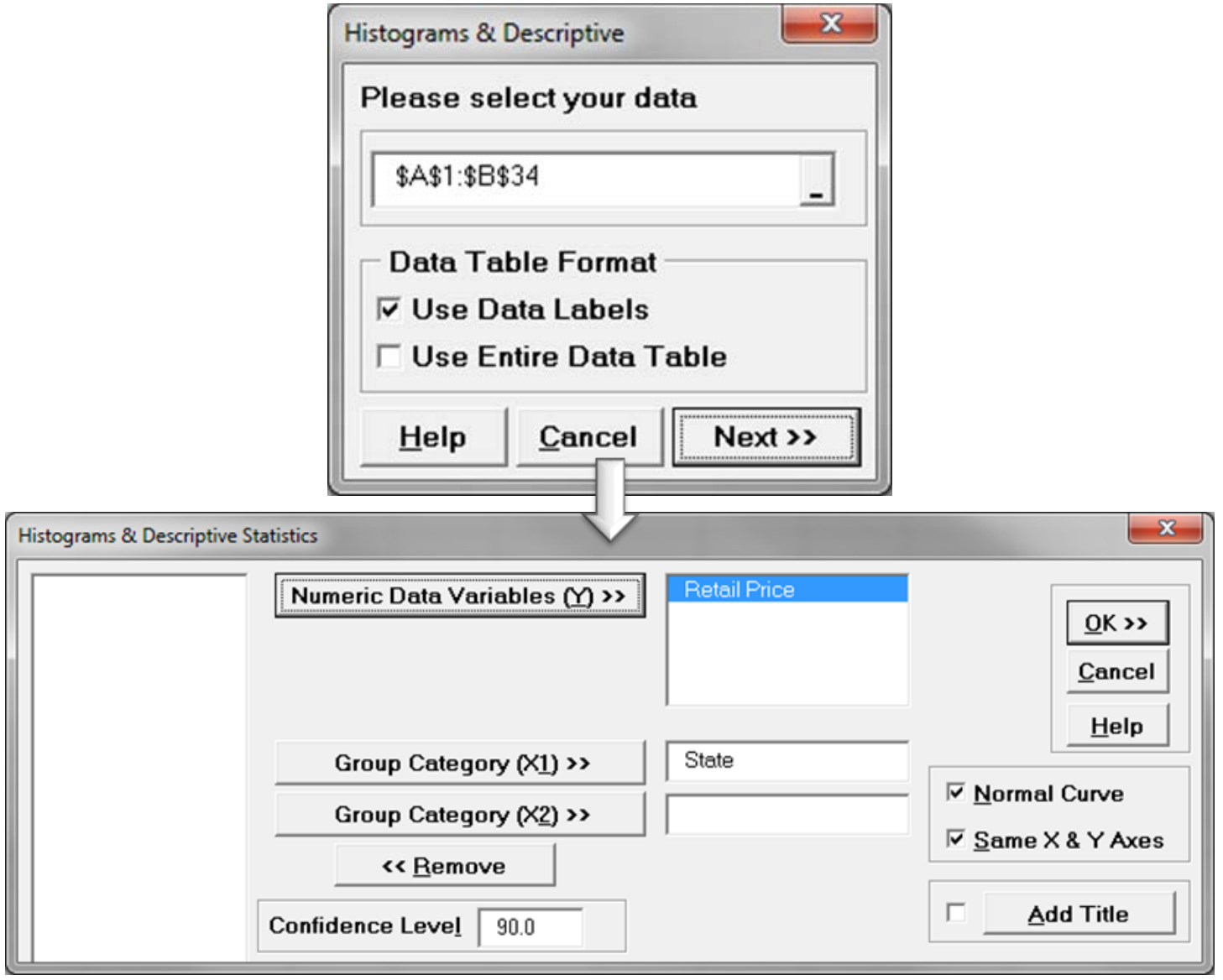


- Null Hypothesis (H_0): $\mu_1 = \mu_2$
- Alternative Hypothesis (H_a): $\mu_1 \neq \mu_2$

Use SigmaXL to Run a Two-Sample T-Test

- Step 1: Test the normality of the retail price for both state A and B.
  - Select the entire range of data (both State and Retail Price)
  - Click SigmaXL -> Graphical Tools -> Histogram & Descriptive Statistics
  - A window named “Histogram & Descriptive” pops up with the selected range pre-populated in the box below “Please select your data”
  - Click “Next >>”
  - Another window named “Histogram & Descriptive Statistics” appears
  - Select “Retail Price” as the “Numeric Data Variables”
  - Select “State” as the “Group Category (X1)”
  - Click “OK>>”
  - The normality test results would appear in the tab “Hist Descript (1)” automatically.

Use SigmaXL to Run a Two-Sample T-Test



Use SigmaXL to Run a Two-Sample T-Test

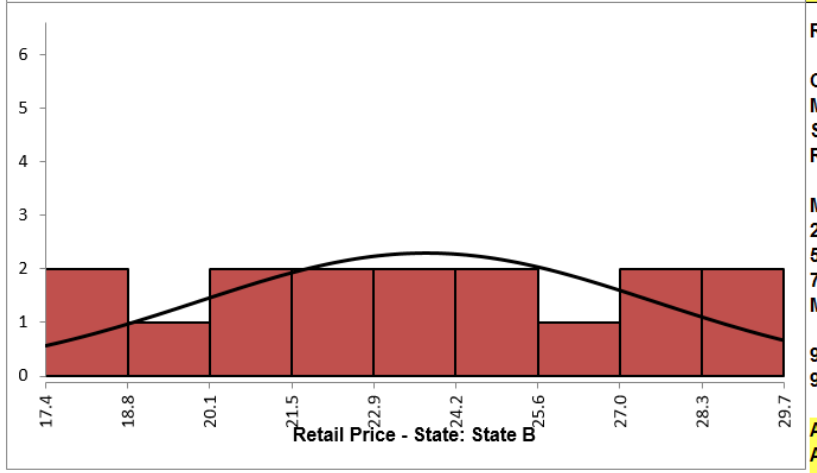
Retail Price - State: State A

Count = 17
Mean = 23.176
Stdev = 3.652
Range = 12.00

Minimum = 17.400
25th Percentile (Q1) = 19.350
50th Percentile (Median) = 23.700
75th Percentile (Q3) = 25.650
Maximum = 29.4

90% CI Mean = 21.63 to 24.72
90% CI Sigma = 2.85 to 5.18

Anderson-Darling Normality Test:
A-Squared = 0.421476; P-Value = 0.2863



Retail Price - State: State B

Count = 16
Mean = 23.744
Stdev = 3.805
Range = 12.20

Minimum = 17.500
25th Percentile (Q1) = 20.875
50th Percentile (Median) = 23.950
75th Percentile (Q3) = 27.225
Maximum = 29.7

90% CI Mean = 22.08 to 25.41
90% CI Sigma = 2.95 to 5.47

Anderson-Darling Normality Test:
A-Squared = 0.200124; P-Value = 0.8580

Null Hypothesis (H_0): The data are normally distributed.
Alternative Hypothesis (H_a): The data are not normally distributed.

Both retail price data of state A and B are normally distributed since the p-values are both greater than alpha level (0.05).

If any of the data series is not normally distributed, we need to use other hypothesis testing methods other than the two sample t-test.

Use SigmaXL to Run a Two-Sample T-Test

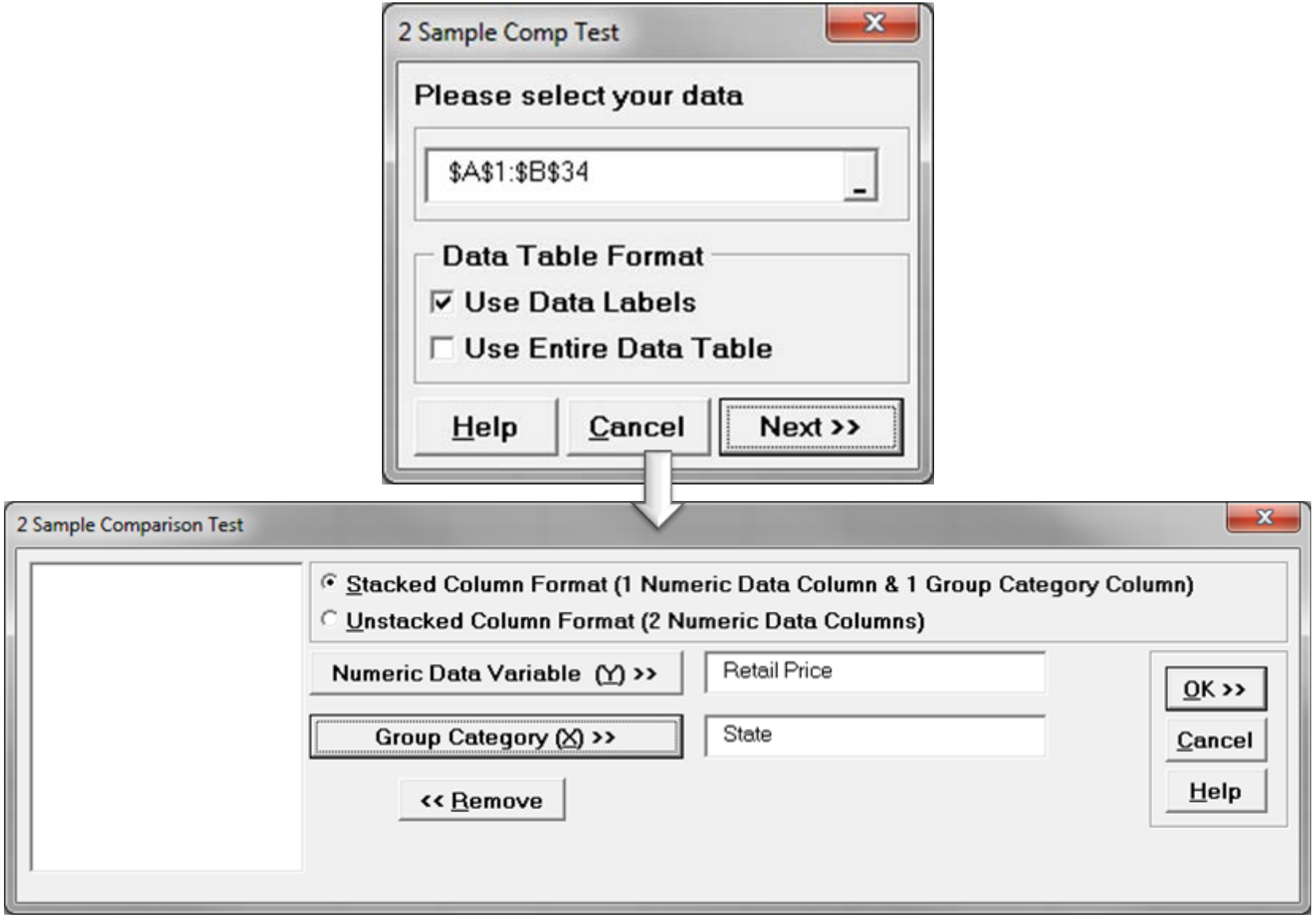
- Step 2: Test whether the variances of the two data sets are equal.

  Null Hypothesis (H_0): $\sigma_1^2 = \sigma_2^2$

  Alternative Hypothesis (H_a): $\sigma_1^2 \neq \sigma_2^2$

  - Select the entire range of data (both “State” and “Retail Price”)
  - Click SigmaXL -> Statistical Tools -> 2 Sample Comparison Tests
  - A new window named “2 Sample Comp Test” pops up with the selected range pre-populated in the box under “Please select your data”.
  - Click “Next>>”
  - Another window named “2 Sample Comparison Test” appears
  - Click the radio button “Stacked Column Format (1 Numeric Data Column & 1 Group Category Column)
  - Select “Retail Price” as the “Numeric Data Variable (Y)”
  - Select “State” as the “Group Category (X)”
  - Click “OK>>”
  - The results show up in the tab “2 Sample Comparison Test (1)”

Use SigmaXL to Run a Two-Sample T-Test



Use SigmaXL to Run a Two-Sample T-Test

2 Sample Comparison - Retail Price		
State	**State A**	**State B**
Count	17	16
Mean	23.176	23.744
Median	23.700	23.950
Standard Deviation	3.652	3.805
AD Normality Test P-Value	0.2863	0.8580
Test for Equal Variances:		
F-test (use with normal data):		
F (test statistic)	1.08510047	
P-Value (2-sided)	0.8700	
Levene's test (use with non-normal data):		
P-Value (2-sided)	0.6900	
2 Sample t-Test for means:		
Assume Equal Variance:		
t (test statistic)	-0.436992	
P-Value (2-sided)	0.6651	
P-Value (1-sided)	0.3326	
Assume Unequal Variance:		
t (test statistic)	-0.436434	
P-Value (2-sided)	0.6656	
P-Value (1-sided)	0.3328	
2 Sample Mann-Whitney test for medians:		
P-Value (2-sided)	0.5764	
P-Value (1-sided)	0.2882	

Because the retail prices at state A and state B are both normally distributed, F test is used to test their variance equality.

The p-value of F test is 0.87 greater than the alpha level (0.05), so we fail to reject the null and we claim that the variances of the two data sets are equal. We will use the two sample t-test (when $\sigma_1 = \sigma_2$) to compare the means of the two groups.

If $\sigma_1 \neq \sigma_2$, we will use the two sample t-test (when $\sigma_1 \neq \sigma_2$) to compare the means of the two groups.

Use SigmaXL to Run a Two-Sample T-Test

- Step 3: Run two-sample t-test to compare the means of two groups.
  - The two sample comparison test we ran in step 2 also automatically generates the two-sample t-test result.

2 Sample Comparison - Retail Price		
State	**State A**	**State B**
Count	17	16
Mean	23.176	23.744
Median	23.700	23.950
Standard Deviation	3.652	3.805
AD Normality Test P-Value	0.2863	0.8580
Test for Equal Variances:		
F-test (use with normal data):		
F (test statistic)	1.08510047	
P-Value (2-sided)	0.8700	
Levene's test (use with non-normal data):		
P-Value (2-sided)	0.6900	
2 Sample t-Test for means:		
Assume Equal Variance:		
t (test statistic)	-0.436992	
P-Value (2-sided)	0.6651	
P-Value (1-sided)	0.332	
Assume Unequal Variance:		
t (test statistic)	-0.436434	
P-Value (2-sided)	0.6656	
P-Value (1-sided)	0.3328	
2 Sample Mann-Whitney test for medians:		
P-Value (2-sided)	0.5764	
P-Value (1-sided)	0.2882	

Since the p-value of t test (assuming equal variance) is 0.6651 greater than the alpha level (0.05)

we fail to reject the null hypothesis and we claim that the means of the two data sets are equal.

Use SigmaXL to Run a Two-Sample T-Test

- If the variances of the two groups do not equal, we will need to use the two-sample t-test (when $\sigma_1 \neq \sigma_2$) to compare the means of the two groups.

2 Sample Comparison - Retail Price		
State	**State A**	**State B**
Count	17	16
Mean	23.176	23.744
Median	23.700	23.950
Standard Deviation	3.652	3.805
AD Normality Test P-Value	0.2863	0.8580
Test for Equal Variances:		
F-test (use with normal data):		
F (test statistic)	1.08510047	
P-Value (2-sided)	0.8700	
Levene's test (use with non-normal data):		
P-Value (2-sided)	0.6900	
2 Sample t-Test for means:		
Assume Equal Variance:		
t (test statistic)	-0.436992	
P-Value (2-sided)	0.6651	
P-Value (1-sided)	0.3326	
Assume Unequal Variance:		
t (test statistic)	-0.436434	
P-Value (2-sided)	0.6656	
P-Value (1-sided)	0.3328	
2 Sample Mann-Whitney test for medians:		
P-Value (2-sided)	0.5764	
P-Value (1-sided)	0.2882	

Since the p-value of the t test (assuming unequal variance) is 0.6656 greater than the alpha level (0.05)

we fail to reject the null hypothesis and we claim that the means of two groups are equal.

Use SigmaXL to Run a Paired T-Test

- Case Study: We are interested to know whether the average salaries ($1000/yr) of male and female professors at the same university are the same.
  - Data File: “Paired T-Test” tab in “Sample Data.xlsx”
  - The data were randomly collected from 22 universities. For each university, the salaries of a male and female professors were randomly selected.
  - The differences were calculated displayed in the data file.

- Null Hypothesis (H_0): $\mu_{male} - \mu_{female} = 0$
- Alternative Hypothesis (H_a): $\mu_{male} - \mu_{female} \neq 0$

Use SigmaXL to Run a Paired T-Test

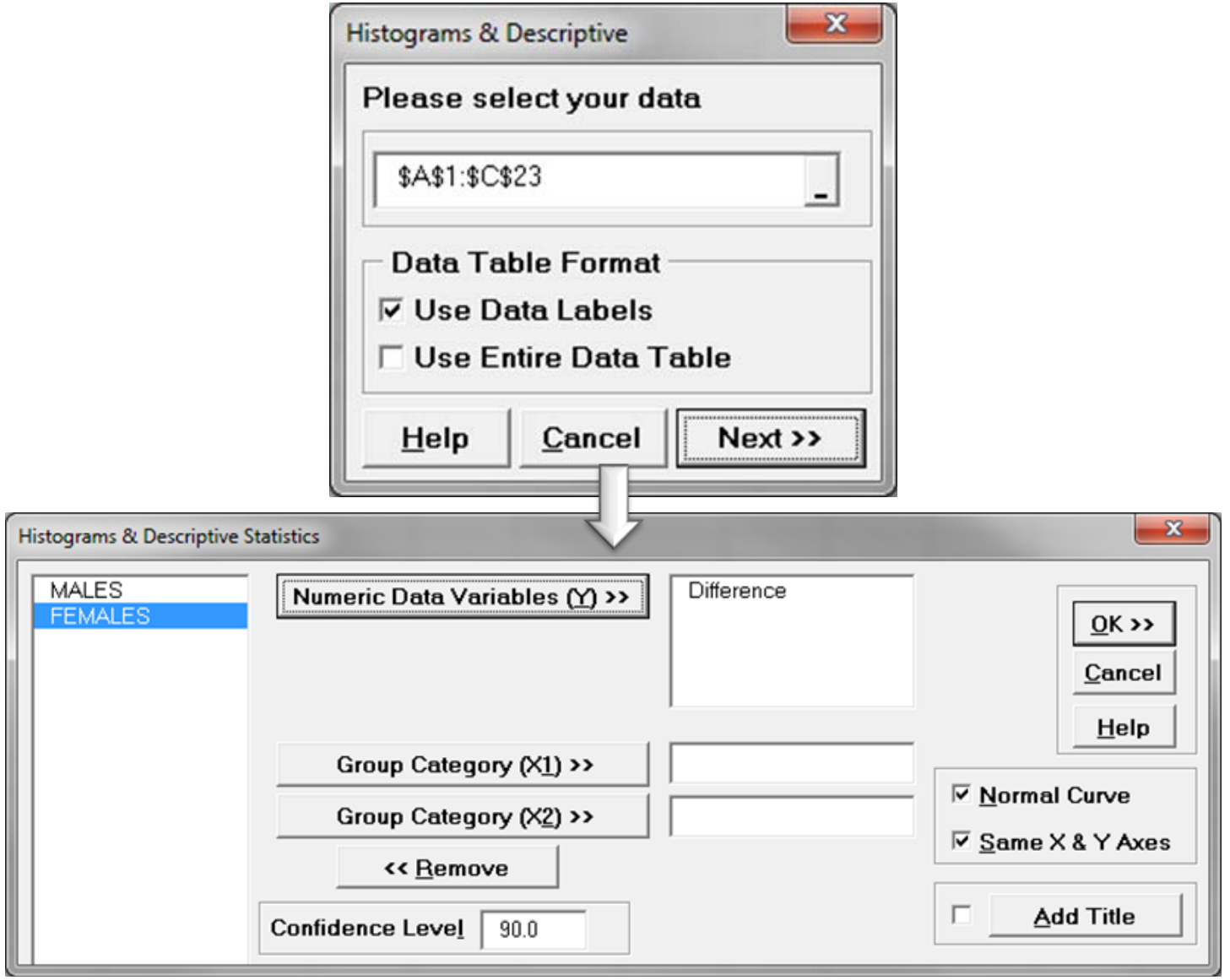
- Step 1: Test whether the difference is normally distributed.

  Null Hypothesis (H_0): The difference between two data sets is normally distributed.

  Alternative Hypothesis (H_a): The difference between two data sets is not normally distributed.

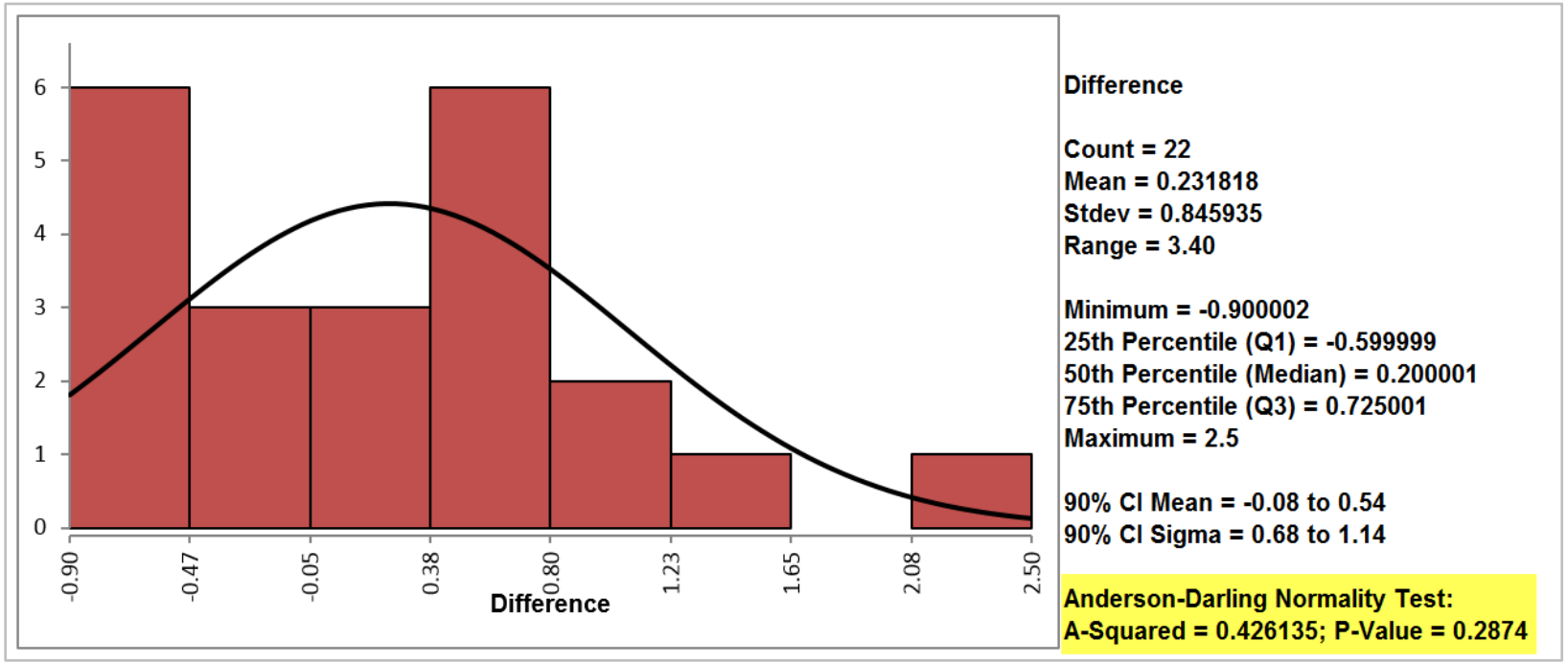
  - Select the entire range of “Difference”
  - Click SigmaXL -> Graphical Tools -> Histograms & Descriptive Statistics
  - A new window named “Histogram & Descriptive” pops up with the selected range pre-populated in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Histograms & Descriptive Statistics” appears.
  - Select “Difference” as the “Numeric Data Variables (Y)”
  - Click “OK>>”

Use SigmaXL to Run a Paired T-Test



Use SigmaXL to Run a Paired T-Test

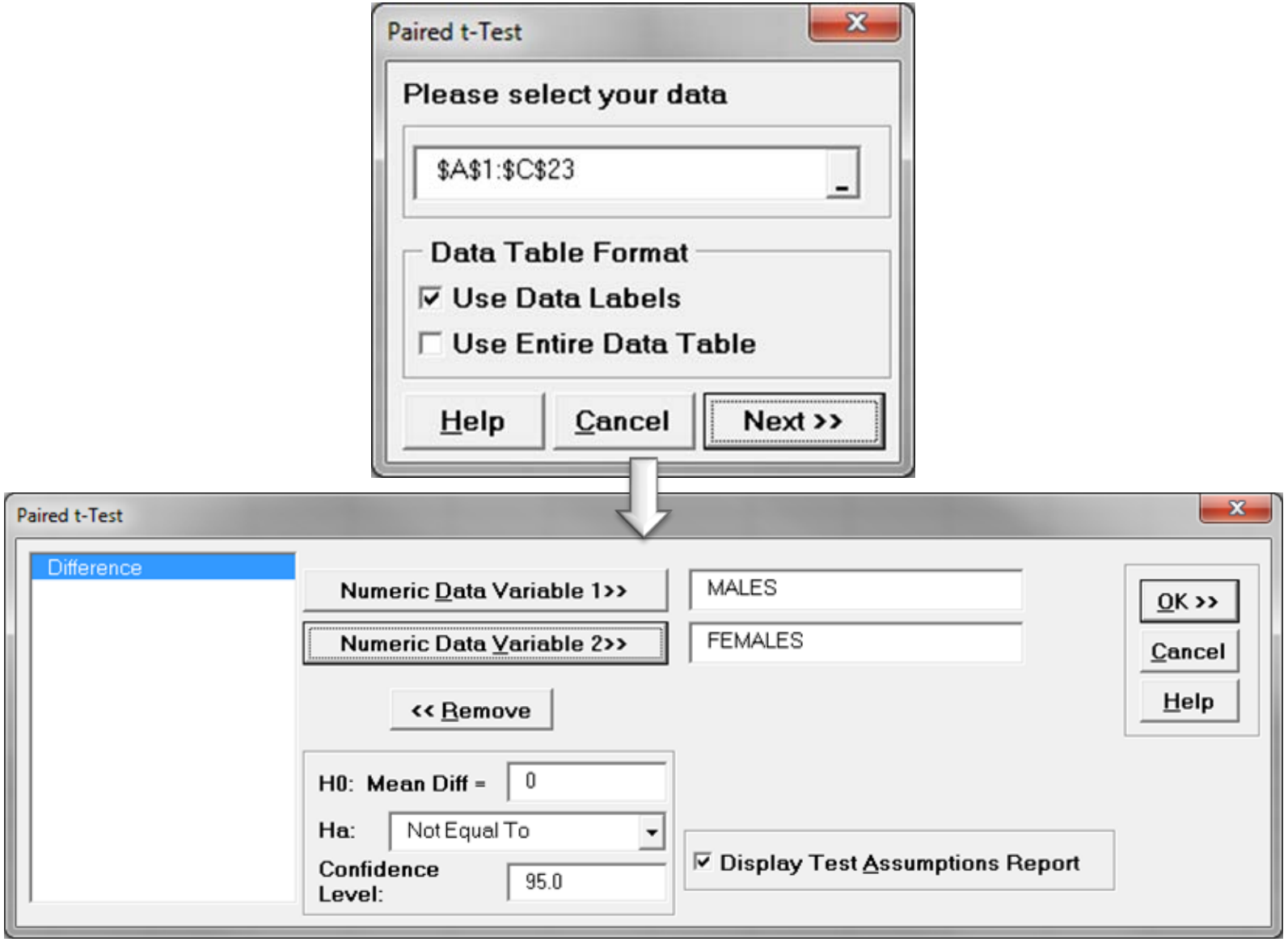
- The p-value of the normality test is 0.2874 greater than the alpha level (0.05), so we fail to reject the null hypothesis and we claim that the difference is normally distributed.
- If the difference is not normally distributed, we need other hypothesis testing methods.



Use SigmaXL to Run a Paired T-Test

- Step 2: Run the paired t-test to compare the means of two dependent data sets.
  - Select the entire range of both MALES and FEMALES data
  - Click SigmaXL -> Statistical Tools -> Paired t-Test
  - A new window named “Paired t-Test” pops up with the selected range pre-populated in the box under “Please select your data”
  - Click “Next >>”
  - Another new window named “Paired t-Test” pops up
  - Select “MALES” as “Numeric Data Variable 1”
  - Select “FEMALES” as “Numeric Data Variable 2”
  - Click “OK>>”
  - The paired t-test results appears in the tab “Paired t-Test (1)”

Use SigmaXL to Run a Paired T-Test



Use SigmaXL to Run a Paired T-Test

- The p-value of the paired t-test is 0.2127 greater than the alpha level (0.05), so we fail to reject the null hypothesis and we claim that there is no statistically significant difference between the salaries of male and female professors' salaries.

Paired t-Test	
Test Information	
H_0: Mean Difference = 0	
H_a: Mean Difference Not Equal To 0	
Results:	**MALES - FEMALES**
Count	22
Mean	0.231818
StDev	0.845935
SE Mean	0.180354
t	1.285
P-Value (2-sided)	0.2127
UC (2-sided, 95%)	0.606884
LC (2-sided, 95%)	-0.143249

Paired t-Test Assumptions Report	
Normality:	Anderson Darling P-Value = 0.287. Fail to reject null hypothesis: "data are sampled from a normal distribution," so conclude that the assumption of normality is not violated.
Robustness:	Not applicable for normal data.
Outliers (Boxplot Rules):	No outliers found.
Randomness (Independence):	Nonparametric Runs Test (Exact) P-Value = 0.270. Fail to reject null hypothesis: "data are random," so conclude that the assumption of randomness (independence) is not violated.

3.4.2 One Sample Variance

What is One Sample Variance Test?

- **One sample variance test** is a hypothesis testing method to study whether there is a statistically significant difference between a population variance and a specified value.

  - Null Hypothesis (H_0): $\sigma^2 = \sigma_0^2$
  - Alternative Hypothesis (H_a): $\sigma^2 \neq \sigma_0^2$

where σ^2 is the variance of a population of our interest and σ_0^2 is the specific value we want to compare against.

Chi-Square Test

- A **chi-square test** is a statistical hypothesis test in which the test statistic follows a chi-square distribution when the null hypothesis is true.

- A chi-square test can be used to test the equality between the variance of a normally distributed population and a specified value.

Chi-Square Test

- Test Statistic

$$\chi_{calc}^{2} = \frac{(n-1)s^2}{\sigma_0^2}$$

where

s^2 is the observed variance and n is the sample size.

σ_0^2 is the specified value we compare against.

- Critical Value
  - χ_{crit}^{2} is the χ^2 value in a χ^2 distribution with the predetermined significance level α and degrees of freedom (n – 1).
  - χ_{crit}^{2} values for a two-sided and a one-sided χ^2-test with the same significance level α and different degrees of freedom (*n –1).*

Chi-Square Test

- Based on the sample data, we calculated the test statistic χ_{crit}^2, which is compared against χ_{calc}^2 to make a decision of whether to reject the null.
  - Null Hypothesis (H_0): $\sigma^2 = \sigma_0^2$
  - Alternative Hypothesis (H_a): $\sigma^2 \neq \sigma_0^2$

- If $\left|\chi_{calc}^2\right| > \chi_{crit}^2$, we reject the null and claim there is a statistically significant difference between the population variance and the specified value.

- If $\left|\chi_{calc}^2\right| < \chi_{crit}^2$, we fail to reject the null and claim there is not any statistically significant difference between the population variance and specified value.

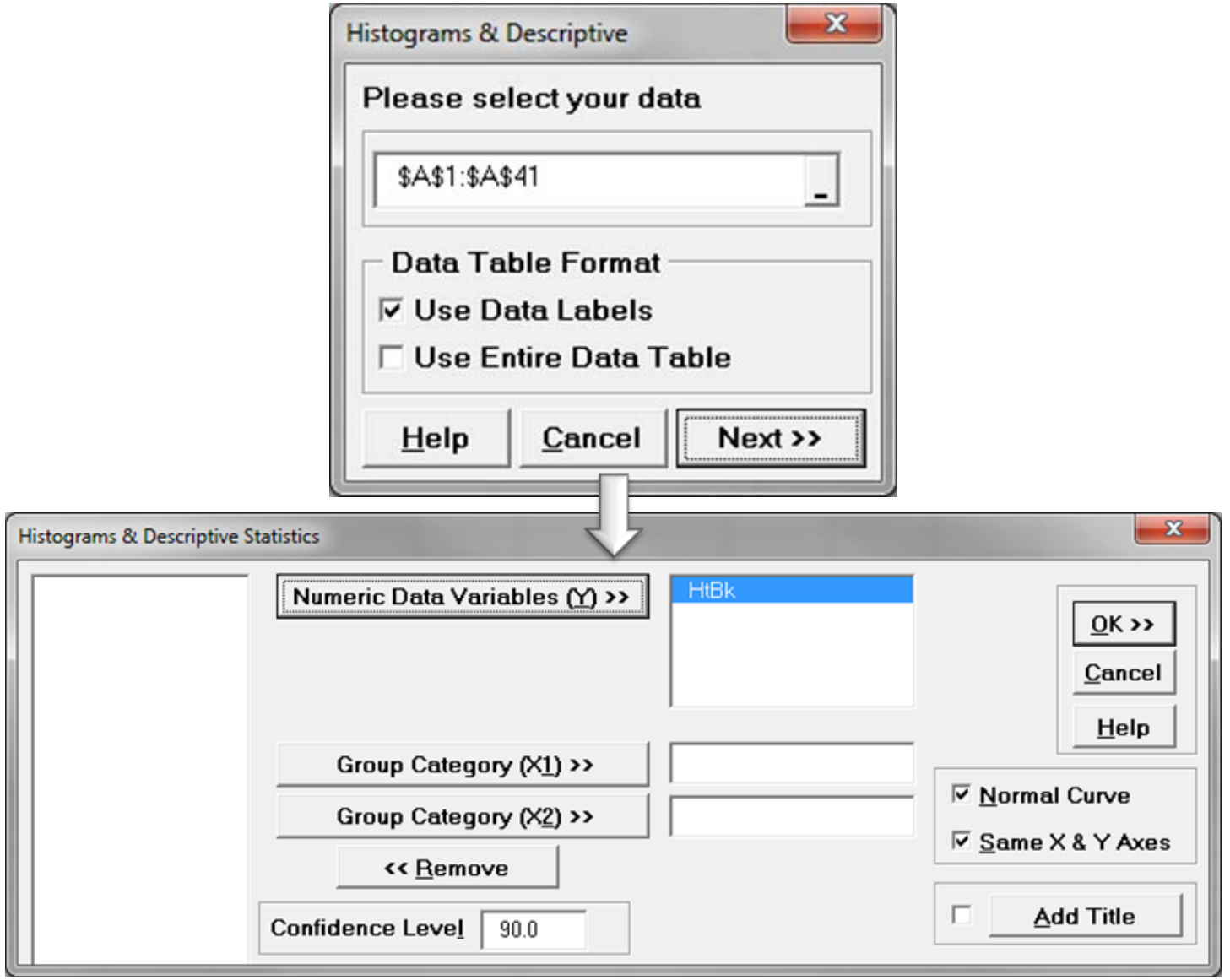
Use SigmaXL to Run a One Sample Variance Test

- Case Study: we are interested in comparing the variance of the height of basketball players with zero.
  - Data File: “One Sample T-Test” tab in “Sample Data.xlsx”
- Null Hypothesis (H_0): $\sigma^2 = 0$
- Alternative Hypothesis (H_a): $\sigma^2 \neq 0$

Use SigmaXL to Run a One Sample Variance Test

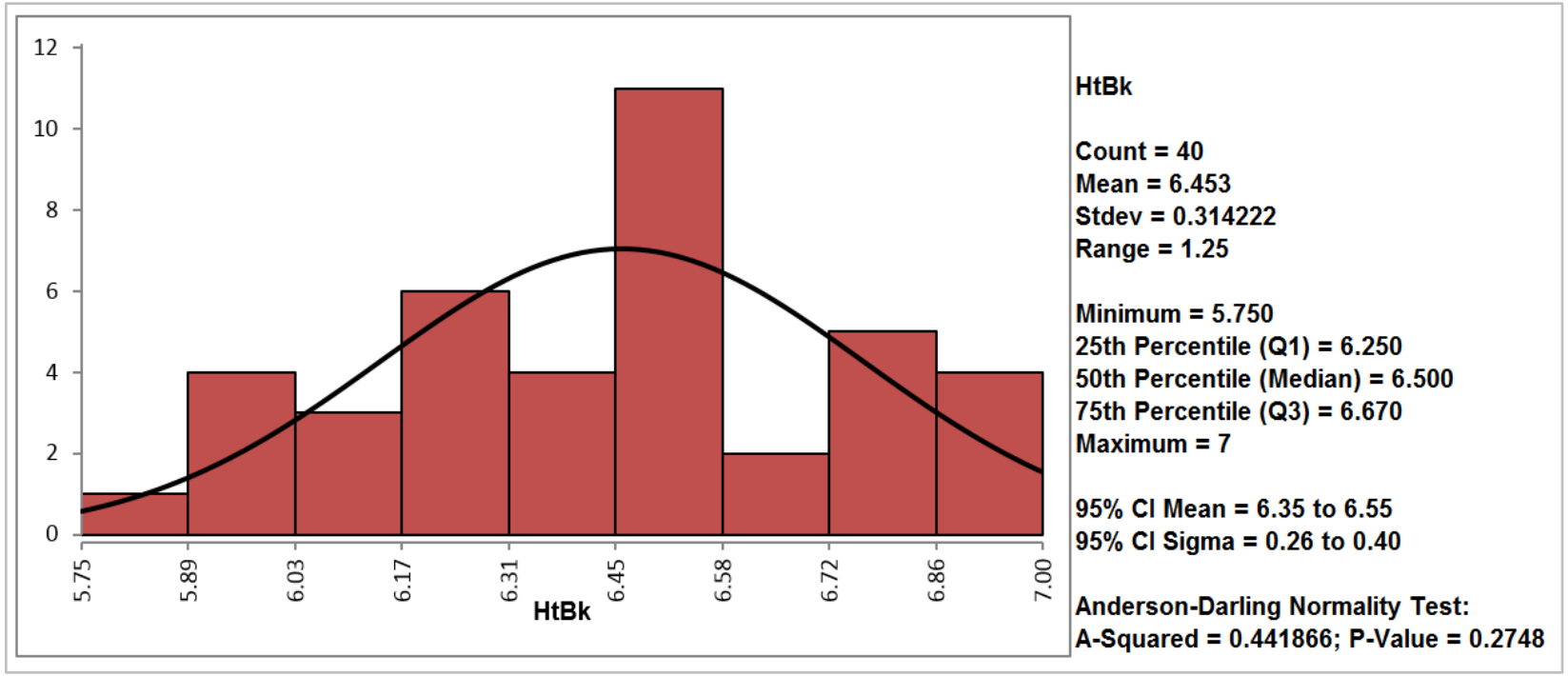
- Step 1: Test whether the data are normally distributed
  - Select the entire range of “HtBk”
  - Click SigmaXL -> Graphical Tools -> Histograms & Descriptive Statistics
  - A new window named “Histogram & Descriptive” pops up with the selected range pre-populated in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Histograms & Descriptive Statistics” appears.
  - Select “Difference” as the “Numeric Data Variables (Y)”
  - Click “OK>>”
  - The normality test results appears in the tab “Hist Descript (1)”

Use SigmaXL to Run a One Sample Variance Test



Use SigmaXL to Run a One Sample Variance Test

- Null Hypothesis (H_0): The data are normally distributed.
- Alternative Hypothesis (H_a): The data are not normally distributed.
- Since the p-value of the normality is 0.2748 greater than alpha level (0.05), we fail to reject the null and claim that the data are normally distributed.
- If the data are not normally distributed, you need to use other hypothesis tests.



Use SigmaXL to Run a One Sample Variance Test

- Step 2: Check if the specified value is different from 0.01
  [we really want to know if it's different from zero but we need a value other than 0 for SigmaXL to calculate so we will use 0.01].
  - Select the entire range of "HtBk"
  - Click SigmaXL -> Statistical Tools -> Basic Statistical Templates -> 1 Sample Chi-Square Test and CI for Standard Deviation
  - The one sample variance test appears in the newly generated tab "1 Sample CI StDev"
  - A New Template Opens, Enter 40 in Sample Size
  - Enter 0.324222 in Sample Standard Deviation
  - Enter 0.01 in (hypothesized StdDev)

Use SigmaXL to Run a One Sample Variance Test

- If the square root of the specified value is between the upper and lower confidence interval boundaries then we fail to reject the null hypothesis and claim that the variance of the population is not statistically different from the specified value.
- In this case, the square root of 0.01 is outside the confidence interval of the population standard deviation and we claim that the population variance is statistically different from zero.

1 Sample Standard Deviation Chi-Square Test and Confidence Interval

Sample Data (user inputs):		
Sample Size	n	40
Sample Standard Deviation	s	0.314222
Null Hypothesis (hypothesized StdDev)	H_0: Sigma (σ) =	0.01
Alternative Hypothesis	H_a: Sigma (σ)	Not Equal To
Confidence Level (enter .95 for 95%)	100*(1-α)%	95.0%

Test Information	
Null Hypothesis H_0: Sigma (σ) = 0.01	Reject
Alternative Hypothesis H_a: Sigma (σ) ≠ 0.01	Conclude true at 95.0% confidence level

Results:	
Sample Variance	0.0987
DF	39
Chi-Square test statistic	38506.8315
alpha	0.0500
P-Value (2 sided)	0.0000
Upper Confidence Limit (2 Sided)	0.4035
Lower Confidence Limit (2 Sided)	0.2574

Interpretation of P-Value and Confidence Intervals
The P-Value is the two-sided (two-tail) probability of obtaining a test statistic at least as extreme as 38,506.8315, given that the null hypothesis is true. Since the P-Value is less than alpha (0.05), we reject the null hypothesis: Sigma (σ) = 0.01 and conclude that the alternative hypothesis Ha: Sigma (σ) ≠ 0.01 is true at the confidence level of 95.0%.
0.4035).

3.4.3 One Way ANOVA

What is One-Way ANOVA?

- **One-way ANOVA** (one-way analysis of variance) is a statistical method to compare means of two or more populations.
  - Null Hypothesis (H_0): $\mu_1 = \mu_2 = ... = \mu_k$
  - Alternative Hypothesis (H_a): at least one μ_i is different, where *i* is any value from 1 to k.
- It is a generalized form of the two sample t-test since a two sample t-test compares two population means and one-way ANOVA compares *k* population means where $k \geq 2$.

Assumptions of One-Way ANOVA

- The sample data drawn from *k* populations are unbiased and representative.
- The data of *k* populations are continuous.
- The data of *k* populations are normally distributed.
- The variances of *k* populations are equal.

How ANOVA Works

- ANOVA compares the means of different groups by analyzing the variances between and within groups.

- Let us say we are interested in comparing the means of three normally distributed populations. We randomly collected one sample for each population of our interest.

  - Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3$
  - Alternative Hypothesis (H_a): one of the μ is different from the others.

How ANOVA Works

- Based on the sample data, the means of the three populations might look different because of two variation sources.
  1) Variation between groups

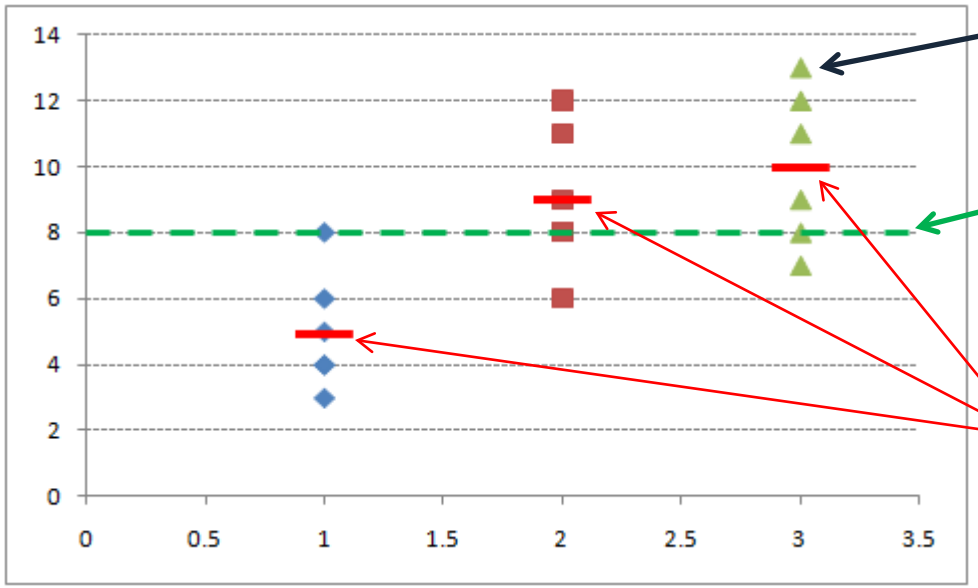
     There are non-random factors leading to the variation between groups.
  2) Variation within groups

     There are random errors resulting in the variation within each individual group.

- What we care about the most is the variation between groups since we are interested in whether the groups are statistically different from each other.

- Variation between groups is the *signal* we want to detect and variation within groups is the *noise* which corrupts the signal.

How ANOVA Works



Individual observation: Y_{ij}

Grand mean of all observations: $\overline{\overline{Y}}$

Group means of each individual sample: $\overline{Y}_j$

Total Variation = SS(Total) = $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{\overline{Y}})^2$

Between Variation = SS(Between) = $\sum_{j=1}^{k} n_j(\overline{Y}_j-\overline{\overline{Y}})^2$

Within Variation = SS(Within) = $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)^2$

How ANOVA Works

- Variation Components
  - Total Variation = Variation Between Groups + Variation Within Groups

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{\bar{Y}})^2 \;=\; \sum_{j=1}^{k} n_j(\bar{Y}_j-\bar{\bar{Y}})^2 \;+\; \sum_{j=1}^{k}\sum_{i=1}^{n_j}(Y_{ij}-\bar{Y}_j)^2$$

    - Total Variation = sums of squares of the vertical difference between the individual observation and the grand mean
    - Variation Between Groups = sums of squares of the vertical difference between the group mean and the grand mean
    - Variation Within Groups = sum of squares of the vertical difference between the individual observation and the group mean

How ANOVA Works

- Degrees of Freedom (DF)
  - In statistics, the degrees of freedom is the number of unrestricted values in the calculation of a statistic.

- Degrees of Freedom Components
  - $DF_{total} = DF_{between} + DF_{within}$
  - $DF_{total} = n - 1$
  - $DF_{between} = k - 1$
  - $DF_{within} = n - k$

where

n is the total number of observations
k is the number of groups.

How ANOVA Works

- Signal-to-Noise Ratio (SNR)
  - SNR denotes the ratio of a signal to the noise corrupting the signal.
  - It measures how much a signal has been corrupted by the noise.
  - When it is higher than 1, there is more signal than noise.
  - The higher the SNR, the less the signal has been corrupted by the noise.
- F-ratio is the SNR in ANOVA

$$F = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between} \big/ DF_{between}}{SS_{within} \big/ DF_{within}} = \frac{\sum_{j=1}^{k} (\bar{Y}_j - \bar{\bar{Y}})^2 \big/ (k-1)}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(Y_{ij} - \bar{Y}_j\right)^2 \big/ (n-k)}$$

  - In ANOVA, we use the F-test to compare the means of different groups. The F-ratio calculated as above is the test statistic F_{calc}.

  - The critical value (F_{cri}) in an F-test can be derived from the F table with predetermined significance level (α) and with (k –1) degrees of freedom in the numerator and (n – k) degrees of freedom in the denominator.

How ANOVA Works

- Null Hypothesis (H_0): $\mu_1 = \mu_2 = \ldots = \mu_k$
- Alternative Hypothesis (H_a): at least one μ_i is different, where *i* is any value from 1 to *k*.
  - If $|F_{calc}| < F_{crit}$, we fail to reject the null and claim that the means of all the populations of our interest are the same.
  - If $|F_{calc}| > F_{crit}$, we reject the null and claim that there is at least one mean different from the others.

Model Validation

- ANOVA is a modeling procedure. To make sure the conclusions made in ANOVA are reliable, we need to perform residuals analysis.

- Good residuals:
  - Have a mean of zero
  - Are normally distributed
  - Are independent of each other
  - Have equal variance.

Use SigmaXL to Run ANOVA

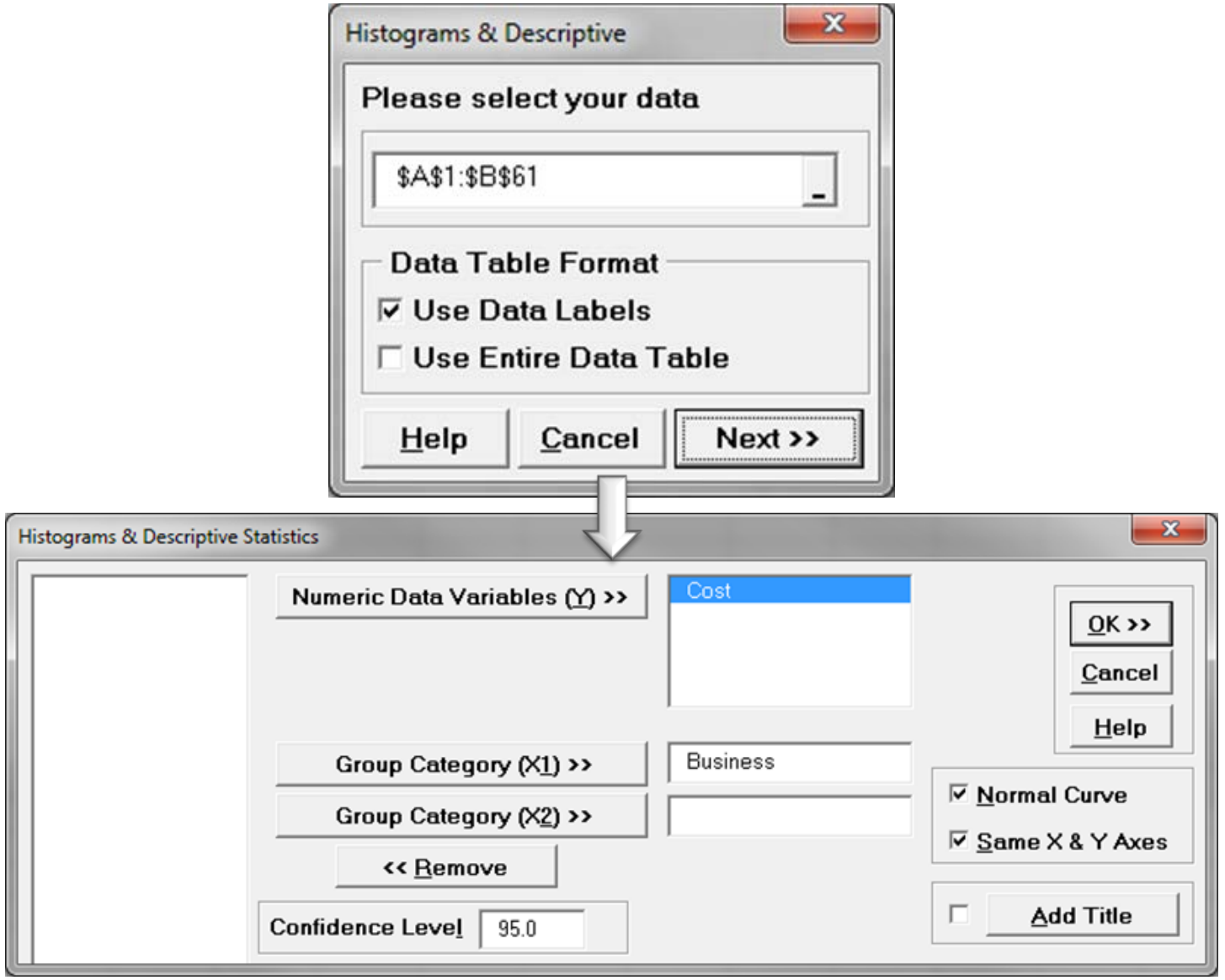
- Case Study: We are interested in comparing the average startup costs of five kinds of business.
  - Data File: “One-Way ANOVA” tab in “Sample Data.xlsx”

- Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- Alternative Hypothesis (H_a): at least one of the five means is different from others.

Use SigmaXL to Run ANOVA

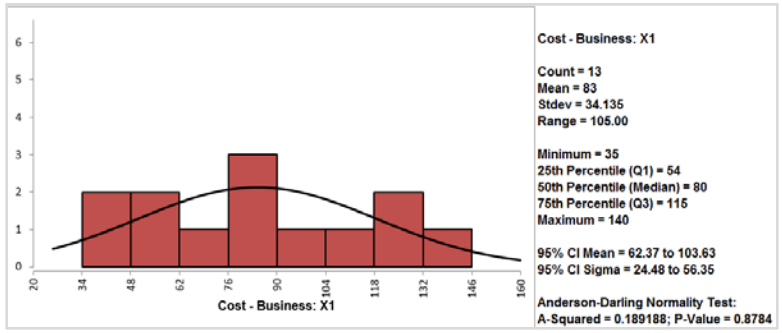
- Step 1 : Test whether the data for each level are normally distributed.
  - Select the entire range of data (both “Business” and “Cost”)
  - Click SigmaXL -> Graphical Tools -> Histogram & Descriptive Statistics
  - A new window named “Histogram & Descriptive” pops up with the selected range pre-populated
  - Click “Next>>”
  - A new window named “Histogram & Descriptive” appears
  - Select “Cost” as “Numeric Data Variables (Y)” and “Business” as “Group Category (X1)”
  - Click “OK>>”
  - The normality results appear in the new tab “Hist Descript (1)”

Use SigmaXL to Run ANOVA

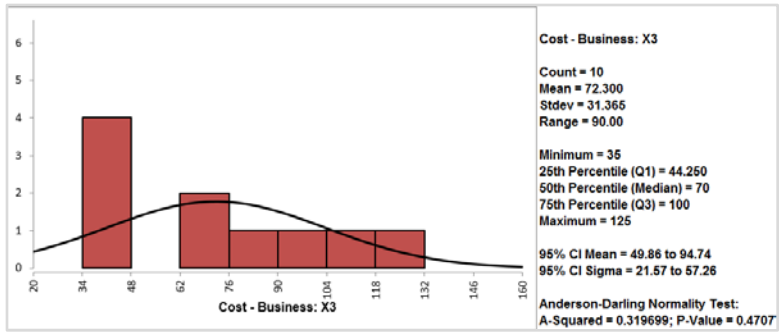


Use SigmaXL to Run ANOVA

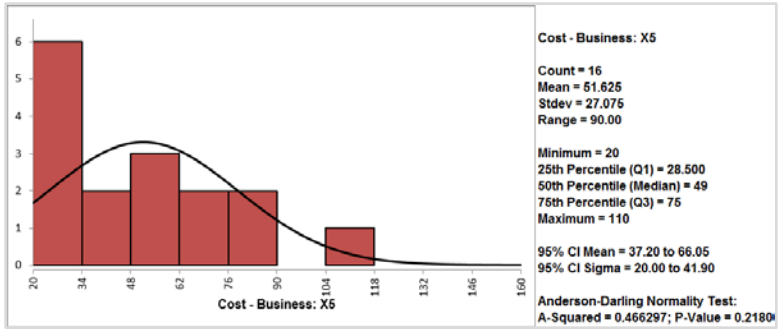
Business = X1:



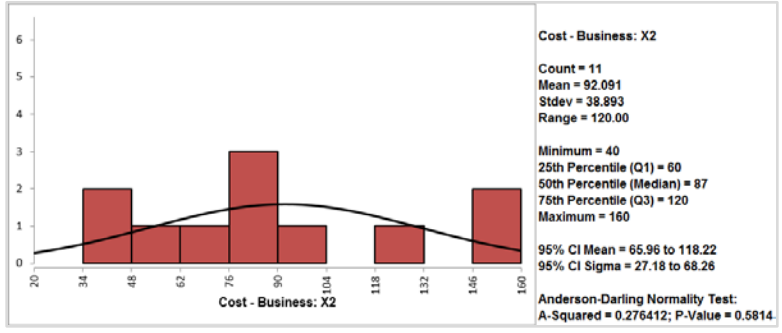
Business = X3:



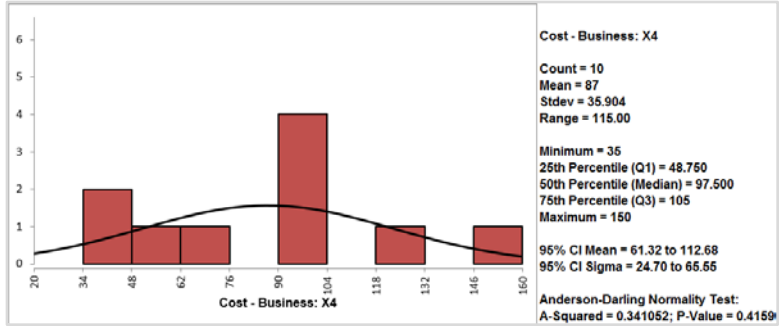
Business = X5:



Business = X2:



Business = X4:



Use SigmaXL to Run ANOVA

- Null Hypothesis (H_0): The data are normally distributed.
- Alternative Hypothesis (H_a): The data are not normally distributed.

- Since the p-values of normality tests for the five data sets are higher than alpha level (0.05), we fail to reject the null hypothesis and claim that the startup costs for any of the five businesses are normally distributed.
- If any of the five data sets are not normally distributed, we need to use other hypothesis testing methods other than one way ANOVA.

Use SigmaXL to Run ANOVA

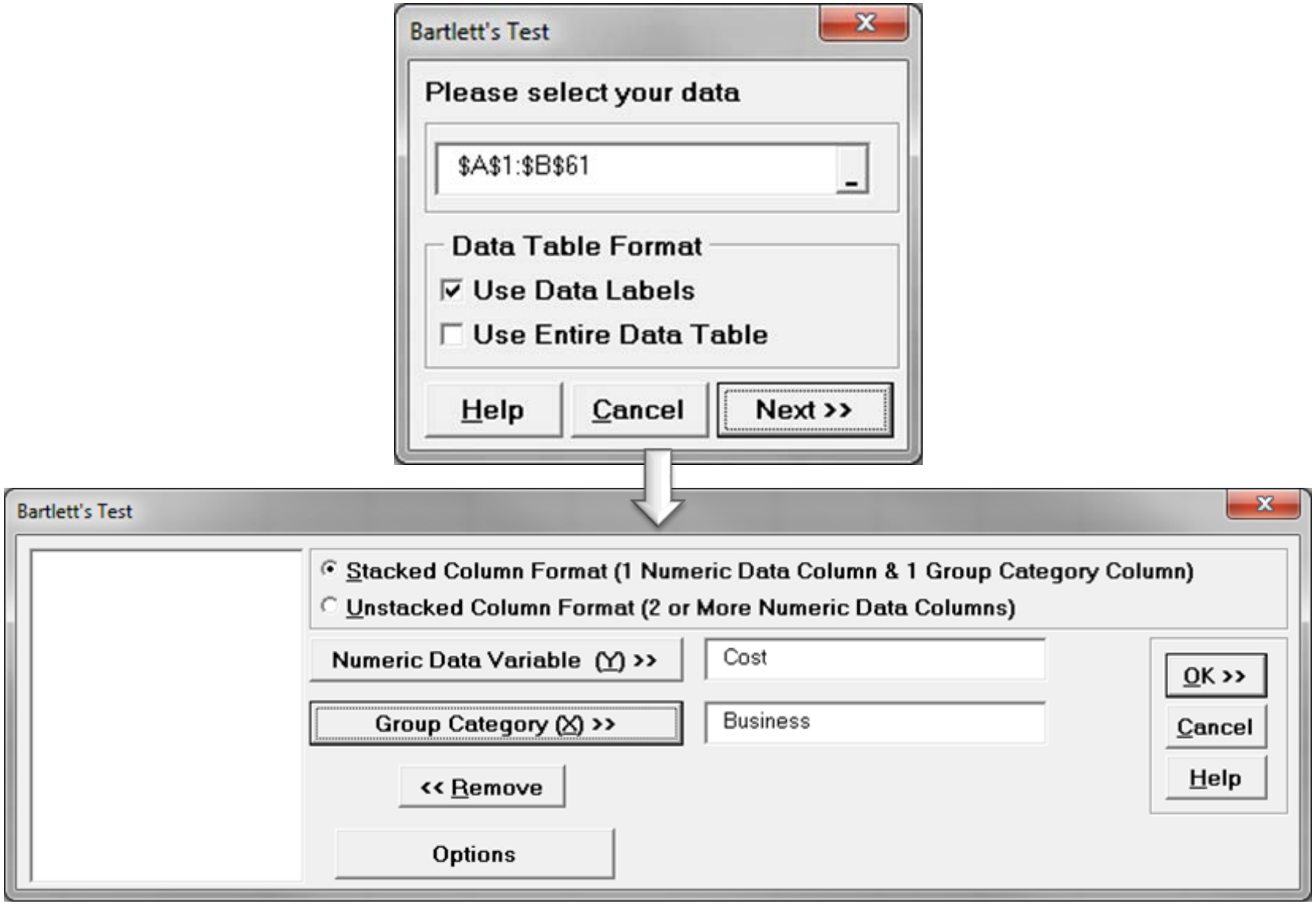
- Step 2: Test whether the variance of the data for each level is equal to the variance of other levels.

  Null Hypothesis (H_0): $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$

  Alternative Hypothesis (H_a): at least one of the variances is different from others.

  - Select the entire range of data (both “Business” and “Cost”)
  - Click SigmaXL -> Statistical Tools -> Bartlett’s Test (Since there are more than two levels in the data and the data of each level are normally distributed, we use Barlett’s test for testing the variances between the five levels.
  - A new window named “Bartlett’s Test” pops up with the selected range pre-populated
  - Click “Next>>”
  - A new window named “Bartlett’s Test” appears
  - Select “Cost” as the “Numeric Data Variable (Y)”
  - Select “Business” as the “Group Category (X)”
  - Click “OK>>”
  - The results shows up in the newly generated tab “Bartlett's Test (1)”

Use SigmaXL to Run ANOVA



Use SigmaXL to Run ANOVA

- The p-value of Barlett's test is 0.7768 greater than the alpha level (0.05), so we fail to reject the null hypothesis and we claim that the variances of five groups are equal.
- If the variances are not all equal, we need to use other hypothesis testing methods other than one way ANOVA.

Bartlett's Test For Equal Variance: Cost
(Use with normal data)

Test Information

H_0: Variance 1 = Variance 2 = ... = Variance k
Ha: At least one pair Variance i ≠ Variance j

Business	X1	X2	X3	X4	X5
Count	13	11	10	10	16
Mean	83	92.091	72.300	87	51.625
Median	80	87	70	97.500	49
StDev	34.135	38.893	31.365	35.904	27.075
AD Normality Test P-Value	0.8784	0.5814	0.4707	0.4159	0.2180

Bartlett's Test Statistic	1.776
P-Value	0.7768

Use SigmaXL to Run ANOVA

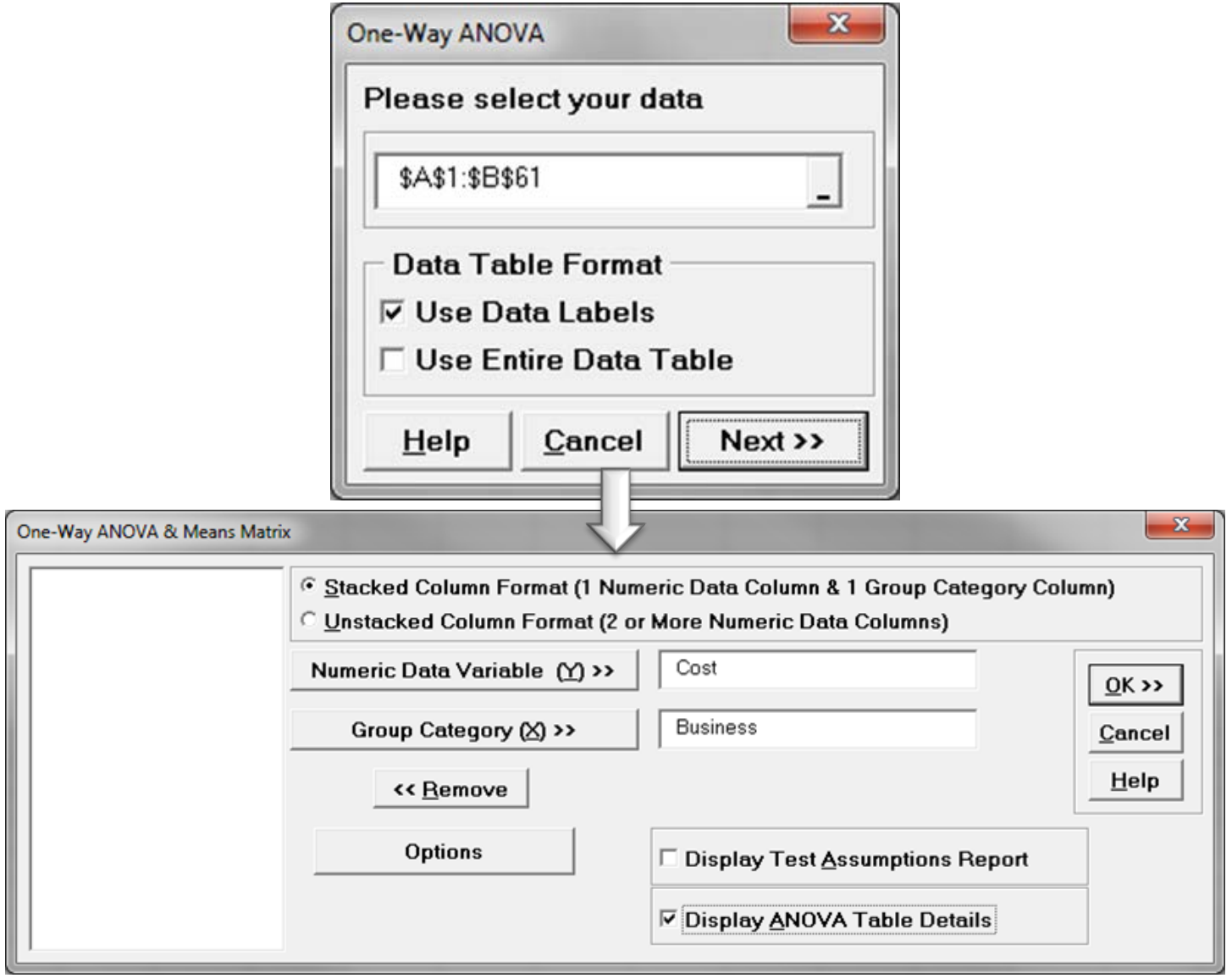
- Step 3: Test whether the mean of the data for each level is equal to the means of other levels.

  Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

  Alternative Hypothesis (H_a): at lease one of the means is different from others.

  - Select the entire range of the data (both “Business” and “Cost”
  - Click SigmaXL -> One-Way ANOVA & Means Matrix
  - A new window named “One-Way ANOVA” pops up with selected range pre-populated
  - Click “Next>>”
  - A new window named “One-Way ANOVA & Means Matrix” appears
  - Select “Cost” as “Numeric Data Variable (Y)”
  - Select “Business” as “Group Category (X)”
  - Check the checkbox “Display ANOVA Table Details”
  - Click “OK>>”
  - The ANOVA results appear in the newly generated tab “One-Way ANOVA (1)”

Use SigmaXL to Run ANOVA



Use SigmaXL to Run ANOVA

One-Way ANOVA & Means Matrix: Cost

H_0: Mean 1 = Mean 2 = ... = Mean k

H_a: At least one pair Mean i ≠ Mean j

Business	X1	X2	X3	X4	X5
Count	13	11	10	10	16
Mean	83	92.091	72.300	87	51.625
Standard Deviation	34.135	38.893	31.365	35.904	27.075
UC (2-sided, 95%, pooled)	101.44	112.14	93.329	108.03	68.250
LC (2-sided, 95%, pooled)	64.556	72.040	51.271	65.971	35.000

ANOVA Table

Source	SS	DF	MS	F	P-Value
Between	14298	4	3574.6	3.246	0.0184
Within	60561	55	1101.1		
Total	74859	59			
Pooled Standard Deviation =	33.183		R-Sq =	19.10%	
DF =	55		R-Sq adj. =	13.22%	

Since the p-value of the F test is 0.0184 lower than alpha level (0.05). The null hypothesis is rejected and we conclude that the at least one of the means of the five groups is different from others.

3.5 Hypothesis Testing: Non-Normal Data

Black Belt Training: Analyze Phase

3.1 Patterns of Variation
- 3.1.1 Multi-Vari Analysis
- 3.1.2 Classes of Distributions

3.2 Inferential Statistics
- 3.2.1 Understanding Inference
- 3.2.2 Sampling Techniques and Uses
- 3.2.3 Sample Size
- 3.2.4 Central Limit Theorem

3.3 Hypothesis Testing
- 3.3.1 Goals of Hypothesis Testing
- 3.3.2 Statistical Significance
- 3.3.3 Risk; Alpha and Beta
- 3.3.4 Types of Hypothesis Tests

3.4 Hypothesis Testing: Normal Data
- 3.4.1 One and Two Sample T-Tests
- 3.4.2 One sample variance
- 3.4.3 One Way ANOVA

3.5 Hyp Testing: Non-Normal Data
- 3.5.1 Mann-Whitney
- 3.5.2 Kruskal-Wallis
- 3.5.3 Moods Median
- 3.5.4 Friedman
- 3.5.5 One Sample Sign
- 3.5.6 One Sample Wilcoxon
- 3.5.7 One and Two Sample Proportion
- 3.5.8 Chi-Squared (Contingency Tables)
- 3.5.9 Test of Equal Variances

3.5.1 Mann-Whitney

What is the Mann-Whitney Test?

- The **Mann-Whitney test** (also called Mann-Whitney U test or Wilcoxon rank-sum test) is a statistical hypothesis test to compare the medians of two populations that are not normally distributed.

  - Null Hypothesis (H_0): $\eta_1 = \eta_2$
  - Alternative Hypothesis (H_a): $\eta_1 \neq \eta_2$

  where η_1 is the median of one population and η_1 is the median of the other population.

Mann-Whitney Test Assumptions

- The sample data drawn from the populations of interest are unbiased and representative.
- The data of both populations are continuous or ordinal when the spacing between adjacent values is not constant.
- The two populations are independent to each other.
- The Mann-Whitney test is robust for the non-normally distributed population.
- The Mann-Whitney test can be used when shapes of the two populations' distributions are different.

How Mann-Whitney Test Works

- Step 1: Group the two samples from two populations (sample 1 is from population 1 and sample 2 is from population 2) into a single data set and then sort the data in ascending order ranked from 1 to n, where n is the total number of observations.

- Step 2: Add up the ranks for all the observations from sample 1 and call it $\mathbf{R_1}$. Add up the ranks for all the observations from sample 2 and call it $\mathbf{R_2}$.

How Mann-Whitney Test Works

- Step 3: Calculate the test statistics

$$U = \min(U_1, U_2)$$

where $$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

n_1 and n_2 are the sample sizes.

R_1 and R_2 are the sum of ranks for observations from sample 1 and 2 respectively.

How Mann-Whitney Test Works

- Step 4: Make a decision on whether to reject the null hypothesis.
  - Null Hypothesis (H_0): $\eta_1 = \eta_2$
  - Alternative Hypothesis (H_a): $\eta_1 \neq \eta_2$

- If both of the sample sizes are smaller than 10, the distribution of U under the null hypothesis is tabulated.
  - The test statistic is U and, by using the Mann-Whitney table, we would find the p-value.
  - If the p-value is smaller than alpha level (0.05), we reject the null hypothesis.
  - If the p-value is greater than alpha level (0.05), we fail to reject the null hypothesis.

How Mann-Whitney Test Works

- If both sample sizes are greater than 10, the distribution of U can be approximated by a normal distribution. In other words, $\frac{U-\mu}{\sigma}$ follows a standard normal distribution.

$$Z_{calc} = \frac{U-\mu}{\sigma}$$

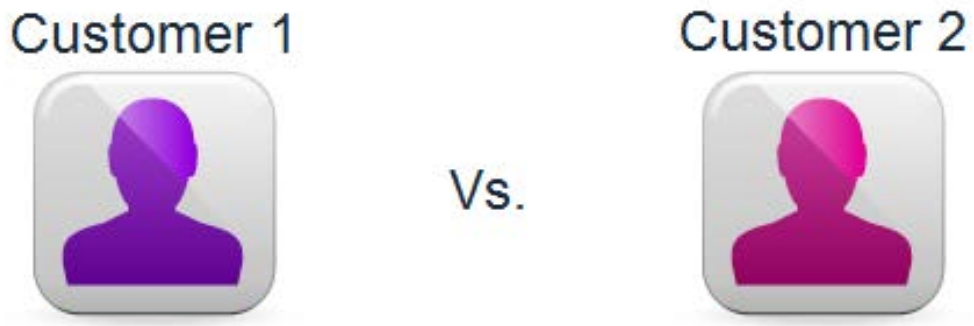
where

$$\mu = \frac{n_1 n_2}{2} \qquad \sigma = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

When $|Z_{calc}|$ is greater than Z value at α/2 level (e.g., when α = 5%, the z value we compare $|Z_{calc}|$ to is 1.96), we reject the null hypothesis.

Use SigmaXL to Run a Mann-Whitney Test

- Case Study: We are interested in comparing the customer satisfaction between two types of customers using nonparametric (i.e. distribution-fee) hypothesis test: Mann-Whitney test.
  - Data File: “Mann-Whitney” tab in “Sample Data.xlsx”

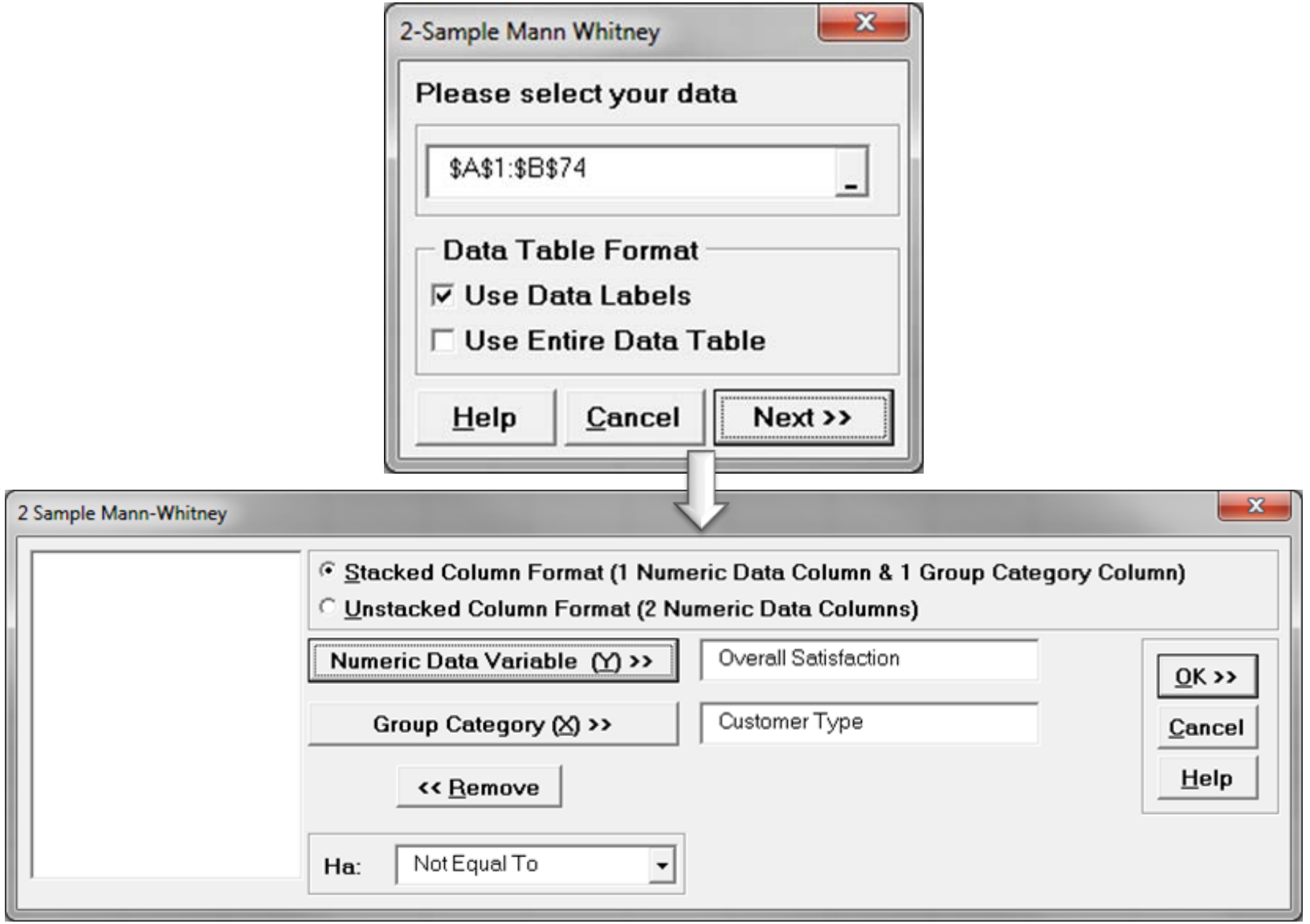


- Null Hypothesis (H_0): $\eta_1 = \eta_2$
- Alternative Hypothesis (H_a): $\eta_1 \neq \eta_2$

Use SigmaXL to Run a Mann-Whitney Test

- Steps to run a Mann-Whitney Test in SigmaXL:
  - Select the entire range of data (both “Customer Type” and “Overall Satisfaction”)
  - Click SigmaXL -> Statistical Tools -> Nonparametric Tests -> 2 Sample Mann-Whitney
  - A new window named “2 Sample Mann Whitney” pops up with the selected range populated in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “2 Sample Mann-Whitney” appears
  - Select “Overall Satisfaction” as the “Numeric Data Variables (Y)”
  - Select “Customer Type” as the “Group Category (X)”
  - Select “Not Equal To” as the “Ha”
  - Click “OK>>”
  - The Mann-Whitney test results appear in the newly generated tab “2 Sample Mann-Whitney (1)”

Use SigmaXL to Run a Mann-Whitney Test



Use SigmaXL to Run a Mann-Whitney Test

- The p-value of the test is lower than alpha level (0.05) so we reject the null hypothesis and conclude that there is statistically significant difference between the overall satisfaction medians of the two customer types.

2 Sample Mann-Whitney - Overall Satisfaction

Test Information		
H_0: Median Difference = 0		
H_a: Median Difference ≠ 0		
Customer Type	**1**	**2**
Count	31	42
Median	3.560	4.340
Mann-Whitney Statistic	772.50	
P-Value (2-sided, adjusted for ties)	0.0000	

3.5.2 Kruskal-Wallis

Kruskal-Wallis One-Way Analysis of Variance

- The **Kruskal-Wallis one-way analysis of variance** is a statistical hypothesis test to compare the medians among more than two groups.
  - Null Hypothesis (H_0): $\eta_1 = \eta_2 = \ldots = \eta_k$
  - Alternative Hypothesis (H_a): at least one of the medians is different from others.

  η_i is the median of population i, and k is the number of groups of our interest.

- It is an extension of Mann-Whitney test.
- If the distributions of k populations are not identically shaped or there are outliers in the distribution, Mood's median test is a more robust than Kruskal-Wallis.

Kruskal-Wallis: Assumptions

- The sample data drawn from the populations of interest are unbiased and representative.
- The data of k populations are continuous or ordinal when the spacing between adjacent values is not constant.
- The k populations are independent to each other.
- The Kruskal-Wallis test is robust for the non-normally distributed population.

How Kruskal-Wallis One-Way ANOVA Works

- Step 1: Group the k samples from k populations (sample i is from population i) into one single data set and then sort the data in ascending order ranked from 1 to N, where N is the total number of observations across k groups.

- Step 2: Add up the ranks for all the observations from sample i and call it $\mathbf{r_i}$, where i can be any integer between 1 and k.

How Kruskal-Wallis One-Way ANOVA Works

- Step 3: Calculate the test statistic

$$T = (N-1)\frac{\sum_{i=1}^{k} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

where

k is the number of groups.
n_i is the sample size of sample i.
N is the total number of all the observations across k groups.
r_{ij} is the rank (among all the observations) of observation j from group i.

$$\bar{r}_i = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i} \qquad \bar{r} = \frac{1}{2}(N+1)$$

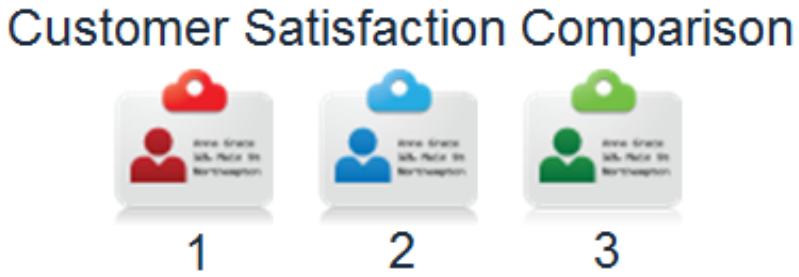
How Kruskal-Wallis One-Way ANOVA Works

- Step 4: Make a decision of whether to reject the null hypothesis.
  - Null Hypothesis (H_0): $\eta_1 = \eta_2 = \ldots = \eta_k$
  - Alternative Hypothesis (H_a): at least one of the medians is different from others.

- The test statistic follows chi-square distribution when the null hypothesis is true.
  - If T is greater than χ^2_{k-1} , we reject the null and claim there is at least one median statistically different from other medians.
  - If T is smaller than χ^2_{k-1} , we fail to reject the null and claim the medians of k groups are equal.

Use SigmaXL to Run a Kruskal-Wallis One-Way ANOVA

- Case Study: We are interested in comparing the customer satisfaction among three types of customers using nonparametric (i.e. distribution-fee) hypothesis test: Kruskal-Wallis one-way ANOVA.
  - Data File: “Kruskal-Wallis” tab in “Sample Data.xlsx”

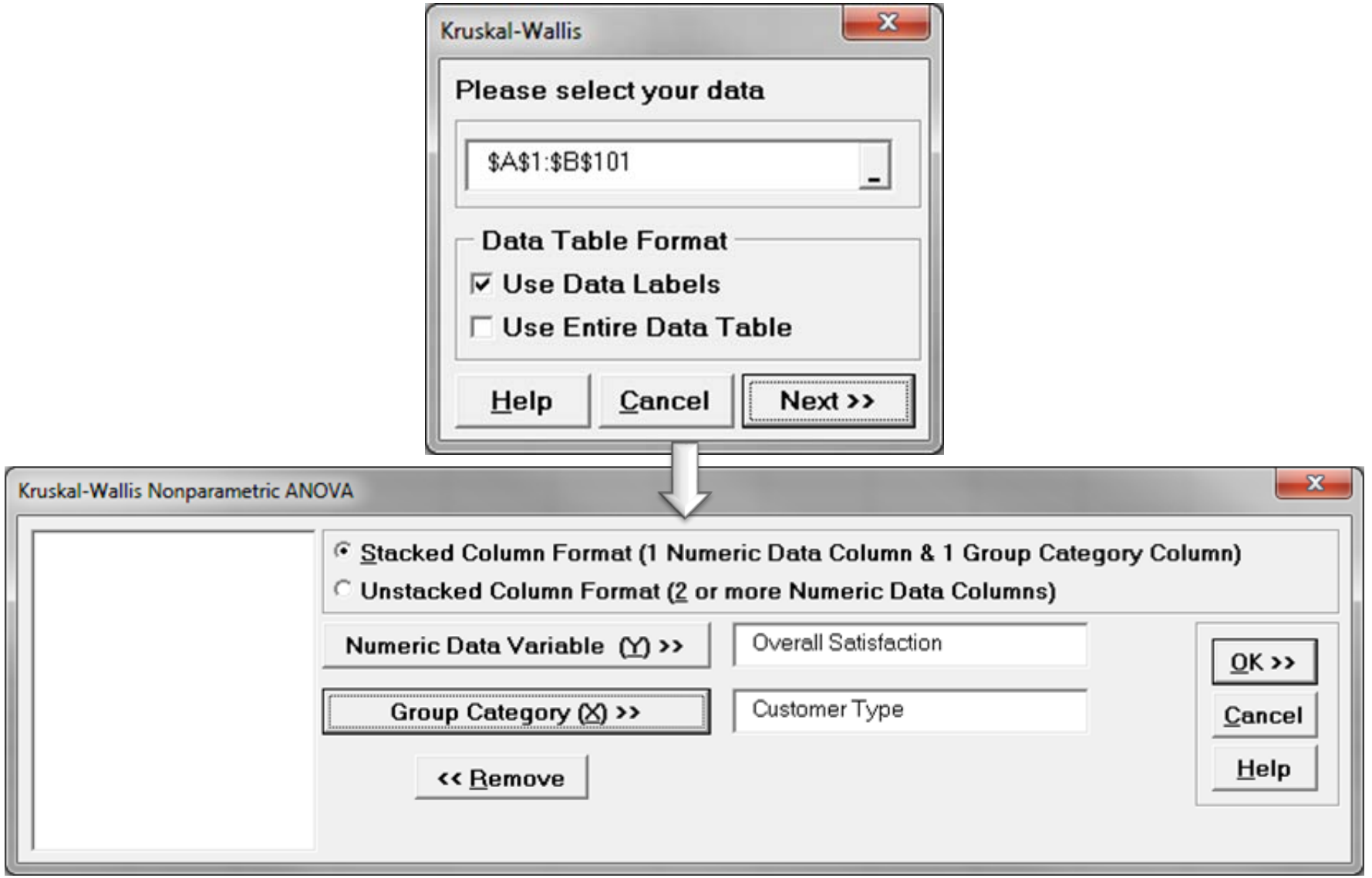


- Null Hypothesis (H_0): $\eta_1 = \eta_2 = \eta_3$
- Alternative Hypothesis (H_a): at least one of the customer type has different overall satisfaction from the others.

Use SigmaXL to Run a Kruskal-Wallis One-Way ANOVA

- Steps to run a Kruskal-Wallis One-Way ANOVA in SigmaXL
  - Select the entire range of data (both “Customer Type” and “Overall Satisfaction”
  - Click SigmaXL -> Statistical Tools -> Nonparametric Tests -> Kruskal-Wallis Test
  - A new window named “Kruskal-Wallis” pops up with the selected range appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Nonparametric ANOVA” appears
  - Select “Overall Satisfaction” as the “Numeric Data Variables (Y)”
  - Select “Customer Type” as the “Group Category (X)”
  - Click “OK”
  - The Kruskal-Wallis test results appear in the newly generated tab “Kruskal-Wallis (1)”

Use SigmaXL to Run a Kruskal-Wallis One-Way ANOVA



Use SigmaXL to Run a Kruskal-Wallis One-Way ANOVA

- The p-value of the test is lower than alpha level (0.05) and we reject the null hypothesis and conclude that at least the overall satisfaction median of one customer type is statistically different from the others.

Kruskal-Wallis Nonparametric ANOVA: Overall Satisfaction

Test Information

H_0: Median 1 = Median 2 = ... = Median k

H_a: At least one pair Median i ≠ Median j

Customer Type	1	2	3
Count (N)	31	42	27
Median	3.56	4.34	3.51
UC Median (2-sided, 95%)	3.936	4.518	4.023
LC Median (2-sided, 95%)	2.954	4.095	3.289
Z	-3.339	4.529	-1.557
Kruskal-Wallis Statistic (H)	21.360		
DF	2		
P-Value (2-sided, adjusted for ties)	0.0000		

3.5.3 Mood’s Median

What is Mood's Median Test?

- **Mood's median test** is a statistical test to compare the medians of two or more populations.
  - Null Hypothesis (H_0): $\eta_1 = \ldots = \eta_k$
  - Alternative Hypothesis (H_a): at least one of the medians is different from the others.
  - k is the number of groups of our interest and is equal to or greater than two.
- Mood's median is an alternative to Kruskal-Wallis.
- It is the extension of one sample sign test.
- For the data with outliers, Mood's median test is more robust than the Kruskal-Wallis test.

Mood's Median Test Assumptions

- The sample data drawn from the populations of interest are unbiased and representative.
- The data of k populations are continuous or ordinal when the spacing between adjacent values is not constant.
- The k populations are independent to each other.
- The distributions of k populations have the same shape.
- Mood's median test is robust for non-normally distributed populations.
- Mood's median test is robust for data with outliers.

How Mood's Median Test Works

- Step 1: Group the k samples from k populations (sample i is from population i) into one single data set and get the median of this combined data set.

- Step 2: Separate the data in each sample into two groups. One consists of all the observations with values higher than the grand median and the other consists of all the observations with values lower than the grand median.

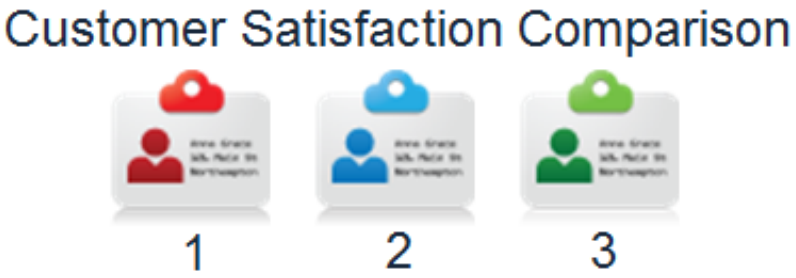
How Mood's Median Test Works

- Step 3: Run a Pearson's chi-square test to determine whether to reject the null hypothesis.
  - Null Hypothesis (H_0): $\eta_1 = \ldots = \eta_k$
  - Alternative Hypothesis (H_a): at least one of the medians is different from the others.

- If χ^2_{calc} is greater than χ^2_{crit}, we reject the null hypothesis and claim that at least one median is different from the others.
- If χ^2_{calc} is smaller than χ^2_{crit}, we fail to reject the null hypothesis and claim that the medians of k populations are not statistically different.

Use SigmaXL to Run a Mood's Median Test

- Case Study: We are interested in comparing the customer satisfaction among three types of customers using nonparametric (i.e. distribution-fee) hypothesis test: Mood's median test.
  - Data File: "Median Test" tab in "Sample Data.xlsx"

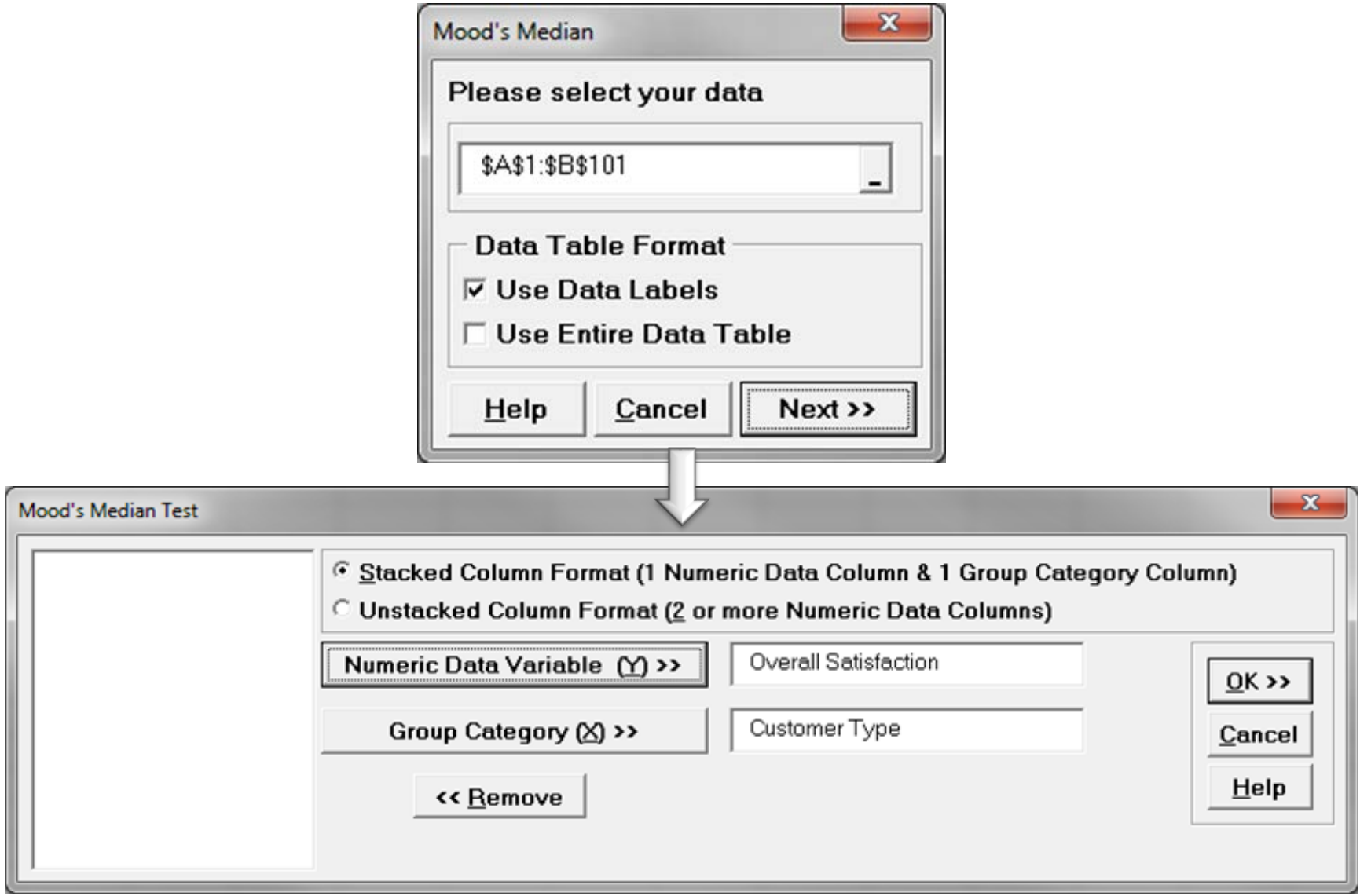


- Null Hypothesis (H_0). $\eta_1 = \eta_2 = \eta_3$
- Alternative Hypothesis (H_a): at least one of the customer type has different overall satisfaction from the others.

Use SigmaXL to Run a Mood's Median Test

- Steps to run a Mood's median test in SigmaXL
  - Select the entire range of data (both "Customer Type" and "Overall Satisfaction"
  - Click SigmaXL -> Statistical Tools -> Nonparametric Tests -> Mood's Median Test
  - A new window named "Mood's Median" pops up with the selected range appearing in the box under "Please select your data"
  - Click "Next>>"
  - A new window named "Mood's Median Test" appears
  - Select "Overall Satisfaction" as the "Numeric Data Variables (Y)"
  - Select "Customer Type" as the "Group Category (X)"
  - Click "OK"
  - The Mood's median test results appear in the newly generated tab "Mood's Median (1)"

Use SigmaXL to Run a Mood's Median Test



Use SigmaXL to Run a Mood's Median Test

- The p-value of the test is lower than alpha level (0.05) and we reject the null hypothesis and conclude that at least the overall satisfaction median of one customer type is statistically different from the others.

Mood's Median Test: Overall Satisfaction

Test Information

H_0: Median 1 = Median 2 = ... = Median k

H_a: At least one pair Median i ≠ Median j

Customer Type	1	2	3
Count (N <= Overall Median)	21	12	17
Count (N > Overall Median)	10	30	10
Median	3.560	4.340	3.510
UC Median (2-sided, 95%)	3.936	4.518	4.023
LC Median (2-sided, 95%)	2.954	4.095	3.289

Overall Median	3.945
Chi-Square	13.432
DF	2
P-Value (2-sided)	0.0012

3.5.4 Friedman

What is the Friedman Test?

- The **Friedman test** is a hypothesis test used to detect the differences in various groups across multiple attempts.
  - Null Hypothesis (H_0): The treatments have identical effects (i.e. $\eta_1 = \ldots = \eta_k$).
  - Alternative Hypothesis (H_a): At least one of the treatments has different effects from the others (i.e., at least one of the medians is statistically different from others).

- It is used as an alternative of the parametric repeated measures ANOVA when the assumption of normality or variance equality is not met.

Friedman Test Assumptions

- The sample data drawn from the populations of interest are unbiased and representative.
- The data are continuous or ordinal when the spacing between adjacent values is not constant.
- Results in one block are independent of the results in another.
- The Friedman test is robust for the non-normally distributed population.
- The Friedman test is robust for populations with unequal variances.

How Friedman Test Works

- Step 1: Organize the data into a tabular view with n rows indicating the blocks and k columns indicating the treatments. Each observation x_{ij} is filled into the intersection of a specific block i and a specific treatment j.

- Step 2: Calculate the ranks of each observation within each block.

- Step 3: Replace the values in the table created in step 1 with the order r_{ij} (within a block) of the corresponding observation x_{ij}.

How Friedman Test Works

- Step 4: Calculate the test statistic

$$Q = \frac{n\sum_{j=1}^{k}(\bar{r}_j - \bar{r})^2}{\frac{1}{n(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k}(r_{ij} - \bar{r})^2}$$

where

$$\bar{r}_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$$

$$\bar{r} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} r_{ij}$$

How Friedman Test Works

- Step 5: Make a decision of whether to reject the null hypothesis.
  - When n > 15 or k > 4, the test statistic Q follows a chi-square distribution if the null hypothesis is true and the p-value is $P\left(\chi^2_{k-1} \geq Q\right)$
  - When n < 15 and k < 4, the test statistic Q does not approximate a chi-square distribution and the p-value can be obtained from tables of Q for the Friedman test.
  - If p-value > alpha level (0.05), we fail to reject the null.

Friedman Test Examples

- A number of n water testers judge the quality of k different water samples, each of which is from a distinct water source. We will apply a Friedman Test to determine whether the water qualities of the k sources are the same.
- There are n blocks and k treatments in this experiment.
- One tester's decision would not have any influence on other testers.
- When running the experiment, each tester judges the water in a random sequence.

3.5.5 One Sample Sign

What is the One Sample Sign Test?

- The **one sample sign test** is a hypothesis test to compare the median of a population with a specified value.
  - Null Hypothesis (H_0): $\eta = \eta_0$
  - Alternative Hypothesis (H_a): $\eta \neq \eta_0$

- It is an alternative test of one sample t-test when the distribution of the data is non-normal. It is robust for the data with non-symmetric distribution.

One Sample Sign Test Assumptions

- The sample data drawn from the population of interest are unbiased and representative.
- The data are continuous or ordinal when the spacing between adjacent values is not constant.
- The one sample sign test is robust for the non-normally distributed population.
- The one sample sign test does not have any assumptions on the distribution. It is a distribution-free test.

How the One Sample Sign Test Works

- Step 1: Separate the sample set of data into two groups: one with values greater than and the other with values less than the hypothesized median η_0. Count the number of observations in each group.

- Step 2: Calculate the test statistic.
  - If the null hypothesis is true, the number of observations in each group should not be significantly different from half of the total sample size.
  - The test statistic follows a binomial distribution when the null is true. When n is large, we use the normal distribution to approximate the binomial distribution.

How the One Sample Sign Test Works

Test Statistic:

$$Z_{calc} = \frac{n_+ - np}{\sqrt{np(1-p)}}$$

where

n_+ is the number of observations with values greater than the hypothesized median.

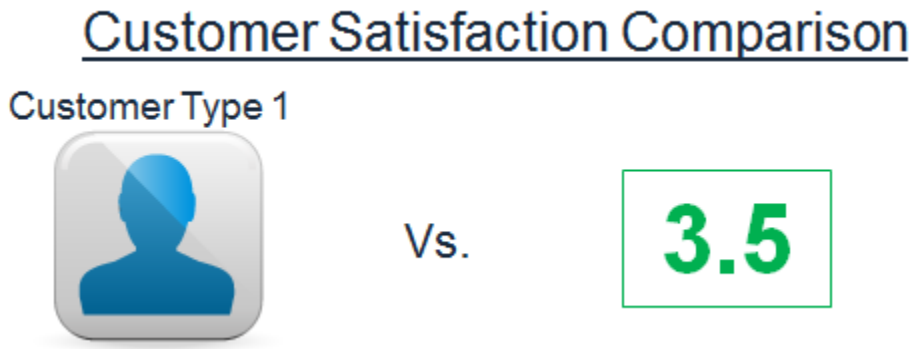
n is the total number of observations .

p is 0.5.

- Step 3: Make a decision on whether to reject the null hypothesis. If the $|Z_{calc}|$ is smaller than Z_{crit}, we fail to reject the null hypothesis and claim that there is no significant difference between the population median and the hypothesized median.

Use SigmaXL to Run a One Sample Sign Test

- Case Study: We are interested in comparing the overall satisfaction of customer type 1 against a specified benchmark satisfaction (3.5) using nonparametric (i.e. distribution-fee) hypothesis test: 1 Sample Sign test.
  - Data File: “One Sample Wilcoxon” tab in “Sample Data.xlsx”

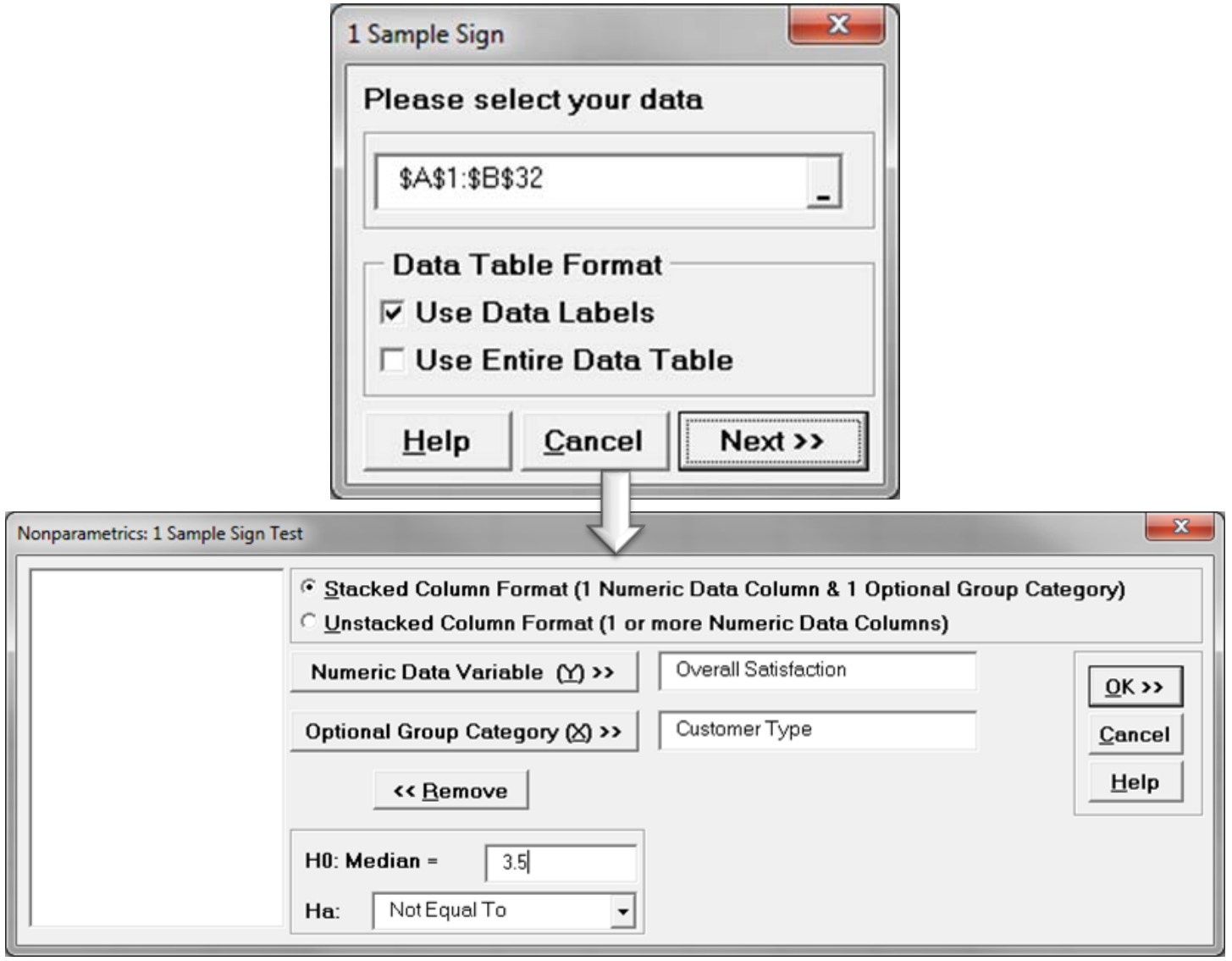


- Null Hypothesis (H_0): $\eta_1 = 3.5$
- Alternative Hypothesis (H_a): $\eta_1 \neq 3.5$

Use SigmaXL to Run a One Sample Sign Test

- Steps to run a one sample sign test in SigmaXL
  - Select the entire range of data (both “Customer Type” and “Overall Satisfaction”
  - Click SigmaXL -> Statistical Tools -> Nonparametric Tests -> 1 Sample Sign
  - A new window named “1 Sample Sign” pops up with the selected range appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “1 Sample Sign Test” appears
  - Select “Overall Satisfaction” as the “Numeric Data Variables (Y)”
  - Select “Customer Type” as the “Group Category (X)”
  - Enter “3.5” in the box next to “H0: Median=“
  - Click “OK”
  - The one sample sign test results appear in the newly generated tab “1 Sample Sign Test (1)”

Use SigmaXL to Run a One Sample Sign Test



Use SigmaXL to Run a One Sample Sign Test

- The p-value of the one sample sign test is 1.00 higher than the alpha level (0.05) and we fail to reject the null. There is not any statistically significant difference between the overall satisfaction of customer type 1 and the benchmark satisfaction level.

1 Sample Sign Test for Medians: Overall Satisfaction

Test Information	
H_0: Median = 3.5	
H_a: Median Not Equal To 3.5	

Customer Type	1
Count (N)	31
Median	3.560
Points Below 3.5	15
Points Equal To 3.5	0
Points Above 3.5	16
P-Value (2-sided)	1.0000

3.5.6 One Sample Wilcoxon

What is the One Sample Wilcoxon Test?

- The **one sample Wilcoxon test** is a hypothesis test to compare the median of one population with a specified value.
  - Null Hypothesis (H_0): $\eta = \eta_0$
  - Alternative Hypothesis (H_a): $\eta \neq \eta_0$

- It is an alternative test of one sample t-test when the distribution of the data is non-normal.

- It is more powerful than one sample sign test but it assumes the distribution of the data is symmetric.

One Sample Wilcoxon Test Assumptions

- The sample data drawn from the population of interest are unbiased and representative.
- The data are continuous or ordinal when the spacing between adjacent values is not constant.
- The distribution of the data is symmetric about a median.
- The one sample Wilcoxon test is robust for the non-normally distributed population.

How the One Sample Wilcoxon Test Works

- Step 1: Create the following columns one by one:
  - Column 1: all the raw observations (X)
  - Column 2: the differences between each observation value and the hypothesized median ($X - \eta_0$)
  - Column 3: the signs (+ or –) of column 2
  - Column 4: the absolute value of column 2
  - Column 5: the ranks of each item in column 4 in ascending order
  - Column 6: the product of column 3 and column 5

How the One Sample Wilcoxon Test Works

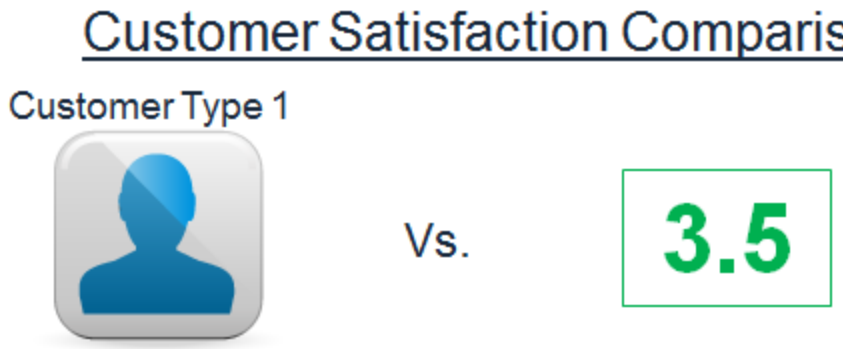
- Step 2: Calculate the test statistic W_{calc}, which is the sum of all the non-negative values in column 6.

- Step 3: Make a decision on whether to reject the null hypothesis. Use the table of critical values for the Wilcoxon test to get the W_{crit} with predetermined alpha level and number of observations.

  - If the W_{calc} is smaller than the W_{crit}, we fail to reject the null hypothesis and claim that there is no significant difference between the population median and the hypothesized median.

Use SigmaXL to Run a One Sample Wilcoxon Test

- Case Study: We are interested in comparing the overall satisfaction of customer type 1 against a specified benchmark satisfaction (3.5) using nonparametric (i.e. distribution-fee) hypothesis test: one sample Wilcoxon test.
  - Data File: “One Sample Wilcoxon” tab in “Sample Data.xlsx”

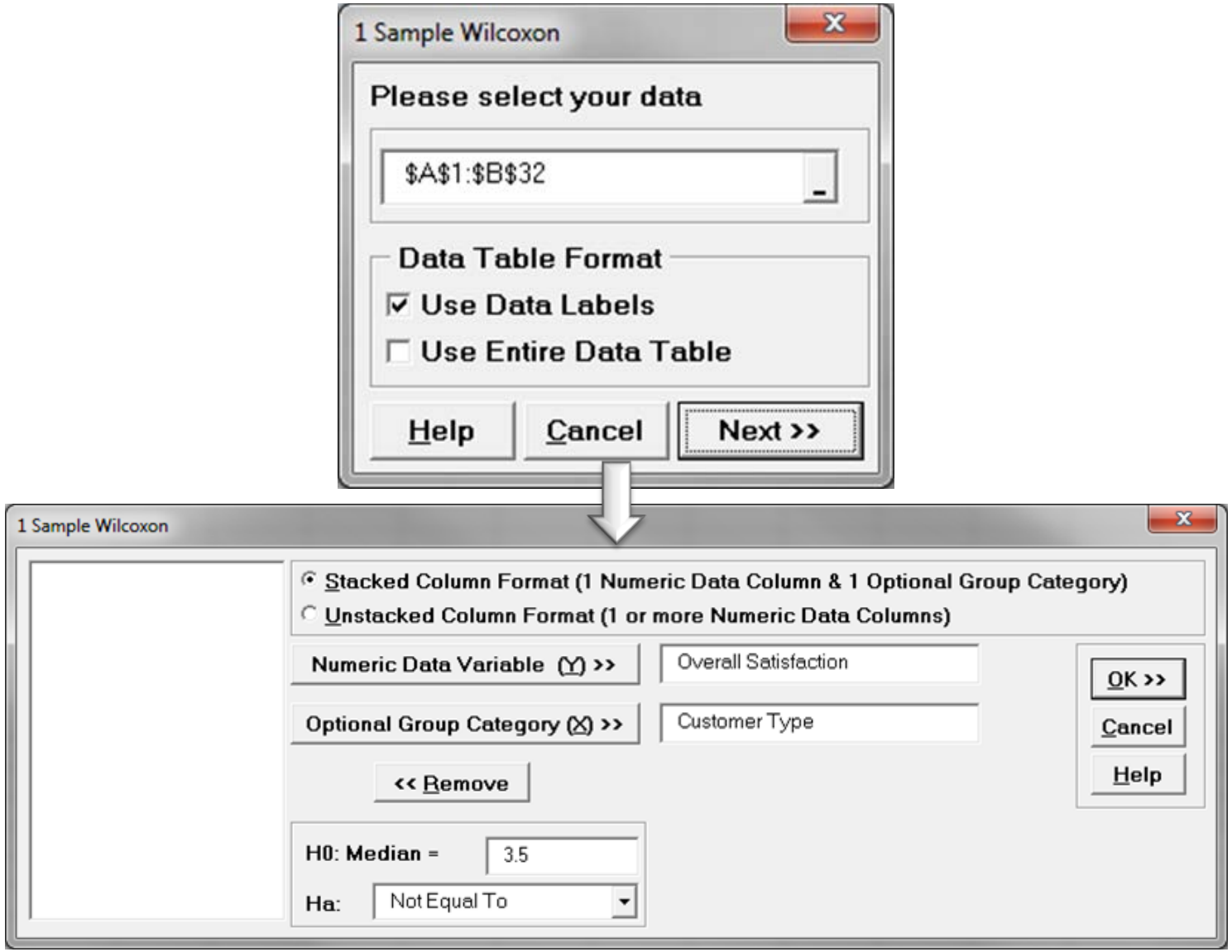


- Null Hypothesis (H_0): $\eta_1 = 3.5$
- Alternative Hypothesis (H_a): $\eta_1 \neq 3.5$

Use SigmaXL to Run a One Sample Wilcoxon Test

- Steps to run a one sample Wilcoxon test in SigmaXL
  - Select the entire range of data (both “Customer Type” and “Overall Satisfaction”
  - Click SigmaXL -> Statistical Tools -> Nonparametric Tests -> 1 Sample Wilcoxon
  - A new window named “1 Sample Wilcoxon” pops up with the selected range appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window also named “1 Sample Wilcoxon” appears
  - Select “Overall Satisfaction” as the “Numeric Data Variables (Y)”
  - Select “Customer Type” as the “Group Category (X)”
  - Enter “3.5” in the box next to “H0: Median=“
  - Click “OK”
  - The one sample Wilcoxon test results appear in the newly generated tab “1 Sample Wilcoxon (1)”

Use SigmaXL to Run a One Sample Wilcoxon Test



Use SigmaXL to Run a One Sample Wilcoxon Test

- The p-value of the one sample Wilcoxon test is 0.5566 higher than the alpha level (0.05) and we fail to reject the null. There is not any statistically significant difference between the overall satisfaction of customer type 1 and the benchmark satisfaction level.

1 Sample Wilcoxon Test: Overall Satisfaction

Test Information	
H_0: Median = 3.5	
H_a: Median Not Equal To 3.5	
Customer Type	**1**
Count (N)	31
Count for Test	31
Median	3.56
Wilcoxon Statistic	217.50
P-Value (2-sided)	0.5566

3.5.7 One & Two Sample Proportion

What is the One Sample Proportion Test?

- **One sample proportion test** is a hypothesis test to compare the proportion of one certain outcome occurring in a population following the binomial distribution with a specified proportion.

  - Null Hypothesis (H_0): $p = p_0$
  - Alternative Hypothesis (H_a): $p \neq p_0$

One Sample Proportion Test Assumptions

- The sample data drawn from the population of interest are unbiased and representative.
- There are only two possible outcomes in each trial: success/failure, yes/no, and defective/non-defective etc.
- The underlying distribution of the population is binomial distribution.
- When $np \geq 5$ and $np(1 - p) \geq 5$, the binomial distribution can be approximated by the normal distribution.

How the One Sample Proportion Test Works

When np ≥ 5 and np(1 – p) ≥ 5, we use normal distribution to approximate the underlying binomial distribution of the population.

Test Statistic: $$Z_{calc} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where

$\hat{p}$ is the observed probability of one certain outcome occurring.

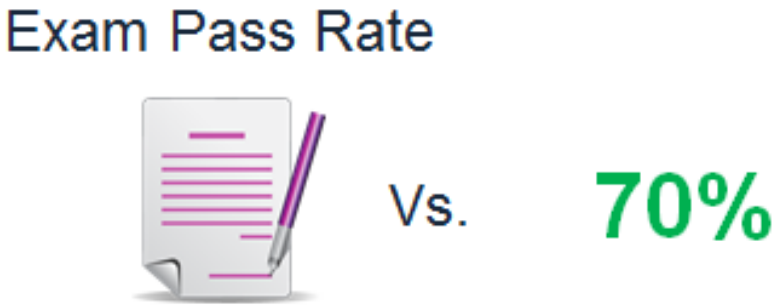
p_0 is the hypothesized probability.

n is number of trials.

When $|Z_{calc}|$ is smaller than Z_{crit}, we fail to reject the null hypothesis and claim that there is no statistically significant difference between the population proportion and the hypothesized proportion.

Use SigmaXL to Run a One Sample Proportion Test

- Case Study: We are interested in comparing the exam pass rate of a high school this month against a specified rate (70%) using nonparametric (i.e. distribution-fee) hypothesis test: one sample proportion test.
  - Data File: “One Sample Proportion” tab in “Sample Data.xlsx”



- Null Hypothesis (H_0): p = 70%
- Alternative Hypothesis (H_a): p ≠ 70%

Use SigmaXL to Run a One Sample Proportion Test

- Steps to run a one sample proportion test in SigmaXL.
  - Click SigmaXL -> Statistical Tools -> Basic Statistical Templates -> 1 Proportion Test and Confidence Interval
  - A new tab named “1 Proportion Test CI” appears
  - Enter “77” in the yellow box of “Number of Events in category of interest x”
  - Enter ‘105” in the yellow box of “Sample Size”
  - Enter 0.70 as “hypothesized proportion” test against 70%
  - Keep 95% as the default confidence level
  - See next page for steps as results..

Use SigmaXL to Run a One Sample Proportion Test

- The interpretation highlighted at bottom right, says “we are 95% confident that the true proportion lies between 0.6381 and 0.8149”
- The p-value is 0.5286, therefore, we claim that there is not any statistically significant difference between the school’s exam passing rate and 70%.

1 Proportion Test and Confidence Interval		
Sample Data (user inputs):		
Number of Events	x	77
Sample Size	n	105
Null Hypothesis (hypothesized proportion)	H_0: Proportion =	0.7
Alternative Hypothesis	H_a: Proportion	Not Equal To
Confidence Level (enter .95 for 95%)	100*(1-α)%	95.0%
Hypothesis Test Method		Binomial Exact
Confidence Interval Method		Exact (Clopper-Pearson Beta)
Results:		
Sample proportion (x/n)		0.7333
alpha		0.0500
npq (npq should be >= 5 for normal approximation; q = 1-p)		20.5333
Z-statistic (normal)		0.7454
Binomial exact probability P-Value (2-sided)		0.5286
Upper Confidence Limit (2-sided)		0.8149
Lower Confidence Limit (2-sided)		0.6381

What is the Two Sample Proportion Test?

- The **two sample proportion test** is a hypothesis test to compare the proportions of one certain event occurring in two populations following the binomial distribution.

  - Null Hypothesis (H_0): $p_1 = p_2$
  - Alternative Hypothesis (H_a): $p_1 \neq p_2$

Two Sample Proportion Test Assumptions

- The sample data drawn from the populations of interest are unbiased and representative.
- There are only two possible outcomes in each trial for both populations: success/failure, yes/no, and defective/non-defective etc.
- The underlying distributions of both populations are binomial distribution.
- When $np \geq 5$ and $np(1 - p) \geq 5$, the binomial distribution can be approximated by the normal distribution.

How the Two Sample Proportion Test Works

- When np ≥ 5 and np(1 – p) ≥ 5, we use normal distribution to approximate the underlying binomial distributions of the populations.

Test Statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0\left(1-\hat{p}_0\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where $\hat{p}_0 = \frac{x_1 + x_2}{n_1 + n_2}$

$\hat{p}_1$ and $\hat{p}_2$ are the observed proportions of events in the two samples.

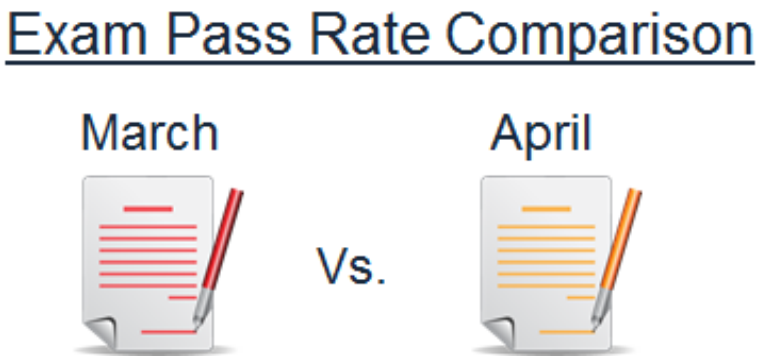
n_1 and n_2 are the number of trials in the two samples respectively.

x_1 and x_2 are the number of events in the two samples respectively.

When $|Z_{calc}|$ is smaller than Z_{crit}, we fail to reject the null hypothesis.

Use SigmaXL to Run a Two Sample Proportion Test

- Case Study: We are interested in comparing the exam pass rates of a high school in March and April using nonparametric (i.e. distribution-fee) hypothesis test: two sample proportion test.
  - Data File: “Two Sample Proportion” tab in “Sample Data.xlsx”



- Null Hypothesis (H_0): $p_{March} = p_{April}$
- Alternative Hypothesis (H_a): $p_{March} \neq p_{April}$

Use SigmaXL to Run a Two Sample Proportion Test

- Steps to run a two sample proportion test in SigmaXL
  - Click SigmaXL -> Statistical Tools -> 2 Proportions Test & Confidence Interval
  - A new tab named “2 Proportions Test and CI” appears automatically.
  - Enter “89” in the yellow box of “Number of Events” in Sample #1 column
  - Enter “112” in the yellow box of “Sample Size in Sample #1 column.
  - Enter “102” in the yellow box of “Number of Events” in Sample #2 column.
  - Enter “130” in the yellow box of “Sample Size” in Sample #2 column.
  - Select “Fisher’s Exact” as the testing method.

Use SigmaXL to Run a Two Sample Proportion Test

- Fisher's exact p-value (2-sided, Ha: P1 ≠ P2) is 0.8756 and higher than the alpha level of 0.05.
- Therefore, we fail to reject the null and we claim that the exam pass rates in March and April are not statistically different.

2 Proportions Test and Confidence Interval

Sample Data (user inputs):		Sample 1	Sample 2
Number of Events	x	89	102
Sample Size	n	112	130
Null Hypothesis (hypothesized difference)	H_0: $P_1 - P_2$ =	0	
Alternative Hypothesis	H_a: $P_1 - P_2$	Not Equal To	
Confidence Level (enter .95 for 95%)	100*(1-α)%	95.0%	
Hypothesis Test Method		Fisher's Exact	
Confidence Interval Method		Newcombe-Wilson Score	

Results:		
Sample proportion (x/n)	0.7946	0.7846
Sample proportion difference	0.0100	
alpha	0.0500	
Minimum expected value (should be >= 5 for normal approximation)	23.6033	
Fisher's Exact probability P-Value (2-sided)	0.8756	
Upper Confidence Limit (2-sided)	0.1114	
Lower Confidence Limit (2-sided)	-0.0943	

3.5.8 Chi-Squared (Contingency Tables)

What is the Chi-Square Test?

- **A chi-square test** is a hypothesis test in which the sampling distribution of the test statistic follows a chi-square distribution when the null hypothesis is true.

- There are multiple chi-square tests available and in this module we will cover the Pearson's chi-square test used in contingency analysis.

  - Null Hypothesis (H_0): $p_1 = p_2 = \ldots = p_k$
  - Alternative Hypothesis (H_a): at least on of the proportions is different from others.
  - k is the number of populations of our interest. $k \geq 2$.

What is the Chi-Square Test?

- The chi-square test can also be used to test whether two factors are independent of each other. In other words, it can be used to test whether there is any statistically significant relationship between two discrete factors.

  - Null Hypothesis (H_0): Factor 1 is independent of factor 2.
  - Alternative Hypothesis (H_a): Factor 1 is not independent of factor 2.

Chi-Square Test Assumptions

- The sample data drawn from the populations of interest are unbiased and representative.
- There are only two possible outcomes in each trial for an individual population: success/failure, yes/no, and defective/non-defective etc.
- The underlying distribution of each population is binomial distribution.
- When $np \geq 5$ and $np(1 - p) \geq 5$, the binomial distribution can be approximated by the normal distribution.

How Chi-Square Test Works

- Test Statistic

$$\chi^2_{calc} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where

O_i is an observed frequency.

E_i is an expected frequency.

N is the number of cells in the contingency table.

If χ^2_{calc} is smaller than χ^2_{crit}, we fail to reject the null hypothesis.

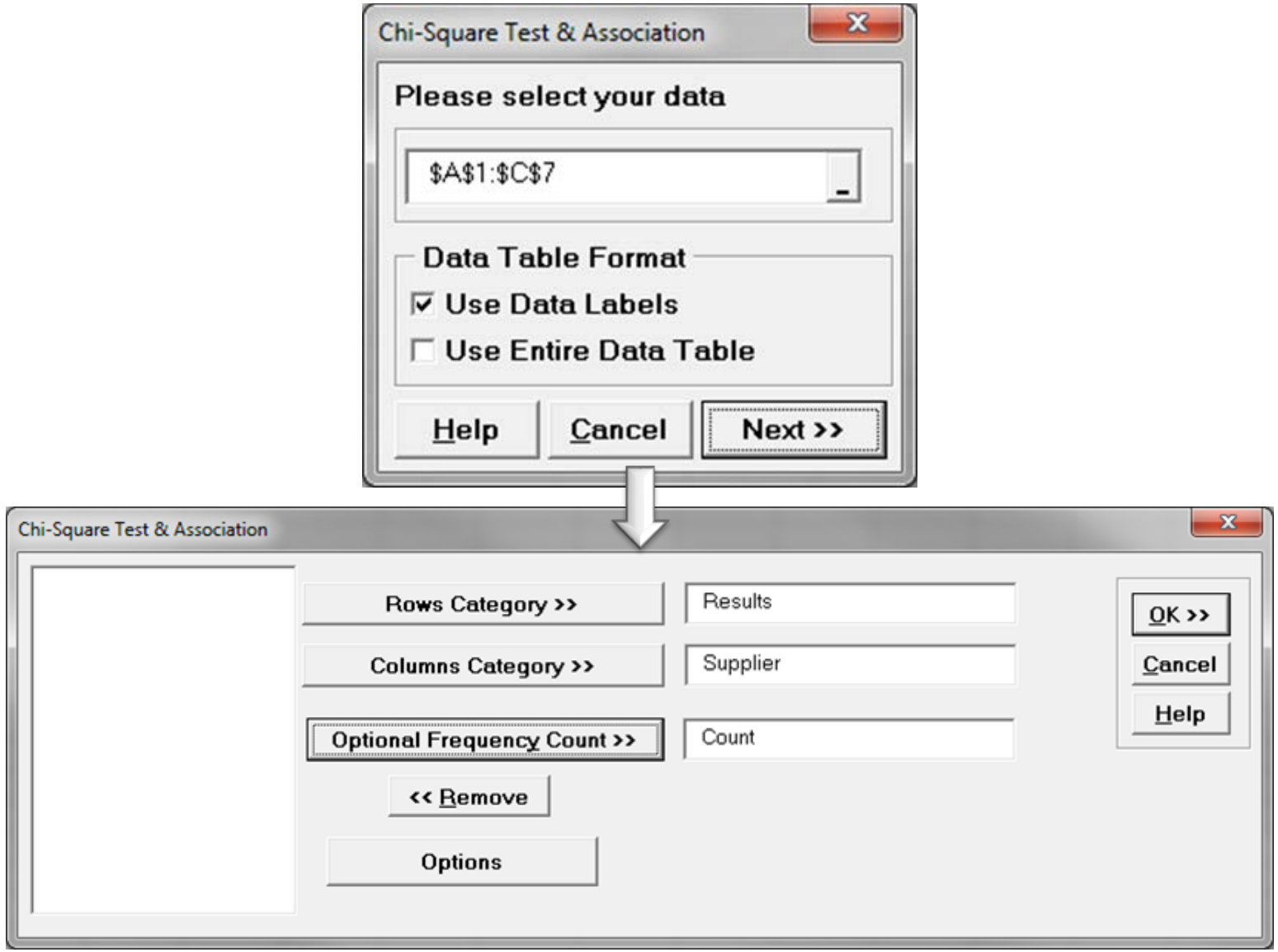
Use SigmaXL to Run a Chi-Square Test

- Case Study 1: We are interested in comparing the product quality exam pass rates of three suppliers A, B and C using nonparametric (i.e. distribution-free) hypothesis test: Chi-square test.
  - Data File: “Chi-Square Test_1” tab in “Sample Data.xlsx”
- Null Hypothesis (H_0): $p_A = p_B = p_C$
- Alternative Hypothesis (H_a): at least one of the suppliers has different pass rates from the others.

Use SigmaXL to Run a Chi-Square Test

- Steps to run a chi-square test in SigmaXL
  - Select the entire range of data (Supplier, Results, & Count)
  - Click SigmaXL -> Statistical Tools ->
    Chi-Square Tests -> Chi-Square Test
  - A new window named “Chi-Square Test & Association” pops up with the selected range appearing in the box under “Please select your data”.
  - Click “Next>>”
  - A new window named “Chi-Square Test & Association” pops up.
  - Select “Results” as “Count Category (X1)”
  - Select “Supplier” as “Group Category (X2)”
  - Select “Count” as “Count (Y)”
  - Click “OK>>”
  - The Chi-square test results appears automatically in the new tab “Chi-Square (1)”

Use SigmaXL to Run a Chi-Square Test



Use SigmaXL to Run a Chi-Square Test

Chi-Square Test

Results - Supplier - Count

Observed Counts	Supplier A	Supplier B	Supplier C
Fail	20	30	10
Pass	160	140	150

Expected Counts	Supplier A	Supplier B	Supplier C
Fail	21.176	20	18.824
Pass	158.82	150	141.18

Std. Residuals	Supplier A	Supplier B	Supplier C
Fail	-0.255655	2.236	-2.034
Pass	0.093352006	-0.816497	0.742611

Chi-Square	10.428
DF	2
P-Value	0.0054

- Observed counts are based on the sample observation.
- Expected counts are based on the assumption that the null hypothesis is true.
- Since the p-value is smaller than alpha level (0.05), we reject the null and claim that at least one supplier has a different pass rate from the others.

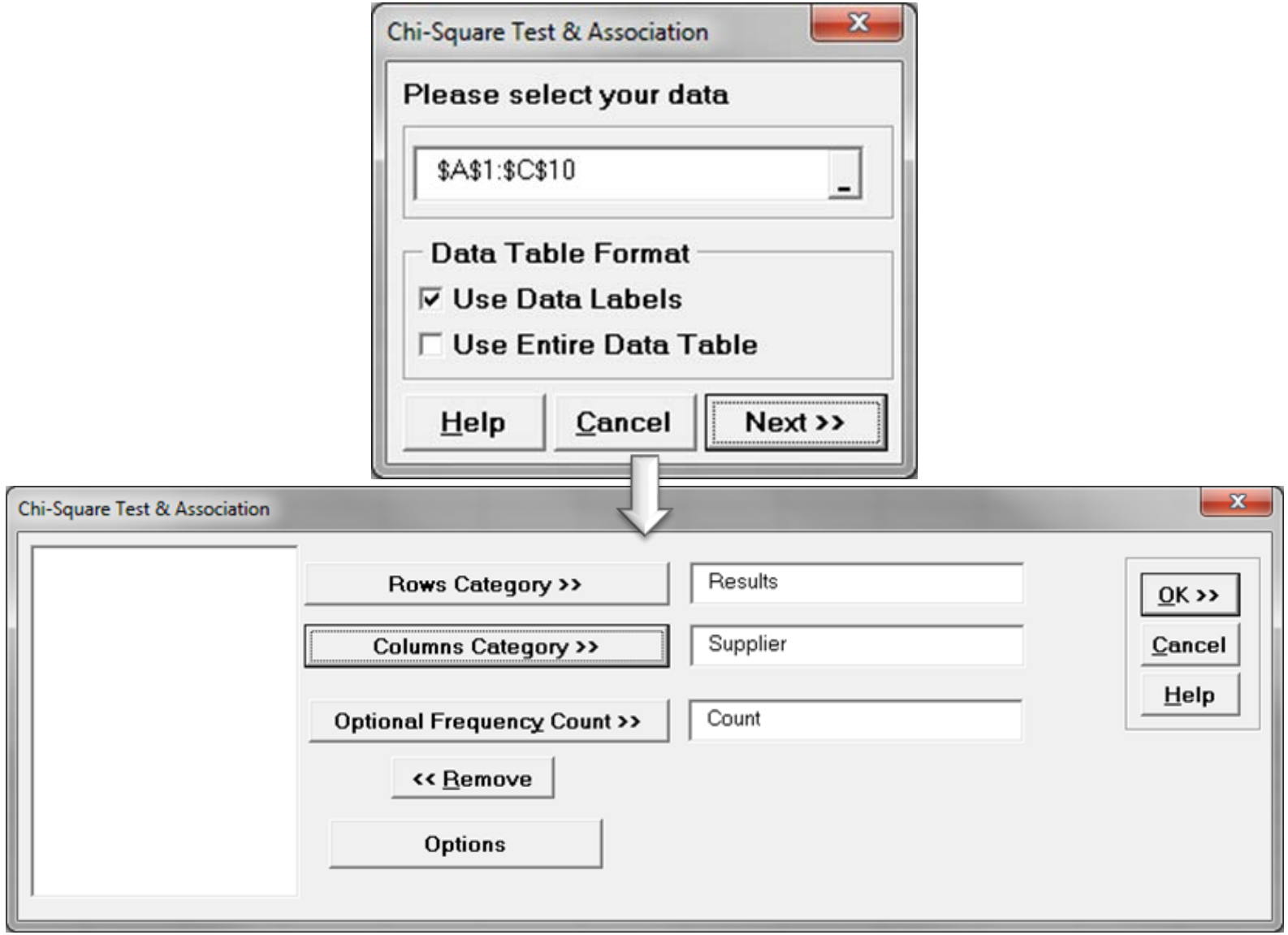
Use SigmaXL to Run a Chi-Square Test

- Case Study 2: We are trying to check whether there is a relationship between the suppliers and the results of product quality exam using nonparametric (i.e. distribution-fee) hypothesis test: Chi-square test.
  - Data File: “Chi-Square Test2” tab in “Sample Data.xlsx”
- Null Hypothesis (H_0): product quality exam results are independent of the suppliers.
- Alternative Hypothesis (H_a): product quality exam results depend on the suppliers.

Use SigmaXL to Run a Chi-Square Test

- Steps to run a chi-square test in SigmaXL
  - Select the entire range of data (Supplier, Results, & Count)
  - Click SigmaXL -> Statistical Tools ->
    Chi-Square Tests -> Chi-Square Test & Association
  - A new window named “Chi-Square Test & Association” pops up with the selected range appearing in the box under “Please select your data”.
  - Click “Next>>”
  - A new window named “Chi-Square Test & Association” pops up.
  - Select “Results” as “Count Category (X1)”
  - Select “Supplier” as “Group Category (X2)”
  - Select “Count” as “Count (Y)”
  - Click “OK>>”
  - The Chi-square test results appears automatically in the new tab “Chi-Square (1)”

Use SigmaXL to Run a Chi-Square Test



Use SigmaXL to Run a Chi-Square Test

Chi-Square Test

Results - Supplier - Count

Observed Counts	Supplier A	Supplier B	Supplier C
Fail	20	30	10
Marginal	20	30	40
Pass	160	140	150

Expected Counts	Supplier A	Supplier B	Supplier C
Fail	20	20	20
Marginal	30	30	30
Pass	150	150	150

Std. Residuals	Supplier A	Supplier B	Supplier C
Fail	0	2.236	-2.236
Marginal	-1.826	0	1.826
Pass	0.816497	-0.816497	0

Chi-Square	18
DF	4
P-Value	0.0012

The p-value is smaller than the alpha level (0.05).

Therefore, we reject the null hypothesis.

The product quality exam results are not independent of the suppliers.

3.5.9 Tests of Equal Variance

What are Tests of Equal Variance?

- **Tests of equal variance** are a family of hypothesis tests used to check whether there is a statistically significant difference between the variances of two or more populations.
  - Null Hypothesis (H_0): $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$
  - Alternative Hypothesis (H_a): at least the variance of one population is different from others.
  - k is the number of populations of interest. k ≥ 2.
- A tests of equal variance can be used alone but most of the time it is used with other statistical methods to verify or support the assumption about the variance equality.

F-Test

- The **F-test** is used to compare the variances between two normally distributed populations.

- It is extremely sensitive to non-normality and serves as a preliminary step for two sample t-test.

- Test Statistic:

$$F_{calc} = \frac{s_1^2}{s_2^2}$$, where s_1 and s_2 are the sample standard deviations.

The sampling distribution of the test statistic follows F distribution when the null is true.

Bartlett's Test

- **Bartlett's test** is used to compare the variances among two or more normally distributed populations.
- It is sensitive to non-normality and it serves as a preliminary step for ANOVA.

- Test Statistic:

$$\chi^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \frac{1}{N-k}\right)}$$, where $N = \sum_{i=1}^{k} n_i$ and $S_p^2 = \frac{1}{N-k}\sum_i (n_i - 1)S_i^2$

The sampling distribution of test statistic follows chi^2 distribution when the null is true.

Brown-Forsythe Test

- The **Brown-Forsythe test** is used to compare the variances between two or more populations with any distributions.

- It is not so sensitive to non-normality as Bartlett's test.

- The Test statistic is the model F statistic from the ANOVA on the transformed response $z_{ij} = \left| y_{ij} - \tilde{y}_i \right|$ where $\tilde{y}_i$ is the median response at i^{th} level.

Levene's Test

- **Levene's test** is used to compare the variances between two or more populations with any distributions.

- It is not so sensitive to non-normality as Bartlett's test.

- The test statistic is the model F statistic from the ANOVA on the transformed response $z_{ij} = \left| y_{ij} - \bar{y}_i \right|$ where $\bar{y}_i$ is the mean response at ith level.

Brown-Forsythe Test vs. Levene's Test

$$F = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2}$$

where

N is the total number of observations.
k is the number of groups.
N_i is the number of observations in the ith group.
$Z_{i.}$ is the group mean of the ith group.
$Z_{..}$ is the grand mean of all the observations.

In Brown-Forsythe Test, $Z_{ij} = \left| Y_{ij} - \tilde{Y}_{ij} \right|$, where $\tilde{Y}_{ij}$ is the group median of the ith group.

In Levene's Test, $Z_{ij} = \left| Y_{ij} - \bar{Y}_{ij} \right|$, where $\bar{Y}_{ij}$ is the group mean of the ith group.

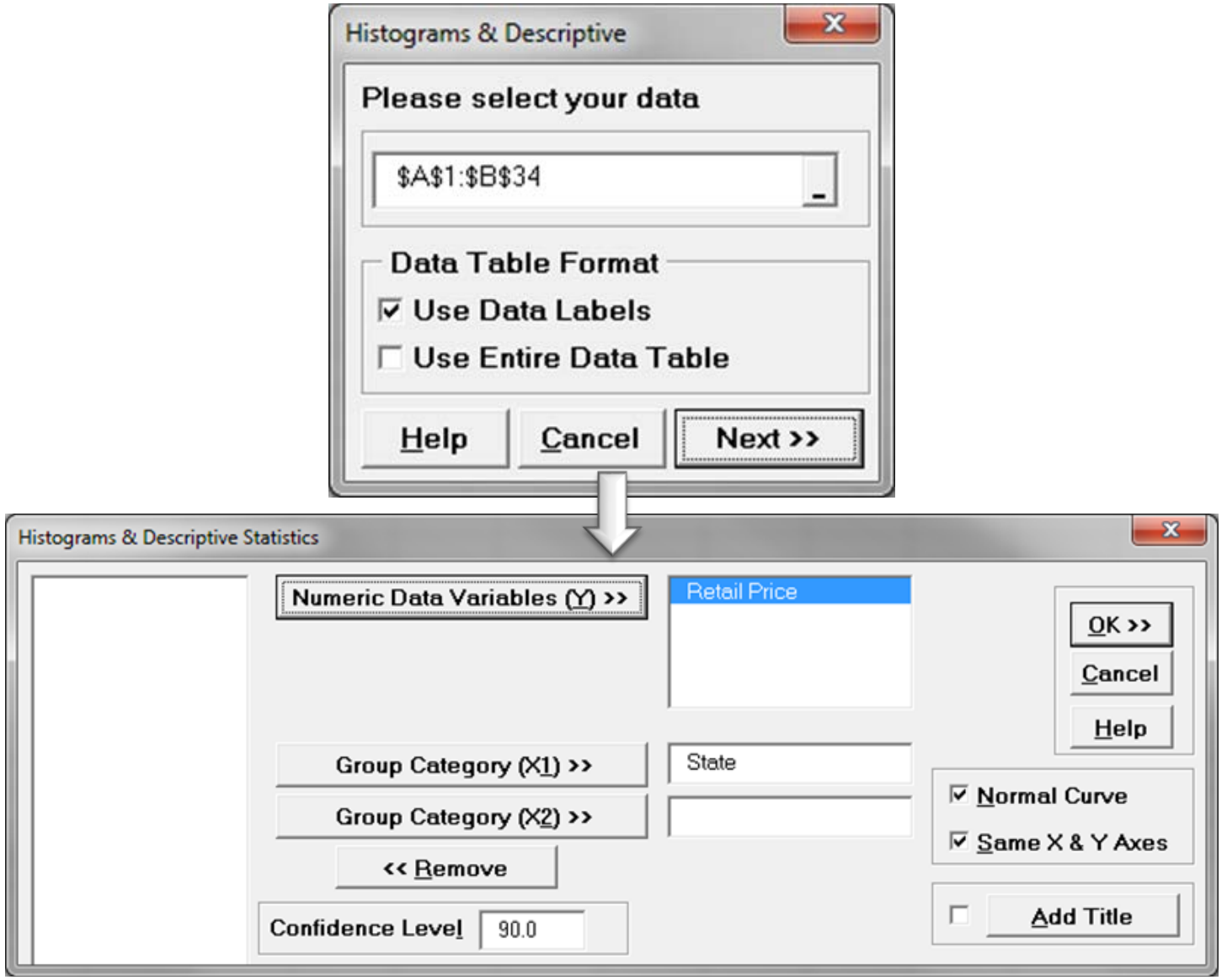
Use SigmaXL to Run Tests of Equal Variance

- Case Study: We are interested in comparing the variances of the retail price of a product in state A and state B.
  - Data File: “Two-Sample T-Test” tab in “Sample Data.xlsx”
- Null Hypothesis (H_0): $\sigma_A^2 = \sigma_B^2$
- Alternative Hypothesis (H_a): $\sigma_A^2 \neq \sigma_B^2$

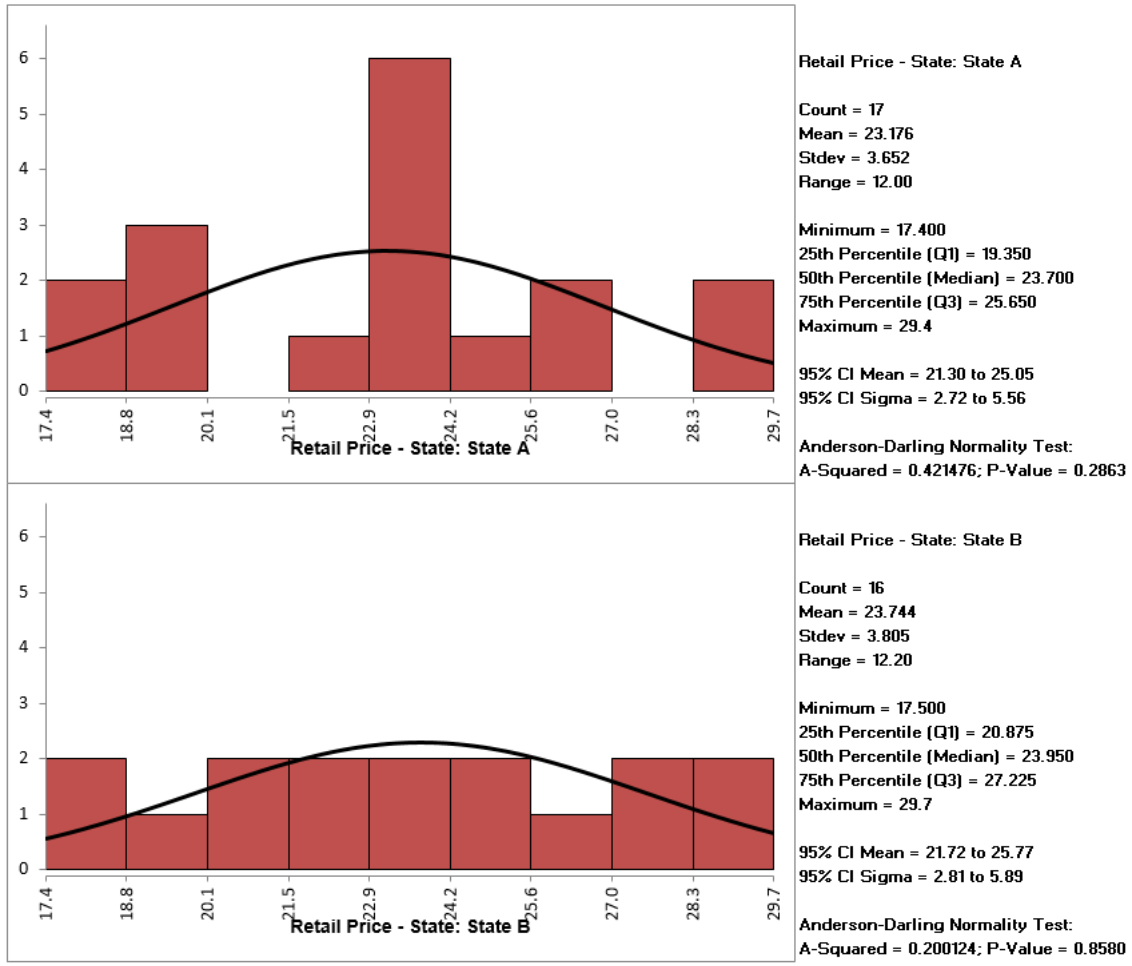
Use SigmaXL to Run Tests of Equal Variance

- Step 1: run the normality test to check whether all levels of data are normally distributed
  - Select the entire range of data (both “State” and “Retail Price” columns)
  - Click SigmaXL -> Graphical Tools -> Histograms & Descriptive Statistics
  - A new window named “Histograms & Descriptive ” pops up with the selected range appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Histograms & Descriptive Statistics” appears
  - Select “Retail Price” as the “Numeric Data Variables (Y)”
  - Select “State” as “Group Category (X1)”
  - Click “OK>>”
  - The normality test results appear automatically in the new tab “Hist Descript (1)”

Use SigmaXL to Run Tests of Equal Variance



Use SigmaXL to Run Tests of Equal Variance



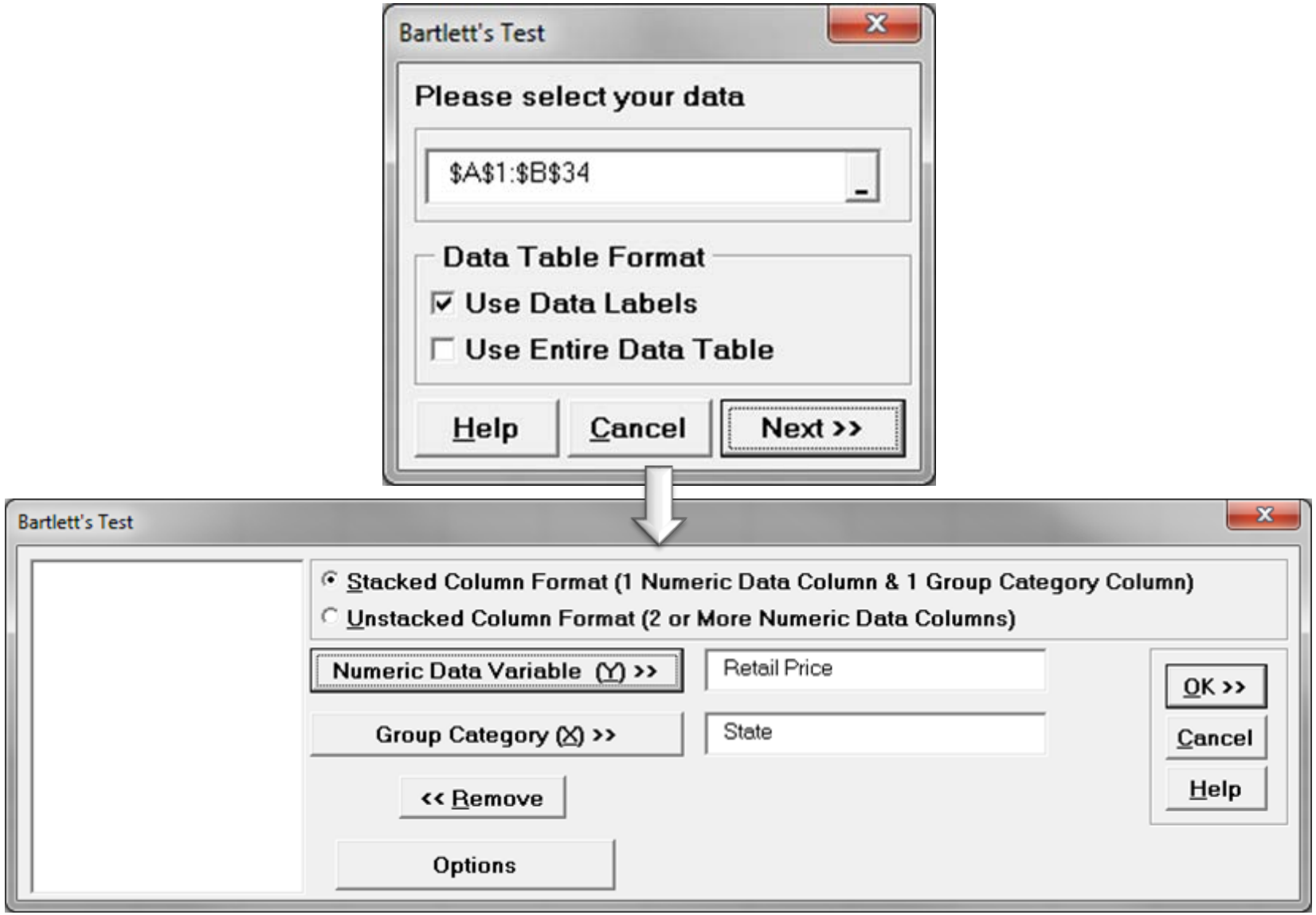
Null Hypothesis (H_0): The data are normally distributed.
Alternative Hypothesis (H_a): The data are not normally distributed.

Both retail price data of state A and B are normally distributed since the p-values are both greater than alpha level (0.05).

Use SigmaXL to Run Tests of Equal Variance

- Step 2: run tests of equal variance in SigmaXL
  - If all the groups of data are normally distributed, run Bartlett's test in SigmaXL to test the equality of the variance.
    - Select the entire range of data (both "State" and "Retail Price" columns)
    - Click SigmaXL -> Statistical Tools -> Equal Variance Tests -> Bartlett
    - A new window named "Bartlett's Test" pops up with the selected range appearing in the box under "Please select your data"
    - Click "Next>>"
    - A new window named "Bartlett's Test" appears
    - Select "Retail Price" as the "Numeric Data Variable (Y)"
    - Select "State" as "Group Category (X)"
    - Click "OK>>"
    - The results of Bartlett's test appear automatically in the new tab "Bartlett's Test (1)"

Use SigmaXL to Run Tests of Equal Variance



Use SigmaXL to Run Tests of Equal Variance

Bartlett's Test For Equal Variance: Retail Price
(Use with normal data)

Test Information

H_0: Variance 1 = Variance 2 = ... = Variance k

Ha: At least one pair Variance i ≠ Variance j

State	State A	State B
Count	17	16
Mean	23.176	23.744
Median	23.700	23.950
StDev	3.652	3.805
AD Normality Test P-Value	0.2863	0.8580
Bartlett's Test Statistic	0.025027978	
P-Value	0.8743	

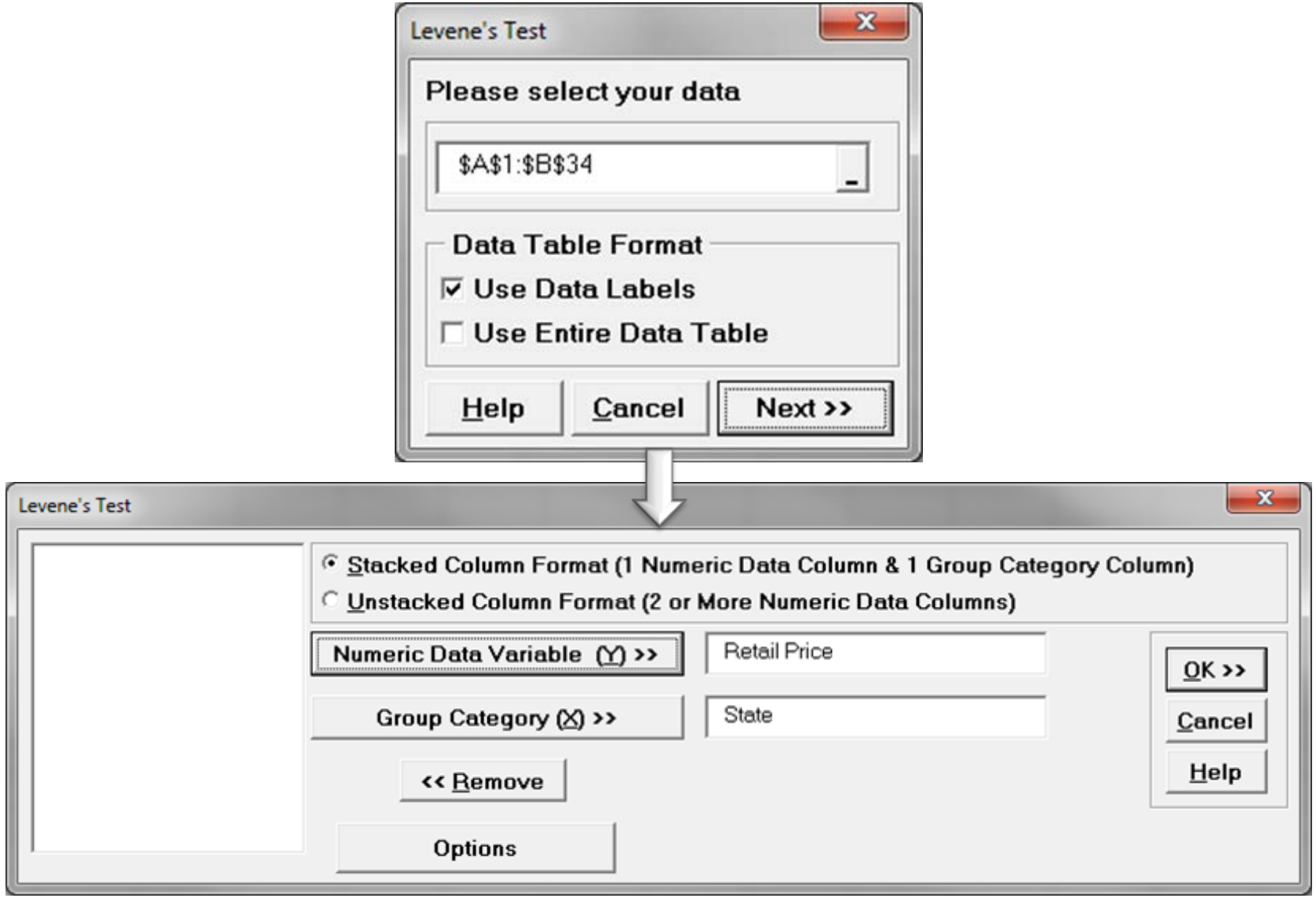
The p-value of Bartlett's test is greater than the alpha level of 0.05.

Therefore, we fail to reject the null hypothesis and claim that the variances of different groups are identical.

Use SigmaXL to Run Tests of Equal Variance

- If at least one of the groups is not normally distributed, run Levene's test in SigmaXL to test the equality of the variance.
  - Select the entire range of data (both "State" and "Retail Price" columns)
  - Click SigmaXL -> Statistical Tools -> Equal Variance Tests -> Levene
  - A new window named "Levene's Test" pops up with the selected range appearing in the box under "Please select your data"
  - Click "Next>>"
  - A new window named "Levene's Test" appears
  - Select "Retail Price" as the "Numeric Data Variable (Y)"
  - Select "State" as "Group Category (X)"
  - Click "OK>>"
  - The results of Levene's test appear automatically in the new tab "Levene's Test (1)"

Use SigmaXL to Run Tests of Equal Variance



Use SigmaXL to Run Tests of Equal Variance

Levene's Test For Equal Variance: Retail Price
(Use with non-normal data)

Test Information

H_0: Variance 1 = Variance 2 = ... = Variance k

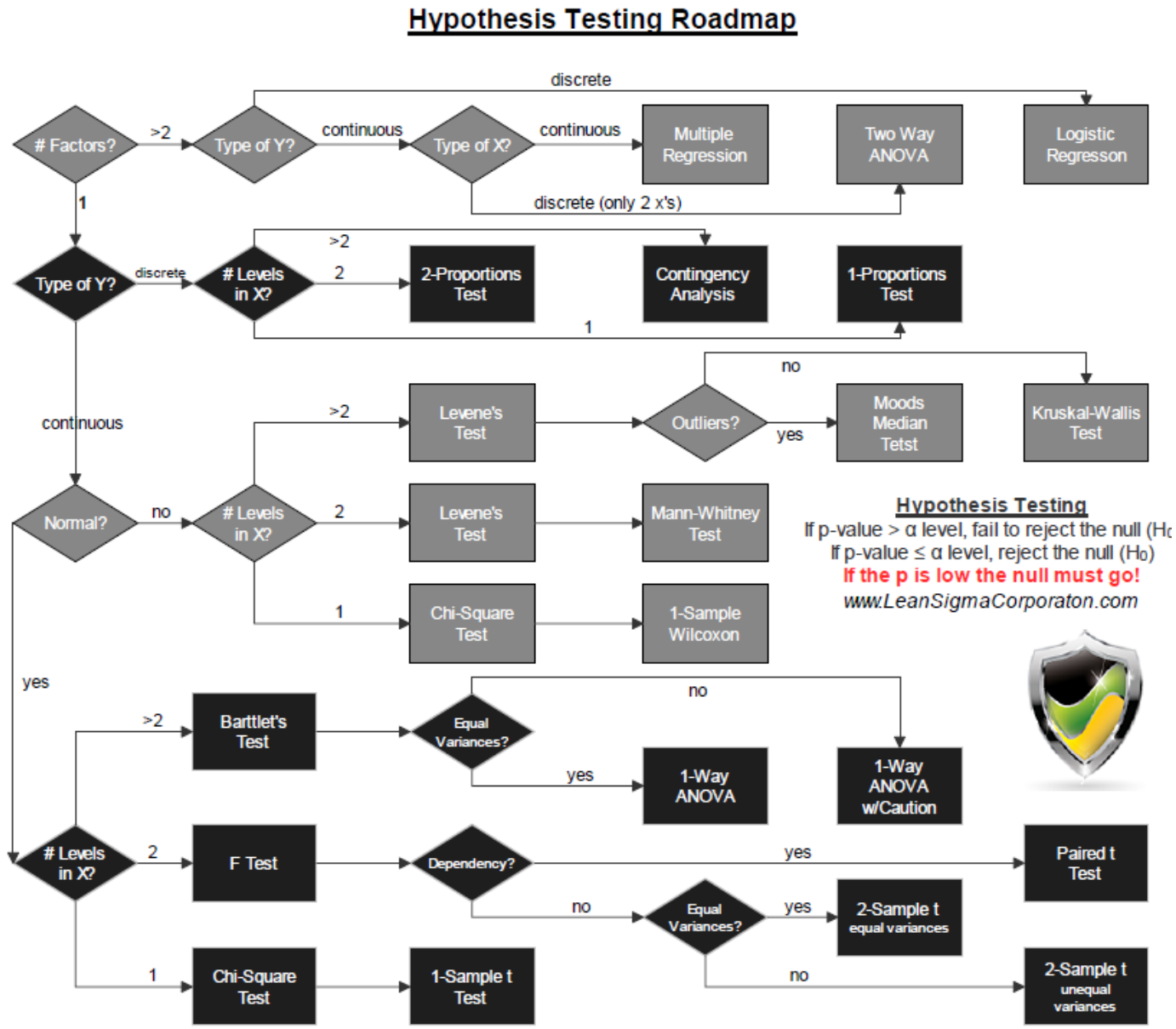
Ha: At least one pair Variance i ≠ Variance j

State	State A	State B
Count	17	16
Mean	23.176	23.744
Median	23.700	23.950
StDev	3.652	3.805
AD Normality Test P-Value	0.2863	0.8580
Levene's Test Statistic	0.162134	
DF Num	1	
DF Den	31	
P-Value	0.6900	

The p-value of Levene's test is greater than the alpha level of 0.05.

Therefore, we fail to reject the null hypothesis and claim that the variances of different groups are identical.

Hypothesis Testing Roadmap: Putting it all together



4.0 Improve Phase

Black Belt Training: Improve Phase

4.1 Simple Linear Regression
- 4.1.1 Correlation
- 4.1.2 X-Y Diagram
- 4.1.3 Regression Equations
- 4.1.4 Residuals Analysis

4.2 Multiple Regression Analysis
- 4.2.1 Non-Linear Regression
- 4.2.2 Multiple Linear Regression
- 4.2.3 Confidence Intervals
- 4.2.4 Residuals Analysis
- 4.2.5 Data Transformation, Box Cox
- 4.2.6 Stepwise Regression
- 4.2.7 Logistic Regression

4.3 Designed Experiments
- 4.3.1 Experiment Objectives
- 4.3.2 Experimental Methods
- 4.3.3 DOE Design Considerations

4.4 Full Factorial Experiments
- 4.4.1 2k Full Factorial Designs
- 4.4.2 Linear and Quadratic Models
- 4.4.3 Balanced and Orthogonal Designs
- 4.4.4 Fit, Model, and Center Points

4.5 Fractional Factorial Experiments
- 4.5.1 Designs
- 4.5.2 Confounding Effects
- 4.5.3 Experimental Resolution

4.1 Simple Linear Regression

Black Belt Training: Improve Phase

4.1 Simple Linear Regression
- 4.1.1 Correlation
- 4.1.2 X-Y Diagram
- 4.1.3 Regression Equations
- 4.1.4 Residuals Analysis

4.2 Multiple Regression Analysis
- 4.2.1 Non-Linear Regression
- 4.2.2 Multiple Linear Regression
- 4.2.3 Confidence Intervals
- 4.2.4 Residuals Analysis
- 4.2.5 Data Transformation, Box Cox
- 4.2.6 Stepwise Regression
- 4.2.7 Logistic Regression

4.3 Designed Experiments
- 4.3.1 Experiment Objectives
- 4.3.2 Experimental Methods
- 4.3.3 DOE Design Considerations

4.4 Full Factorial Experiments
- 4.4.1 2k Full Factorial Designs
- 4.4.2 Linear and Quadratic Models
- 4.4.3 Balanced and Orthogonal Designs
- 4.4.4 Fit, Model, and Center Points

4.5 Fractional Factorial Experiments
- 4.5.1 Designs
- 4.5.2 Confounding Effects
- 4.5.3 Experimental Resolution

4.1.1 Correlation

What is Correlation?

- **Correlation** is a statistical technique that describes whether and how strongly two or more variables are related.

- **Correlation analysis** helps to understand the direction and degree of association between variables, and it suggests whether one variable can be used to predict another.

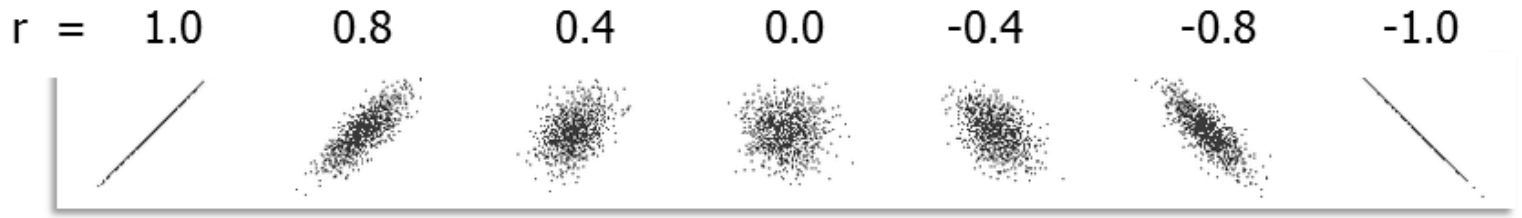
- Of the different metrics to measure correlation, Pearson's correlation coefficient is the most popular. It measures the linear relationship between two variables.

Pearson's Correlation Coefficient

- **Pearson's correlation coefficient** is also called:
  - Pearson's r or coefficient of correlation
  - Pearson's product moment correlation coefficient (r)
- "r" is a statistic measuring the linear relationship between two variables.
- Correlation coefficients range from -1 to 1.
  - If r = 0, there is no linear relationship between the variables.
  - The *sign* of r indicates the *direction* of the relationship:
    - If r < 0, there is a negative linear correlation.
    - If r > 0, there is a positive linear correlation.
  - The *absolute value* of r describes the *strength* of the relationship:
    - If $|r| \leq 0.5$, there is a weak linear correlation.
    - If $|r| > 0.5$, there is a strong linear correlation.
    - If $|r| = 1$, there is a perfect linear correlation.

Pearson's Correlation Coefficient

- When the correlation is *strong*, the data points on a scatter plot will be close together (tight).
  - The closer "r" is to -1 or 1, the stronger the relationship.
    - -1 Strong inverse relationship
    - +1 Strong direct relationship
- When the correlation is *weak*, the data points are spread apart more (loose).
  - The closer the correlation is to 0 the weaker the relationship.



Pearson's Correlation Coefficient

- Pearson's correlation coefficient is only sensitive to the *linear* dependence between two variables.

- It is possible that two variables have a perfect non-linear relationship when the correlation coefficient is low.

- Notice the scatter plots below with correlation equal to 0. There are clearly *relationships* but they are not linear and therefore can not be determined with Pearson's correlation coefficient.

Correlation and Causation

- Correlation *does not* imply causation.

- If variable A is highly correlated with variable B, it does not necessarily mean A causes B or vice versa. It is possible that an unknown third variable C is causing both A and B to change.

- For example, if ice cream sales at the beach are highly correlated with the number of shark attacks, it does not imply that increased ice cream sales causes increased shark attacks. They are triggered by a third factor: summer.

Correlation and Dependence

- If two variables are independent, the correlation coefficient is zero.
- WARNING! If the correlation coefficient of two variables is zero, it does not imply they are independent.
- The correlation coefficient only indicates the linear dependence between two variables. When variables are non-linearly related, they are not independent of each other but their correlation coefficient could be zero.

Correlation Coefficient and X-Y Diagram

- The correlation coefficient indicates the direction and strength of the linear dependence between two variables but it does not cover all the existing relationship patterns.
- With the same correlation coefficient, two variables might have completely different dependence patterns.
- A scatter plot or X-Y diagram can help to discover and understand additional characteristics of the relationship between variables.
- Correlation coefficient is not a replacement for examining the scatter plot to study the variables' relationship.

Statistical Significance of the Correlation Coefficient

- The correlation coefficient could be high or low by chance (randomness). It may have been calculated based on two small samples that do not provide good inference on the correlation between two populations.

- In order to test whether there is a statistically significant relationship between two variables, we need to run a hypothesis test to determine whether the correlation coefficient is statistically different from zero.
  - Hypothesis Test Statements
    - H_0: r = 0: Null Hypothesis: There is *no* correlation.
    - H_1: r ≠ 0: Alternate Hypothesis: There is a correlation.

Statistical Significance of the Correlation Coefficient

- Hypothesis tests will produce p-values as a result of the statistical significance test on r.
  - When the p-value for a test is low (less than 0.05), we can reject the null hypothesis and conclude that “r” is significant; there is a correlation.
  - When the p-value for a test is > 0.05, then we fail to reject the null hypothesis; there is no correlation.
- We can also use the t statistic to draw the same conclusions regarding our test for significance of the correlation coefficient.

Statistical Significance of the Correlation Coefficient

- Test Statistic: $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

- Critical Statistic: t-value in t-table with (n – 2) degrees of freedom

- If $|t| \leq t_{critical}$, we fail to reject the null. There is no statistically significant linear relationship between X and Y.

- If $|t| > t_{critical}$, we reject the null. There is a statistically significant linear relationship between X and Y.

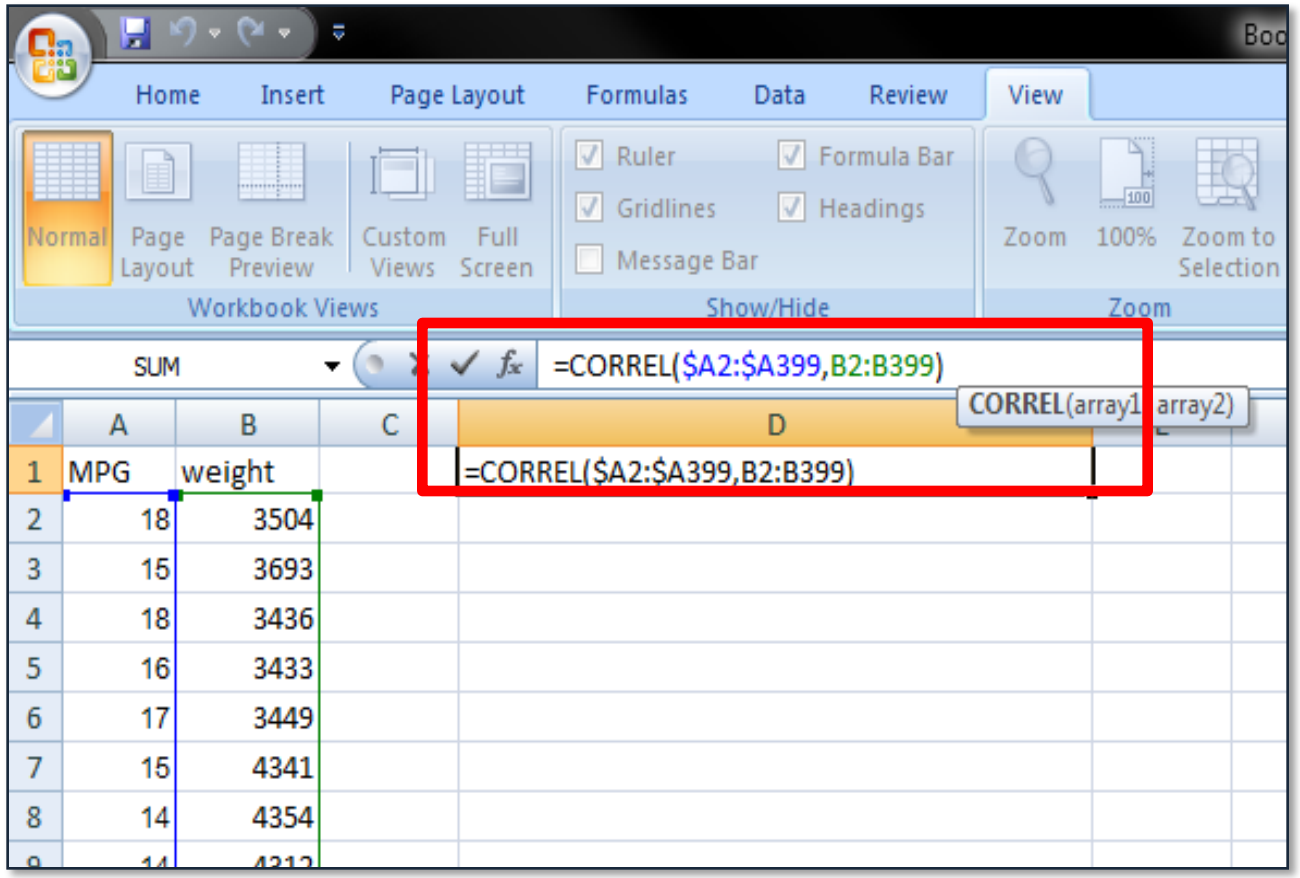
Using Software to Calculate the Correlation Coefficient

- We are interested in understanding whether there is linear dependence between a car's MPG and its weight and if so, how they are related.

- The MPG and weight data are stored in the "Correlation Coefficient" tab in "Sample Data.xlsx." We will discuss three ways to get the results.

Use Excel to Calculate the Correlation Coefficient

- The formula CORREL in Excel calculates the sample correlation coefficient of two data series.
- The correlation coefficient between the two data series is -0.83, which indicates a strong negative linear relationship between MPG and weight.



Interpreting Results

- How do we interpret results and make decisions based Pearson's correlation coefficient (r) and p-values?
  - Let us look at a few examples:
    - r = -0.832, p = 0.000 (previous example). The two variables are inversely related and the linear relationship is strong. Also, this conclusion is significant as supported by p-value of 0.00.
    - r = -0.832, p = 0.71. Based on r, you should conclude the linear relationship between the two variables is strong and inversely related. However, with a p-value of 0.71, you should then conclude that r is not significant and that your sample size may be too small to accurately characterize the relationship.
    - r = 0.5, p = 0.00. Moderately positive linear relationship, r is statistically significant.
    - r = 0.92, p = 0.61. Strong positive linear relationship but r is not statistically significant. Get more data.
    - r = 1.0, p = 0.00. The two variables have a perfect linear relationship and r is significant.

Correlation Coefficient Calculation

- Population Correlation Coefficient (ρ)

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

- Sample Correlation Coefficient (r)

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- It is only defined when the standard deviations of both X and Y are non-zero and finite.
- When covariance of X and Y is zero, the correlation coefficient is zero.

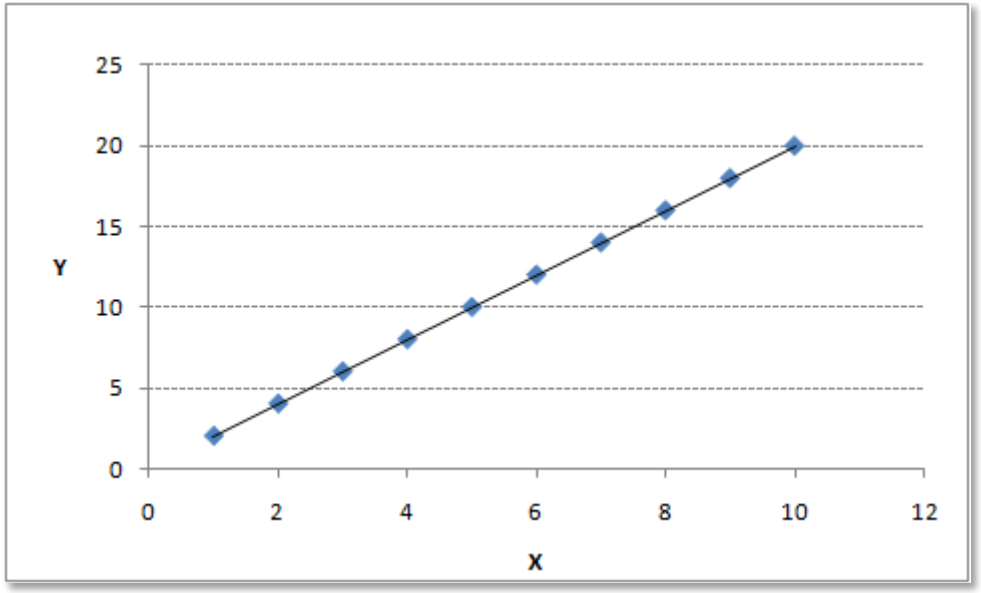
4.1.2 X-Y Diagram

What is an X-Y Diagram?

- An **X-Y diagram** is a scatter plot depicting the relationship between two variables (i.e., X and Y).
- Each point on the X-Y diagram represents a pair of X and Y values, with X plotted on the horizontal axis and Y plotted on the vertical axis.
- With an X-Y diagram, you can qualitatively assess both the strength and direction of the relationship between X and Y.
- To quantitatively measure the relationship between X and Y, you may need to calculate the correlation coefficient.

Example 1: Perfect Linear Correlation

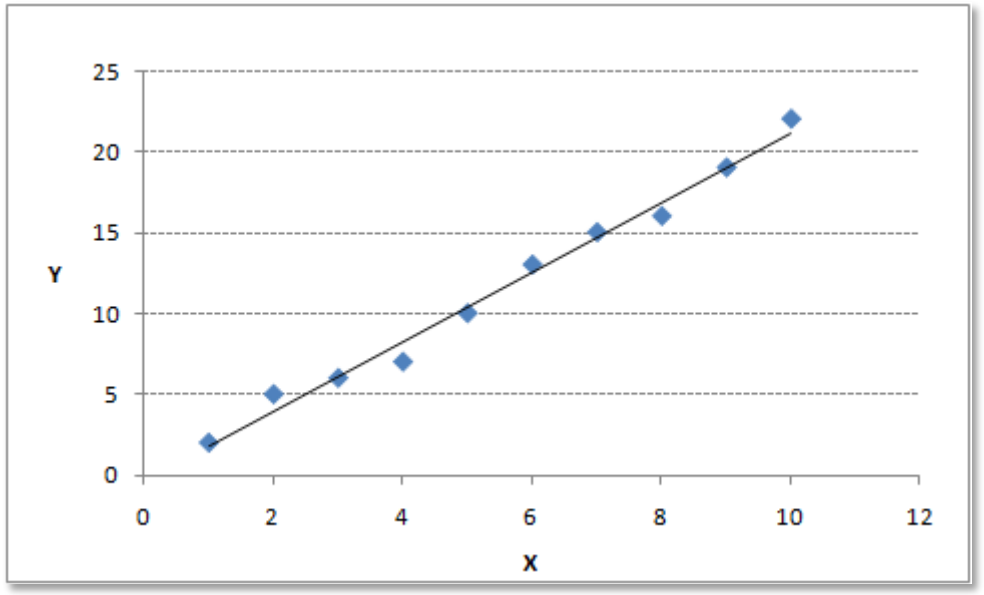
- In the chart below, there are 10 data points depicted (10 pairs of X and Y values), and they were created using the equation Y = 2X.
- As a result, all the data points fall onto the straight line of Y = 2X.
- The chart demonstrates a perfect positive linear correlation between X and Y since the relationship between X and Y can be perfectly described by a linear equation in a format of Y = a × X + b where a ≠ 0.



X	Y
1	2
2	4
3	6
4	8
5	10
6	12
7	14
8	16
9	18
10	20

Example 2: Strong Linear Correlation

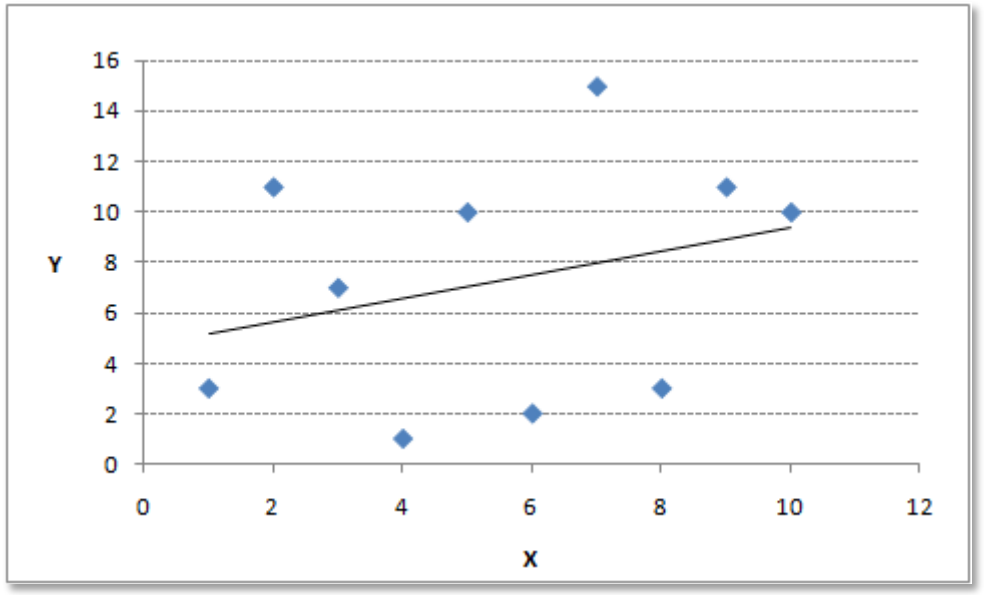
- In this chart, the data points scatter closely around a straight line.
- When X increases, Y increases accordingly.
- This chart demonstrates a strong positive linear correlation between X and Y.
- The straight line is the trend line showing how Y's trend goes with changes in X.



X	Y
1	2
2	5
3	6
4	7
5	10
6	13
7	15
8	16
9	19
10	22

Example 3: Weak Linear Correlation

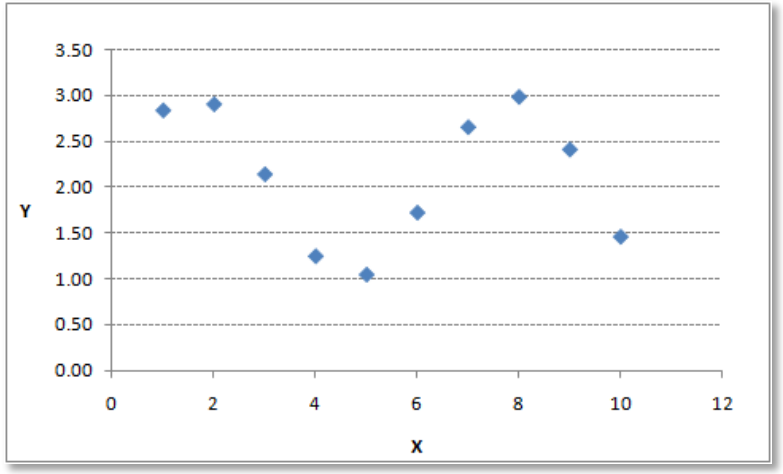
- In this chart, the data points scatter remotely around a straight line.
- When X increases, Y increases accordingly.
- This chart demonstrates a weak positive linear correlation between X and Y since the distance between the data points and the trend line is relatively far on average



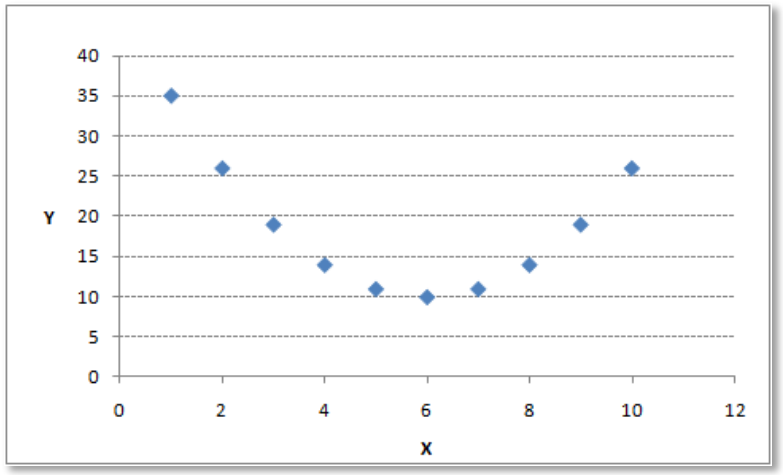
X	Y
1	3
2	11
3	7
4	1
5	10
6	2
7	15
8	3
9	11
10	10

Example 4: Non-Linear Correlation

- The X-Y diagram also helps to identify any nonlinear relationship between X and Y.



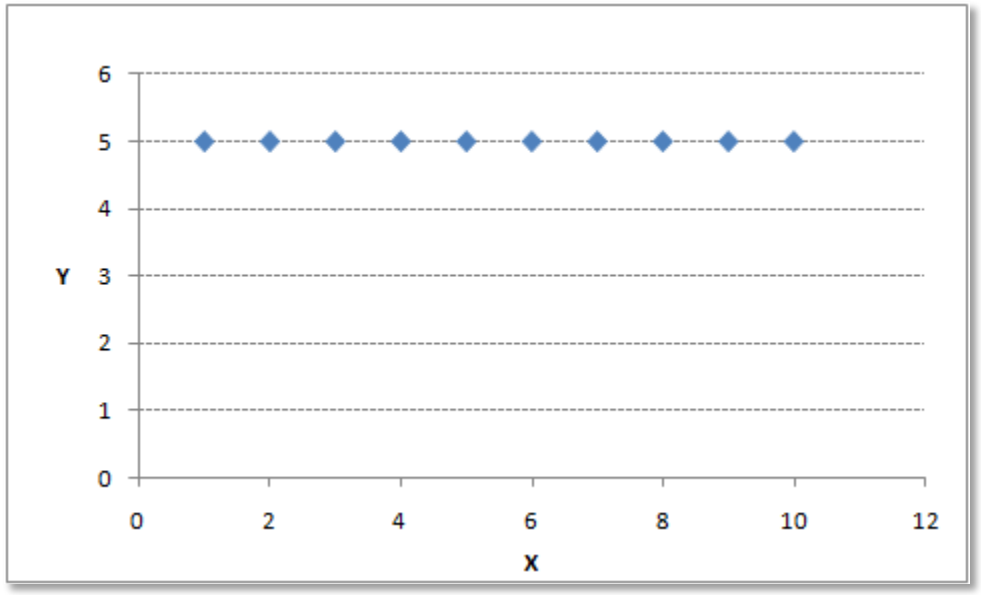
X	Y
1	2.84
2	2.91
3	2.14
4	1.24
5	1.04
6	1.72
7	2.66
8	2.99
9	2.41
10	1.46



X	Y
1	35
2	26
3	19
4	14
5	11
6	10
7	11
8	14
9	19
10	26

Example 5: Uncorrelated

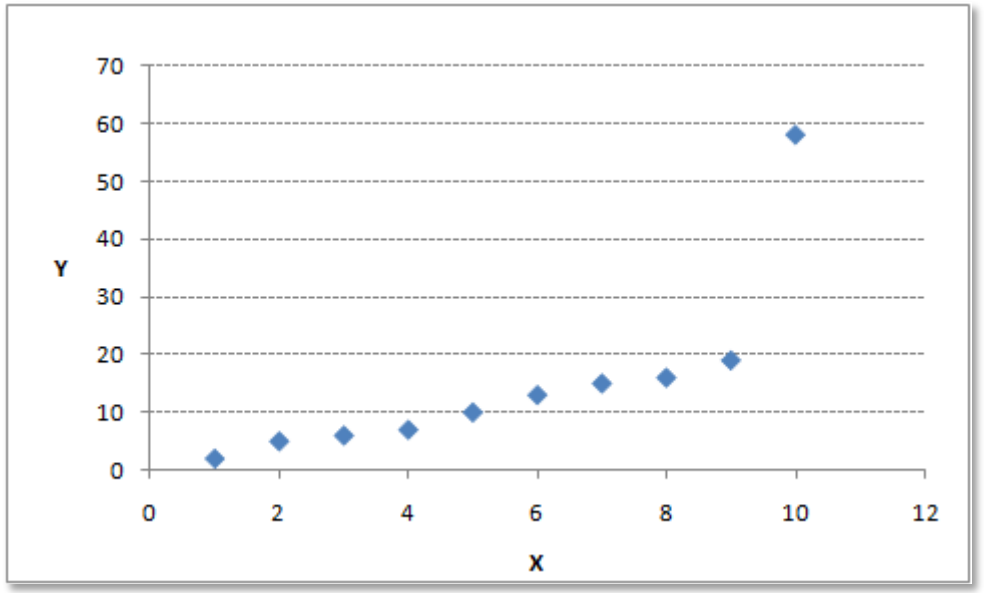
- In this chart, the Y value of each data point is a constant regardless of what the X value is.
- Changes in X do not show any relative impact on Y. As a result, there is no correlation between X and Y.



X	Y
1	5
2	5
3	5
4	5
5	5
6	5
7	5
8	5
9	5
10	5

Example 6: Outlier Identification

- Using X-Y diagram, you may identify outliers in the data.
- In this chart, the last data point does not seem to follow the trend of other data points.
- This should require further investigation on the last data point to determine whether it is an outlier.



X	Y
1	2
2	5
3	6
4	7
5	10
6	13
7	15
8	16
9	19
10	58

Benefits of Using an X-Y Diagram

- An X-Y diagram graphically demonstrates the relationship between two variables.
- It suggests whether two variables are associated and helps to identify the linear or nonlinear correlation between X and Y.
- It captures the strength and direction of the relationship between X and Y.
- It helps identify any outliers in the data.

Limitations of the X-Y Diagram

- Although the X-Y diagram helps to “spot” interesting features in the data, it does not provide any quantitative conclusions about the data and further statistical analysis is needed to:
  - Assess whether the association between variables is statistically significant.
  - Measure the strength of the relationship between variables.
  - Determine whether outliers exist in the data.
  - Quantitatively describe the pattern of the data.

4.1.3 Regression Equations

Correlation and Regression Analysis

- The correlation coefficient answers the following questions:
  - Are two variables correlated?
  - How strong is the relationship between two variables?
  - When one variable increases, does the other variable increase or decrease?
- The correlation coefficient *cannot* address the following questions:
  - How much does one variable changes when the other variable changes by one unit?
  - How can we set the value of one variable to obtain a targeted value of the other variable?
  - How can we use the relationship between two variables to make predictions?
- The simple linear regression analysis helps to answer these questions.

What is Simple Linear Regression?

- **Simple linear regression** is a statistical technique to fit a straight line through the data points.
- It models the quantitative relationship between two variables.
- It describes how one variable changes according to the change of another variable.
- Both variables need to be continuous.
- It is simple because only one predictor variable is involved.

Simple Linear Regression Equation

- The simple linear regression analysis fits the data to a regression equation in the form

$$Y = \alpha \times X + \beta + e$$

where:

- Y is the dependent variable (the response) and X is the single independent variable (the predictor).
- α is the slope describing the steepness of the fitting line. β is the intercept indicating the Y value when X is equal to 0.
- e stands for error (residual). It is the difference between the actual Y and the fitted Y (i.e., the vertical difference between the data point and the fitting line).

Ordinary Least Squares

- The **ordinary least square** is a statistical method used in linear regression analysis to find the best fitting line for the data points.

- It estimates the unknown parameters of the regression equation by minimizing the sum of squared residuals (i.e., the vertical difference between the data point and the fitting line).

- In mathematical language, we look for α and β that satisfy the following criteria:

$$\min_{\alpha,\beta} Q(\alpha,\beta) \text{ where } Q(\alpha,\beta) = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Ordinary Least Squares

- The actual value of the dependent variable:

$$Y_i = \alpha * X_i + \beta + e_i \quad \text{where } i = 1, 2, \ldots, n$$

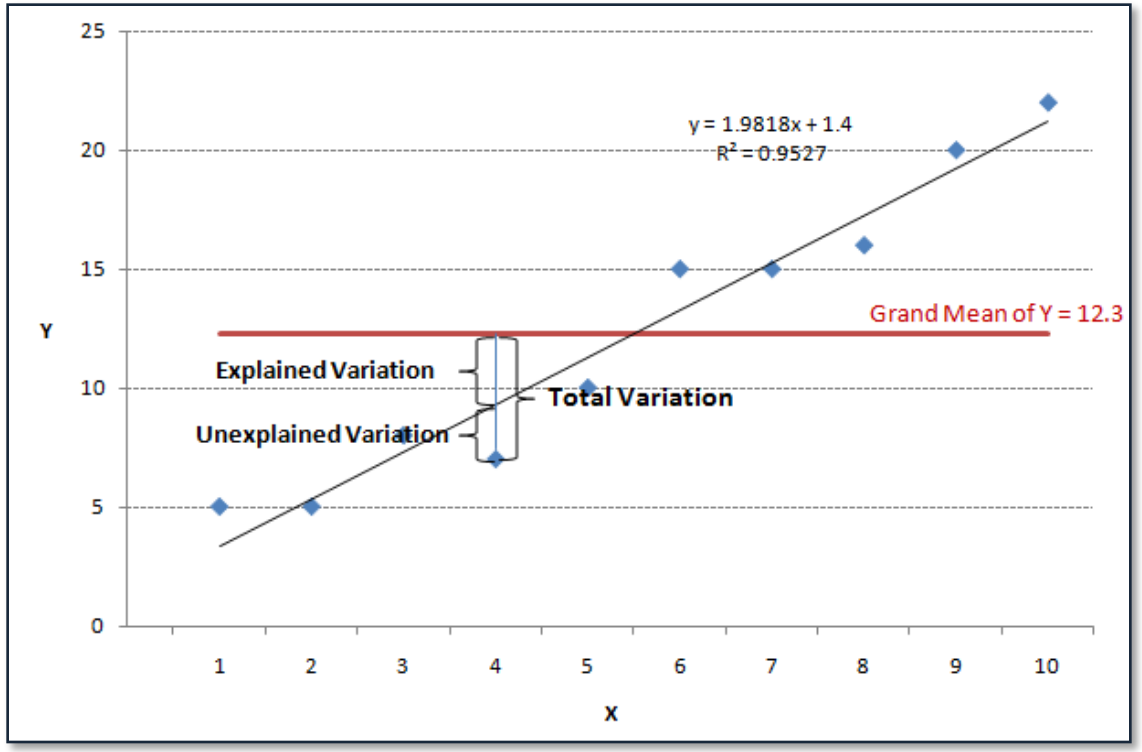
- The fitted value of the dependant variable:

$$\hat{Y}_i = \alpha * X_i + \beta \qquad \text{where } i = 1, 2, \ldots, n$$

- By using calculus, it can be shown the sum of squared error is minimal when

$$\beta = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \qquad \alpha = \bar{Y} - \beta \bar{X}$$

ANOVA in Simple Linear Regression



X: the independent variable that we use to predict;
Y: the dependent variable that we want to predict.

X	Y
1	5
2	5
3	8
4	7
5	10
6	15
7	15
8	16
9	20
10	22

Total Variation = Total Sums of Squares = $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Explained Variation = Regression Sums of Squares = $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

Unexplained Variation = Error Sums of Squares = $\sum_{i=1}^{n}(Y_i - \hat{Y})^2$

ANOVA in Simple Linear Regression

- Linear regression is also analysis of variance (ANOVA).
  - Variation Components:
    - Total Variation = Explained Variation + Unexplained Variation

      i.e., Total Sums of Squares = Regression Sums of Squares + Error Sums of Squares
  - Degrees of Freedom Components
    - Total Degrees of Freedom = Regression Degrees of Freedom + Residual Degrees of Freedom

      i.e., $n - 1 = (k - 1) + (n - k)$, where n is the number of data points, k is the number of predictors

ANOVA in Simple Linear Regression

- Whether the overall model is statistically significant can be tested by using F-test of ANOVA.
  - H_0: The model is not statistically significant.
  - H_a: The model is statistically significant.
- Test Statistic: $$F = \frac{MSR}{MSE} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \Big/ (k-1)}{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \Big/ (n-k)}$$
- Critical Statistic: F value in F table with (k – 1) degrees of freedom in the numerator and (n – k) degrees of freedom in the denominator.

- If $F \leq F_{critical}$, we fail to reject the null. There is no statistically significant relationship between X and Y.
- If $F > F_{critical}$, we reject the null. There is a statistically significant relationship between X and Y.

Coefficient of Determination

- R^2 (also called coefficient of determination) measures the proportion of variability in the data that can be explained by the model.

- R^2 ranges from 0 to 1. The higher R^2 is, the better the model can fit the actual data.

- How to calculate R^2:

$$R^2 = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

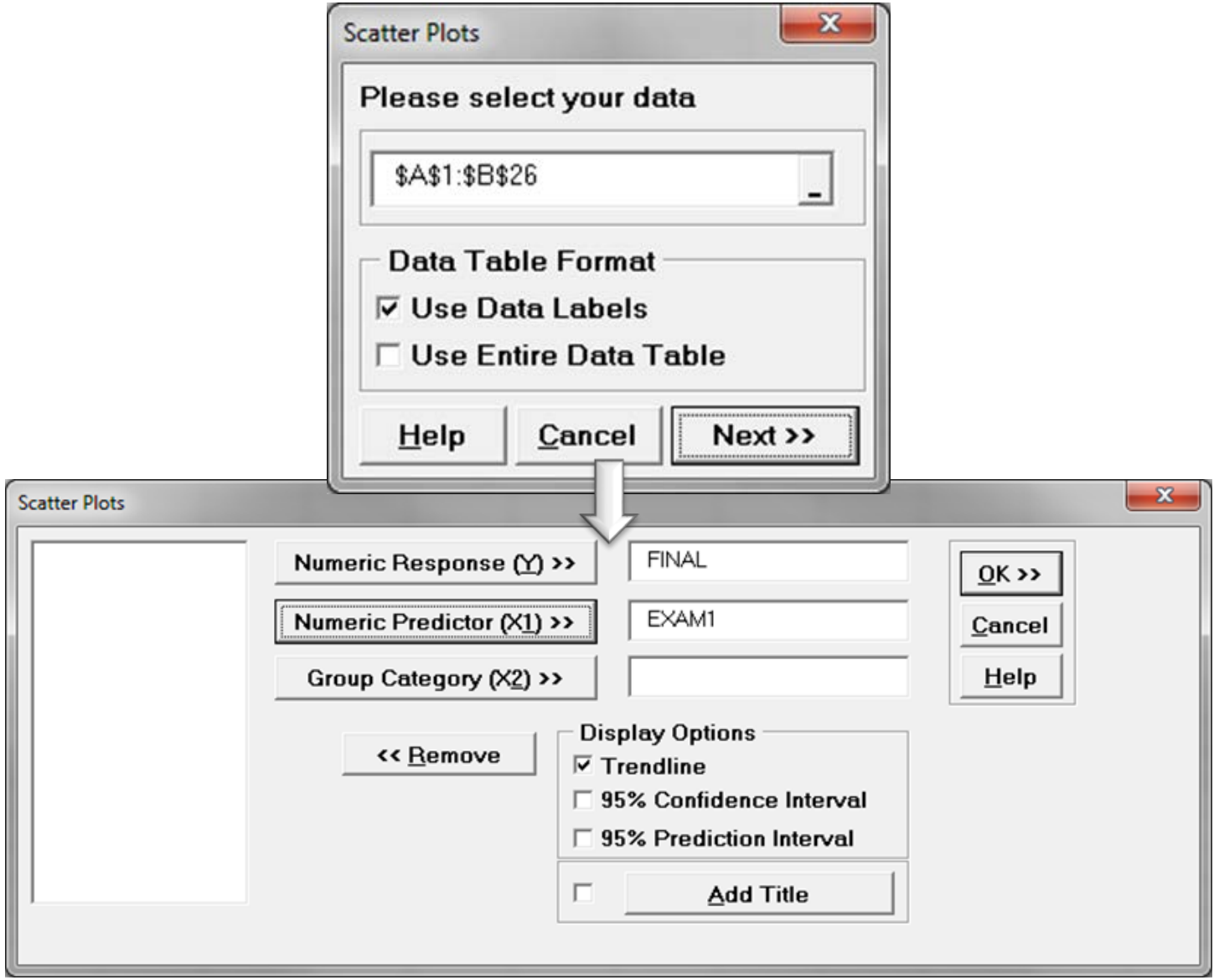
Use SigmaXL to Run a Simple Linear Regression

- Case Study
  - We are trying to see whether the score in exam one has any statistically significant relationship with the score in final exam. If yes, how much impact does exam one have on the final exam?
  - Data File: “Simple Linear Regression” tab in “Sample Data.xlsx”
- Step 1: Determine the dependent and independent variables. Both should be continuous variables.
  - Y (dependent variable) is the score of final exam.
  - X (independent variable) is the score of exam one.
  - Both variables are continuous.

Use SigmaXL to Run a Simple Linear Regression

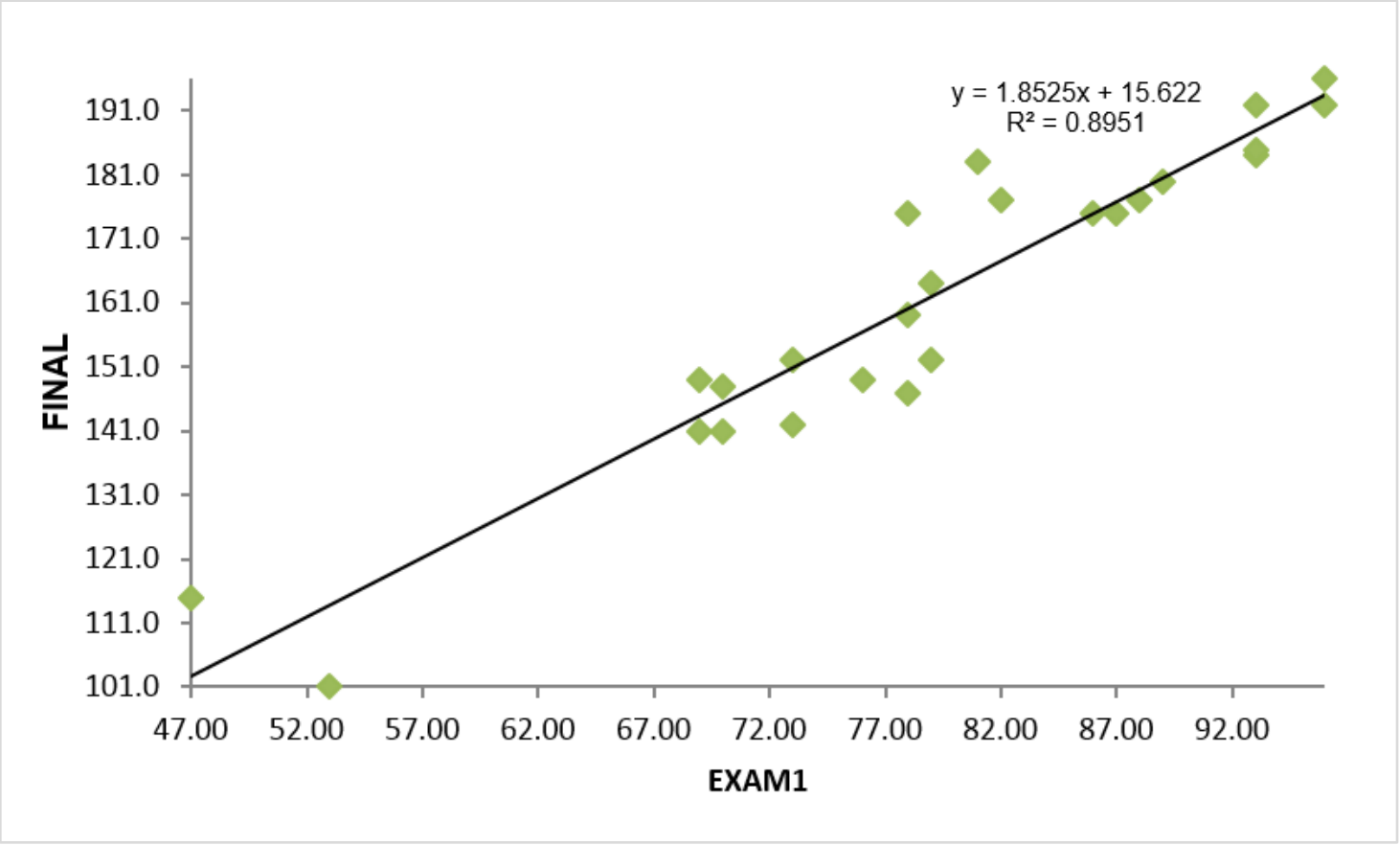
- Step 2: Create a scatter plot to eyeball whether there seems to be a linear relationship between X and Y.
  - Select the range of both independent and dependent variables in Excel.
  - Click SigmaXL -> Graphical Tools -> Scatter Plots
  - A new window named “Scatter Plots” pops up and the selected range appears automatically in the box below “Please select your data”.
  - Click “Next >>”
  - A new window also named “Scatter Plots” pops up.
  - Select “FINAL” as Numeric Response (Y)” and “EXAM1” as “Numeric Predictor (X1) >>”
  - Click “OK>>”
  - A scatter plot is generated in a new spreadsheet “Scatterplot(1)”.

Use SigmaXL to Run a Simple Linear Regression



Use SigmaXL to Run a Simple Linear Regression

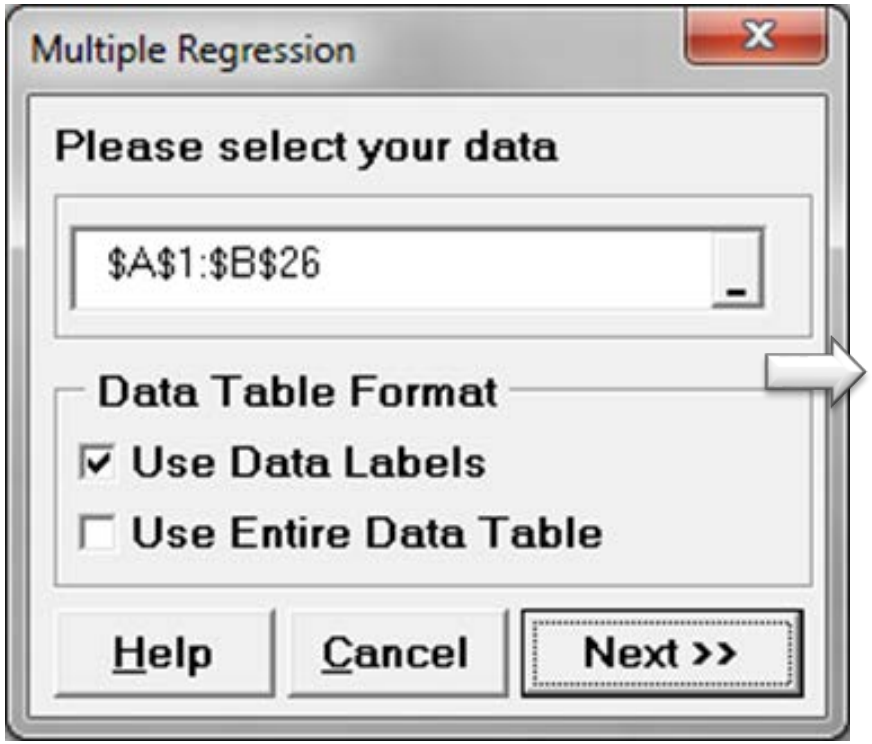
- Based on the scatter plot, the relationship between exam one and final seems linear. The higher the score on exam one, the higher the score on the final.

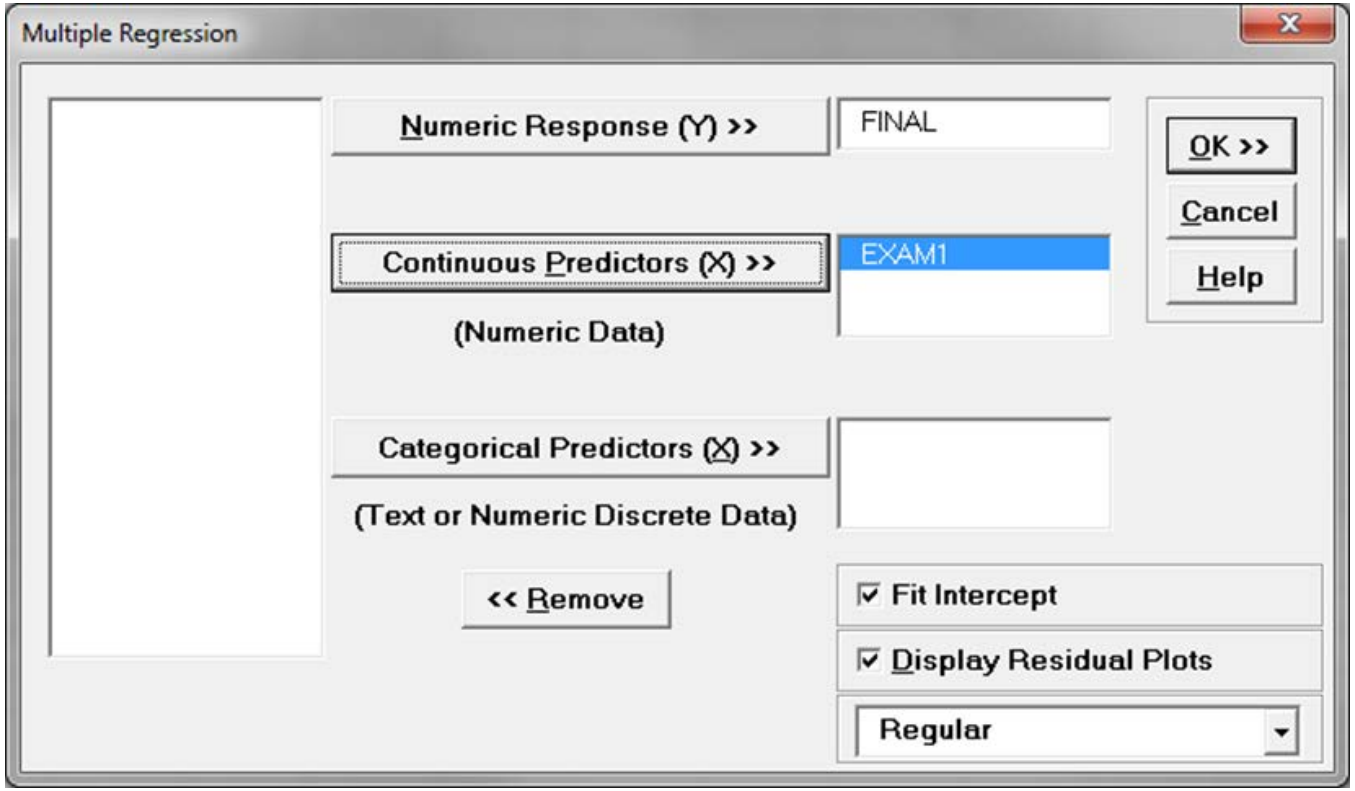


Use SigmaXL to Run a Simple Linear Regression

- Step 3: Run the Simple Linear Regression analysis.
  - Select the range of both independent and dependent variables in Excel.
  - Click SigmaXL -> Statistical Tools -> Regression -> Multiple Regression
  - A new window named “Multiple Regression” pops up and the selected range appears automatically in the box below “Please select your data”
  - Click “Next >>”
  - A new window also named “Multiple Regression” pops up
  - Select “FINAL” as “Numeric Response (Y)” and “EXAM1” as “Continuous Predictor (X)”
  - Click “OK>>”
  - The regression analysis results appear in the newly generated spreadsheet “Multiple Regression” and the residual analysis results appear in another new spreadsheet “Mult Reg Residuals (1)”.

Use SigmaXL to Run a Simple Linear Regression





Use SigmaXL to Run a Simple Linear Regression

- Step 4: Check whether the model is statistically significant. If not significant, we will need to re-examine the predictor or look for new predictors before continuing.

Multiple Regression Model: FINAL = (15.622) + (1.852) * EXAM1

Model Summary:	
R-Square	89.51%
R-Square Adjusted	89.05%
S (Root Mean Square Error)	7.957

The "Model Summary" section provides R^2 which measures the percentage of variation in the data set which can be explained by the model. 89.51% of the variability in the data can be accounted for by this linear regression model.

Parameter Estimates:

Predictor Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	15.622	10.575	1.477	0.1532		
EXAM1	1.852	0.132267	14.005	0.0000	1	1

Analysis of Variance for Model:

Source	DF	SS	MS	F	P
Model	1	12419	12419	196.15	0.0000
Error	23	1456.2	63.312		
Lack of Fit	14	837.01	59.786	0.869036	0.6074
Pure Error	9	619.17	68.796		
Total (Model + Error)	24	13875	578.12		

Durbin-Watson Test for Autocorrelation in Residuals:

DW Statistic	2.425
P-Value Positive Autocorrelation	0.8551
P-Value Negative Autocorrelation	0.1294

The "Analysis of Variance for Model" section provides a ANOVA table covering degrees of freedom, sum of squares and mean square information for total, model and error. The p-value of the F-test is lower than α level (0.05) indicating that the model is statistically significant.

Use SigmaXL to Run a Simple Linear Regression

- Step 5: Understand regression equation

Multiple Regression Model: FINAL = (15.622) + (1.852) * EXAM1

Model Summary:	
R-Square	89.51%
R-Square Adjusted	89.05%
S (Root Mean Square Error)	7.957

Parameter Estimates:

Predictor Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	15.622	10.575	1.477	0.1532		
EXAM1	1.852	0.132267	14.005	0.0000	1	1

Analysis of Variance for Model:

Source	DF	SS	MS	F	P
Model	1	12419	12419	196.15	0.0000
Error	23	1456.2	63.312		
Lack of Fit	14	837.01	59.786	0.869036	0.6074
Pure Error	9	619.17	68.796		
Total (Model + Error)	24	13875	578.12		

Durbin-Watson Test for Autocorrelation in Residuals:

DW Statistic	2.425
P-Value Positive Autocorrelation	0.8551
P-Value Negative Autocorrelation	0.1294

The estimates of slope and intercept are shown in the "Parameter Estimate" section.

In this example, Y = 15.622 + 1.852 * X.
A one unit increase in the score of Exam1 would increase the final score by 1.852.

Interpreting the Results

- Rsquare Adj = 89.0%
  - 89% of the variation in FINAL can be explained by EXAM1
- P-value of the F-test = 0.000
  - We have a statistically significant model
- Prediction Equation: 15.6 + 1.85 × EXAM1
  - 15.6 is the Y intercept, all equations will start with 15.6
  - 1.85 is the EXAM1 Coefficient: multiply it by EXAM1 score

- Let us say you are the professor and you want to use this prediction equation to estimate what two of your students might get on their final exam.

Interpreting the Results

- Let us assume the following:
  - Student “A” exam 1 results were: 79
  - Student “B” exam 1 results were: 94.

- Remember our prediction equation?
  - 15.6 + 1.85 × Exam1
  - Now apply the equation to each student
    - Student “A” Estimate: 15.6 + (1.85 × 79) = 161.8
    - Student “B” Estimate: 15.6 + (1.85 × 94) = 189.5

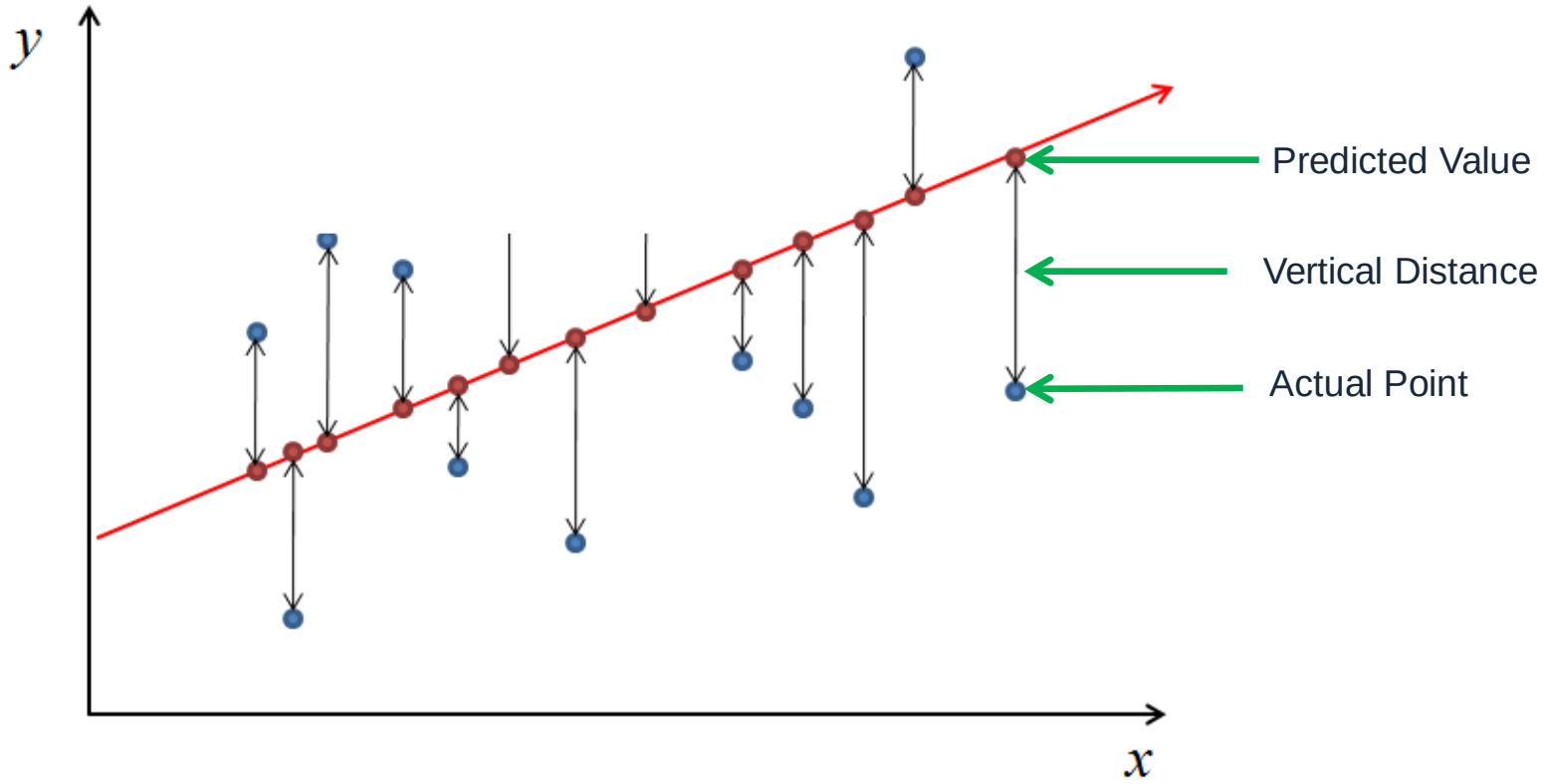
Now you can use your “magic” as the professor and allocate your time appropriately to the student(s) who you predict need the most help

Nice Work!

4.1.4 Residuals Analysis

What are Residuals?

- **Residuals** are the vertical differences between actual values and the predicted values or the “fitted line” created by the regression model.



Why Perform Residuals Analysis?

- *Regression equations* are generated on the basis of certain statistical assumptions.
- *Residuals analysis* helps to determine the validity of these assumptions.
- The *assumptions* are:
  - The residuals are normally distributed, mean equal to zero.
  - The residuals are independent.
  - The residuals have a constant variance.
  - The underlying population relationship is linear.

- If residuals performance does not meet the requirements, we will need to rebuild the model by replacing the predictor with a new one, adding new predictors, building non-linear models, and so on.

Use SigmaXL to Perform Residual Analysis

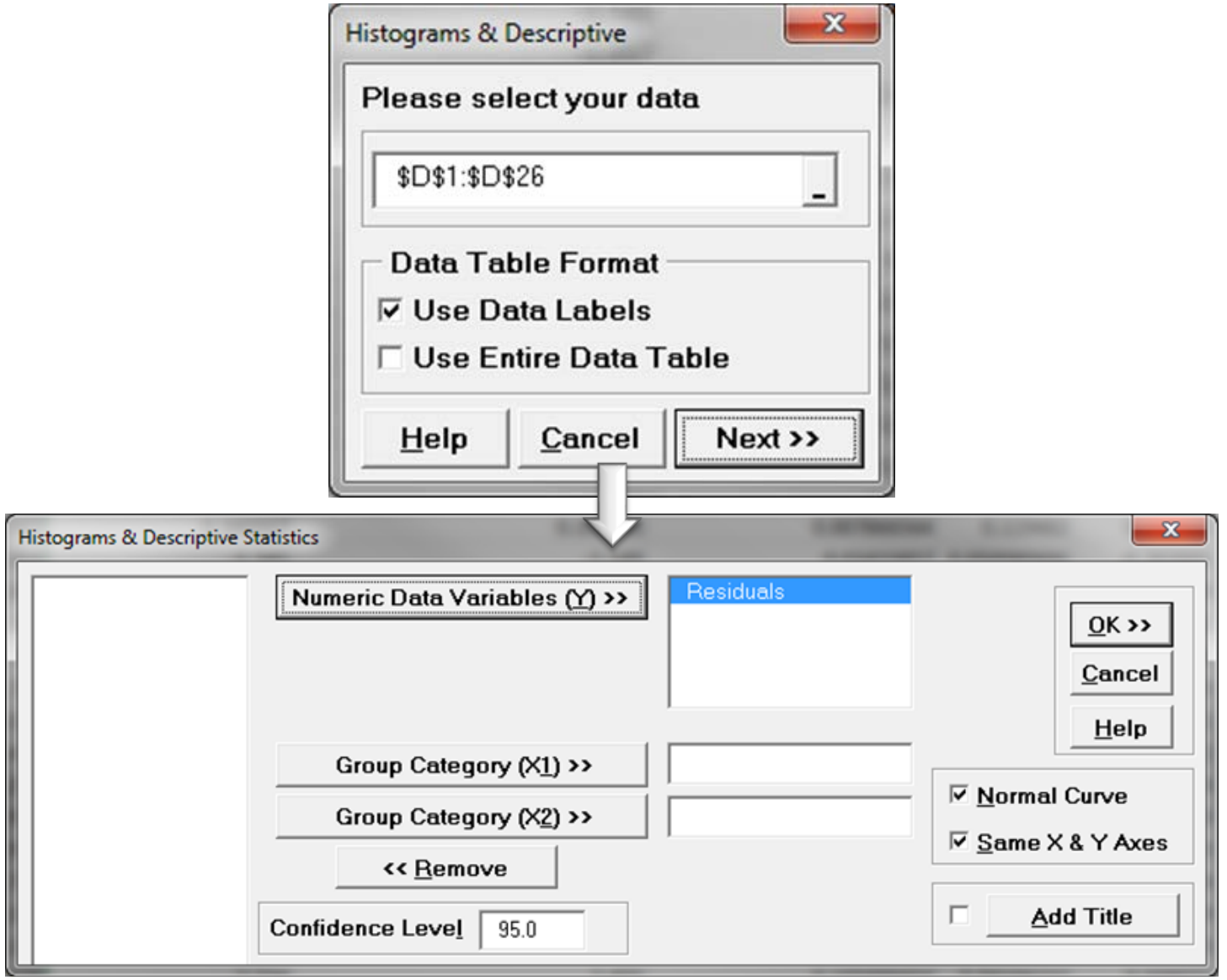
- The residuals of the model are automatically generated in a new tab “Mult Reg Residuals (1)”

EXAM1	FINAL	Predicted (Fitted) Values	Residuals	Standardized Residuals	Studentized (Deleted t) Residuals	Cook's Distance (Influence)	Leverage	DFITS
73.00	152	150.85	1.149	0.148139	0.144952	0.000578482	0.050080686	0.033282408
93.00	185	187.90	-2.900	-0.382909	-0.375691	0.007592687	0.093850167	-0.120906
89.00	180	180.49	-0.490450	-0.063827358	-0.062429917	0.000147241	0.067411632	-0.016784762
96.00	196	193.46	2.542	0.340507	0.333865	0.007866564	0.119482	0.122985
73.00	142	150.85	-8.851	-1.141	-1.149	0.03433857	0.050080686	-0.263885
53.00	101	113.80	-12.802	-1.830	-1.937	0.492983	0.227369	-1.050669249
69.00	149	143.44	5.559	0.723576	0.715866	0.019055818	0.067853748	0.193142
47.00	115	102.69	12.313	1.882	2.001	0.847132	0.323662	1.384
87.00	175	176.79	-1.786	-0.231147	-0.226329	0.001630039	0.057508234	-0.055907039
79.00	164	161.97	2.034	0.260912	0.255555	0.001418242	0.040000442	0.052165242
69.00	141	143.44	-2.441	-0.317794	-0.311493	0.00367579	0.067853748	-0.08404142
70.00	141	145.29	-4.294	-0.557356	-0.548824	0.010369267	0.062581515	-0.141804
93.00	184	187.90	-3.900	-0.514934	-0.506544	0.013731185	0.093850167	-0.163018
79.00	152	161.97	-9.966	-1.278	-1.297	0.034043928	0.040000442	-0.264780
70.00	148	145.29	2.706	0.351276	0.344480	0.004118888	0.062581515	0.089006254
93.00	192	187.90	4.100	0.541268	0.532774	0.015171528	0.093850167	0.171459
78.00	147	160.11	-13.113	-1.682	-1.757	0.059420777	0.04029887	-0.360035
81.00	183	165.67	17.329	2.224	2.455	0.105899454	0.04106152	0.508031
88.00	177	178.64	-1.638	-0.212574	-0.208106	0.00149813	0.062183611	-0.053587584
78.00	159	160.11	-1.113	-0.142843	-0.139765	0.000428394	0.04029887	-0.028640261
82.00	177	167.52	9.477	1.217	1.231	0.032812027	0.042421027	0.259021
86.00	175	174.93	0.066914252	0.008643486	0.008453509	2.10668E-06	0.053385503	0.00200753
78.00	175	160.11	14.887	1.910	2.036	0.076576038	0.04029887	0.417255
76.00	149	156.41	-7.409	-0.951551	-0.949513	0.020121332	0.042553662	-0.200176
96.00	192	193.46	-1.458	-0.195225	-0.191093	0.002585867	0.119482	-0.070392374

Use SigmaXL to Perform Residual Analysis

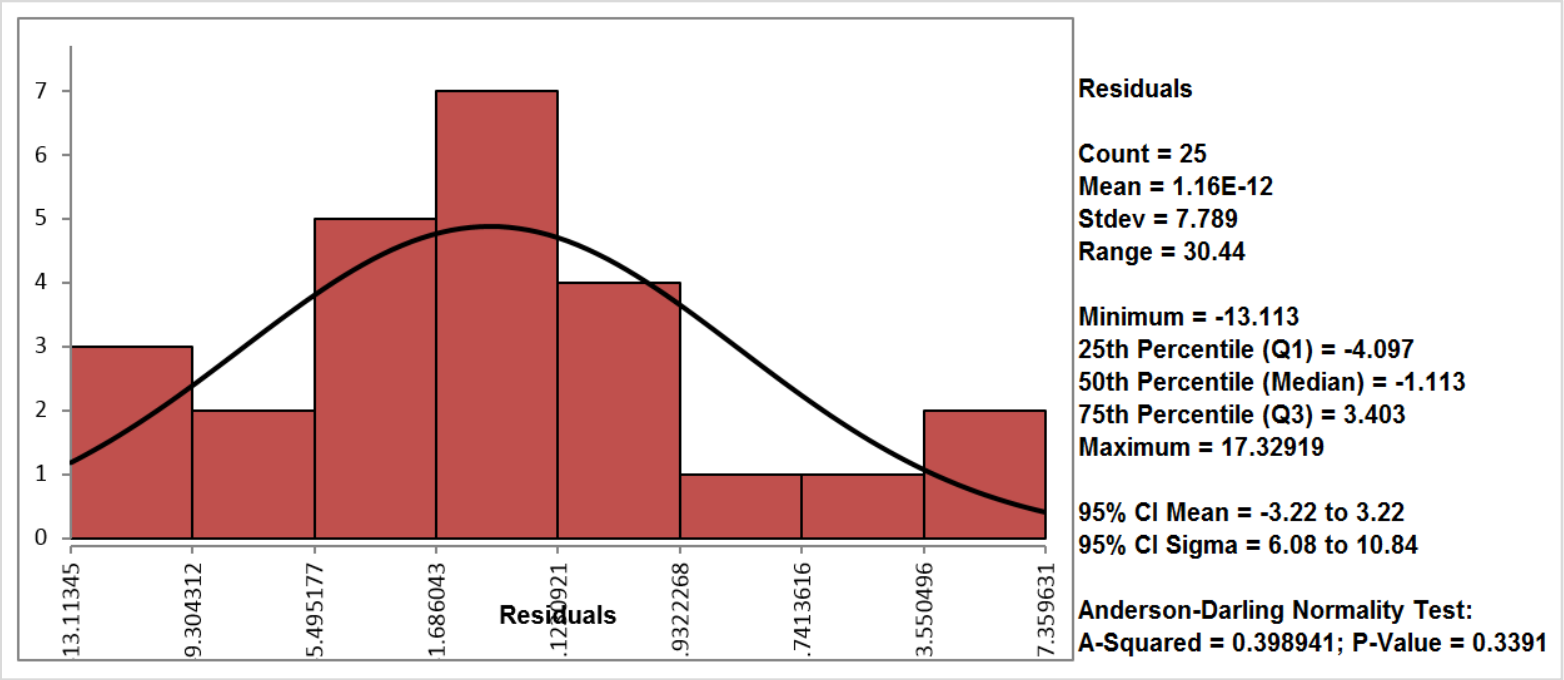
- Step 1: Check whether residuals are normally distributed around the mean of zero.
  - Select the range of residuals in spreadsheet “Mult Reg Residuals (1)”
  - Click SigmaXL -> Graphical Tool -> Histograms & Descriptive Statistics”
  - A new window named “Histograms & Descriptive” pops up and the selected range of residuals automatically appears in the box below “Please select your data”.
  - Click “Next >>”
  - A new window also named “Histograms & Descriptive” pops up.
  - Select “Residuals” as “Numeric Data Variables (Y)”
  - Click “OK>>”
  - The analysis results are shown automatically in the new spreadsheet “Hist Descript(1)”

Use SigmaXL to Perform Residual Analysis



Use SigmaXL to Perform Residual Analysis

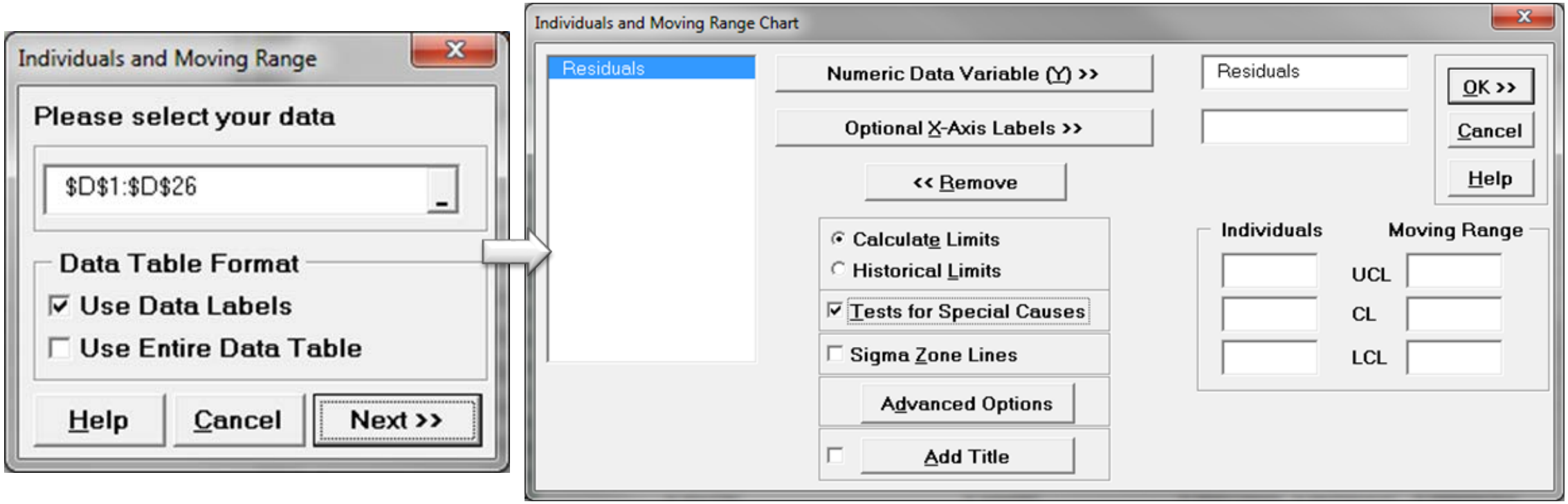
- The mean of residuals is 1.16E-12 which is approximately zero.
- Anderson-Darling Test is used to test the normality. Since the p-value (0.3391) is greater than the alpha level (0.05), we fail to reject the null and the residuals are normally distributed.
  - H_0: The residuals are normally distributed.
  - H_1: The residuals are not normally distributed.



Use SigmaXL to Perform Residual Analysis

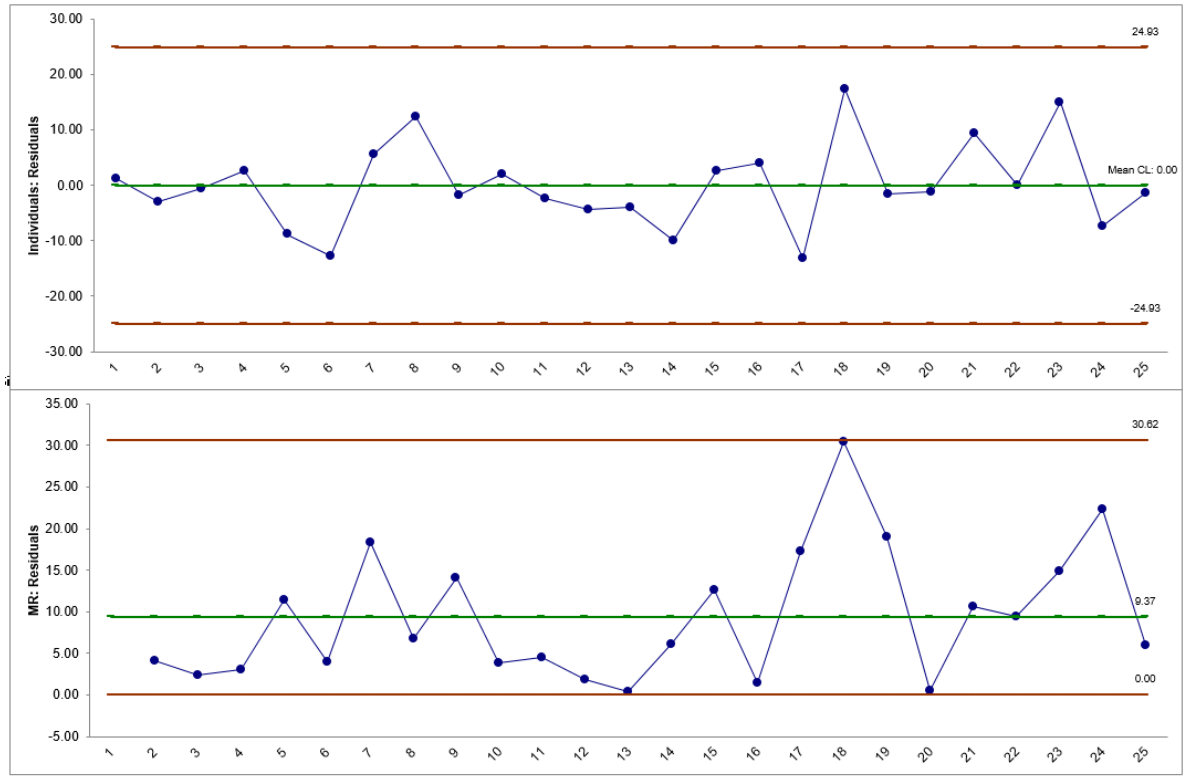
- Step 2: If the data are in time order, run the IR chart to check whether residuals are independent.
  - Select the range of residuals in spreadsheet “Mult Reg Residuals (1)”
  - Click SigmaXL -> Control Charts -> Individuals & Moving Range
  - A new window named “Individuals & Moving Range” pops up and the selected range of residuals automatically appears in the box below “Please select your data”.
  - A new window also named “Individuals & Moving Range” pops up.
  - Select “Residuals” as “Numeric Data Variables (Y)” and check the box of “Test for special causes”
  - Click “OK>>”
  - The analysis results are shown automatically in the new spreadsheet “Indiv & MR Charts (1)”

Use SigmaXL to Perform Residual Analysis



Use SigmaXL to Perform Residual Analysis

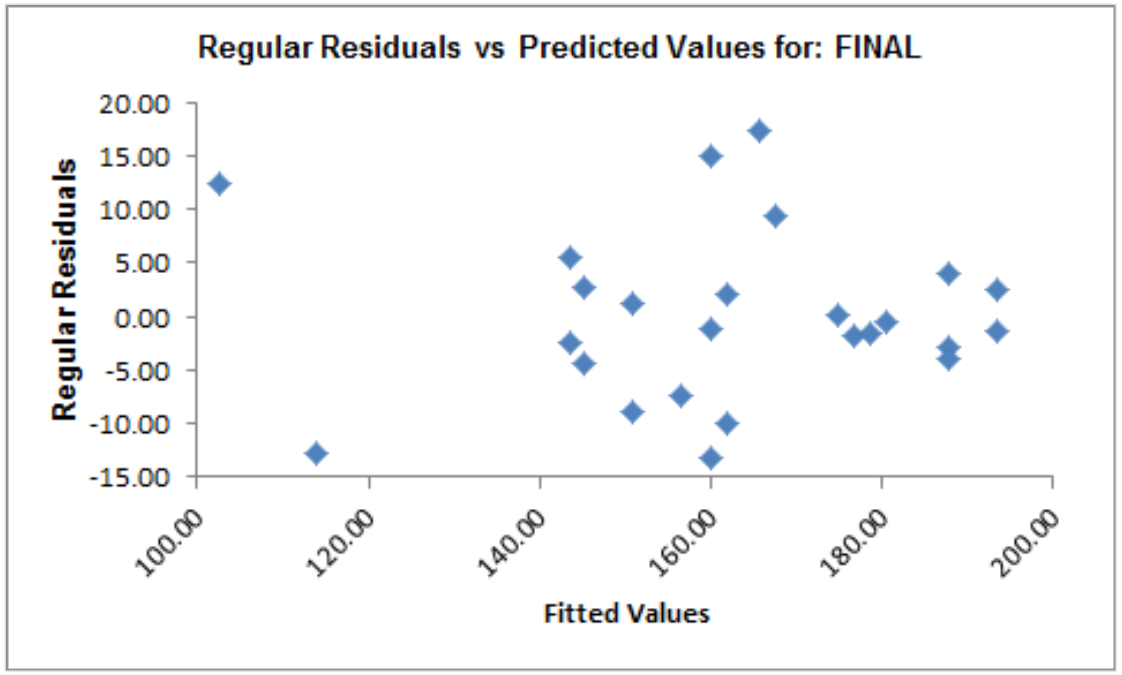
If no data points are out of control in both the I-Chart and MR charts, the residuals are independent of each other. If the residuals are not independent, it is possible that some important predicators are not included in the model. In this example, since the IR chart is in control, residuals are independent.



Tests for Special Causes - Individuals: Residuals
Number of Data Points Failing Tests = 0

Use SigmaXL to Perform Residual Analysis

- Step 3: Check if the residuals have equal variance across the predicted responses.
  - We are looking for the pattern in which residuals spread out evenly around zero from the top to the bottom.
  - The “Regular Residuals vs Predicted Values for: FINAL” chart is automatically generated in the right “Mult Reg Residulas (1)”.



4.2 Multiple Regression Analysis

Black Belt Training: Improve Phase

4.1 Simple Linear Regression
- 4.1.1 Correlation
- 4.1.2 X-Y Diagram
- 4.1.3 Regression Equations
- 4.1.4 Residuals Analysis

4.2 Multiple Regression Analysis
- 4.2.1 Non-Linear Regression
- 4.2.2 Multiple Linear Regression
- 4.2.3 Confidence Intervals
- 4.2.4 Residuals Analysis
- 4.2.5 Data Transformation, Box Cox
- 4.2.6 Stepwise Regression
- 4.2.7 Logistic Regression

4.3 Designed Experiments
- 4.3.1 Experiment Objectives
- 4.3.2 Experimental Methods
- 4.3.3 DOE Design Considerations

4.4 Full Factorial Experiments
- 4.4.1 2k Full Factorial Designs
- 4.4.2 Linear and Quadratic Models
- 4.4.3 Balanced and Orthogonal Designs
- 4.4.4 Fit, Model, and Center Points

4.5 Fractional Factorial Experiments
- 4.5.1 Designs
- 4.5.2 Confounding Effects
- 4.5.3 Experimental Resolution

4.2.1 Non-Linear Regression

Linear and Non-Linear

- The word *linear* originally comes from Latin word *linearis* meaning “created by lines.”

- A linear function in mathematics follows the following pattern (i.e., the output is proportional to its input):

$$f(x) = \alpha * x + \beta$$

$$f(x_1, x_2, ..., x_n) = \alpha_1 * x_1 + \alpha_2 * x_2 + ... + \alpha_n * x_n + \beta$$

- A non-linear function does not follow the above pattern. There are usually exponents, logarithms, power, polynomial components, and other non-linear functions of the independent variables and parameters.

Non-Linear Relationships Using Linear Models

- Many non-linear relationships can be transformed into linear relationships, and from there we can use linear regression methods to model the relationship.

- Some non-linear relationships cannot be transformed to linear ones and we need to apply other methods to build the non-linear models.

- In this section, we will focus on building non-linear regression models using linear transformation (i.e., transforming the independent or dependent variables or parameters to generate a linear function).

Assumptions in Using Non-Linear Regression

- The population relationship is non-linear based on a reliable underlying theory.

- Across the range of all the possible values of the independent variables, the non-linear relationship applies. It is possible that at some extreme values the relationship between variables changes dramatically.

Non-Linear Functions: Transforming to Linear

- Examples of non-linear functions that can be transformed to linear functions:
  - Exponential Function
  - Inverse Function
  - Polynomial Function
  - Power Function.

Exponential Function

- Exponential Function

$$Y = a \times b^{X}$$

- Transformation

$$\log Y = \log a + X \times \log b$$

Inverse Function

- Inverse Function

$$Y = a + b \times \frac{1}{X}$$

- Transformation

$$Y = a + b \times Z \quad \text{where} \quad Z = \frac{1}{X}$$

Polynomial Function

- Polynomial Function

$$Y = a + b \times X + c \times X^2$$

- Transformation

$$Y = a + b \times X + c \times Z \quad \text{where} \quad Z = X^2$$

Power Function

- Power Function

$$Y = a \times X^{b}$$

- Transformation

$$\log Y = \log a + b \times \log X$$

4.2.2 Multiple Linear Regression

What is Multiple Linear Regression?

- **Multiple linear regression** is a statistical technique to model the relationship between one dependent variable and two or more independent variables by fitting the data set into a linear equation.

- The difference between simple linear regression and multiple linear regression:
  - Simple linear regression only has one predictor.
  - Multiple linear regression has two or more predictors.

Multiple Linear Regression Equation

$$Y = \alpha_1 * X_1 + \alpha_2 * X_2 + ... + \alpha_p * X_p + \beta + e$$

- *Y* is the dependent variable (response).
- $X_1, X_2, \ldots, X_p$ are the independent variables (predictors). There are p predictors in total.
- Both dependent and independent variables are continuous.
- β is the intercept indicating the *Y* value when all the predictors are zeros.
- $\alpha_1, \alpha_2, \ldots, \alpha_p$ are the coefficients of predictors. They reflect the contribution of each independent variable in predicting the dependent variable.
- *e* is the residual term indicating the difference between the actual and the fitted response value.

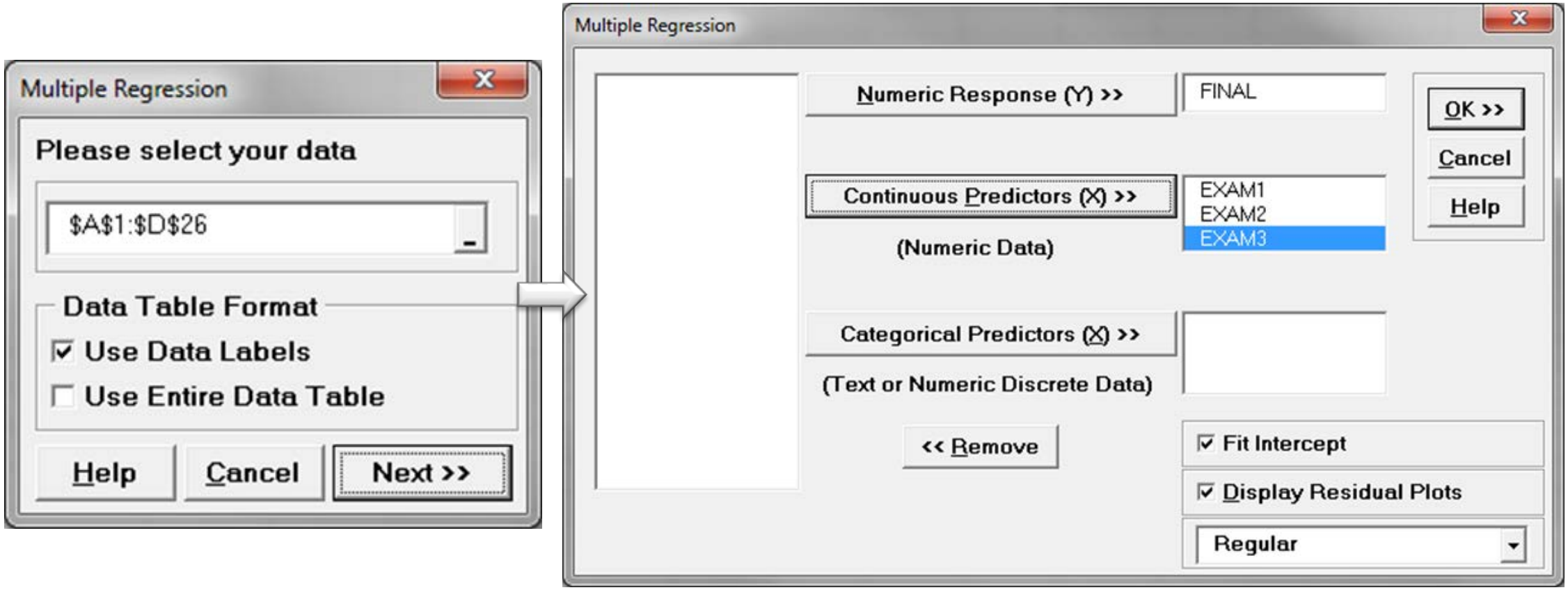
Use SigmaXL to Run a Multiple Linear Regression

- Case Study
  - We are trying to see whether the scores in exam one, two and three have any statistically significant relationship with the score in final exam. If so, how are they related to final exam score? Can we use the scores in exam one, two and three to predict the score in final exam?
  - Data File: “Multiple Regression Analysis” tab in “Sample Data.xlsx”
- Step 1: Determine the dependent and independent variables. All should be continuous.
  - Y (dependent variable) is the score of final exam.
  - X_1, X_2 and X_3 (independent variables) are the scores of exam one, two and three respectively.
  - All the variables are continuous.

Use SigmaXL to Run a Multiple Linear Regression

- Step 2: Start building the multiple linear regression model
  - Select the range of independent and dependent variables in Excel.
  - Click SigmaXL -> Statistical Tools -> Regression -> Multiple Regression
  - A new window named “Multiple Regression” pops up and the selected range appears automatically in the box below “Please select your data”
  - Click “Next >>”
  - A new window also named “Multiple Regression” pops up
  - Select “FINAL” as “Numeric Response (Y)” and “EXAM1”, “EXAM2” and “EXAM3” as “Continuous Predictor (X)”
  - Click “OK>>”
  - The regression analysis results appear in the newly generated spreadsheet “Multiple Regression” and the residual analysis results appear in another new spreadsheet “Mult Reg Residuals (1)”.

Use SigmaXL to Run a Multiple Linear Regression



Use SigmaXL to Run a Multiple Linear Regression

- Step 3: Check whether the whole model is statistically significant. If not, we need to re-examine the predictors or look for new predictors before continuing.

Analysis of Variance for Model:

Source	DF	SS	MS	F	P
Model	3	13732	4577.2	670.09	0.0000
Error	21	143.45	6.831		
Total (Model + Error)	24	13875	578.12		

H_0: The model is not statistically significant (i.e. all the parameters of predictors are not significantly different from zeros).
H_1: The model is statistically significant (i.e. at least one predictor parameter is significantly different from zero).

In this example, p-value is much smaller than alpha level (0.05), hence we reject the null and the model is statistically significant.

Use SigmaXL to Run a Multiple Linear Regression

- Step 4: Check whether multicollinearity exists in the model.
  - The VIF information is automatically generated in the “Parameter Estimate” table in the “Multiple Regression” spreadsheet.

Parameter Estimates:

Predictor Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	-4.336	3.764	-1.152	0.2623		
EXAM1	0.355938	0.121389	2.932	**0.0080**	7.807	0.128093
EXAM2	0.542519	0.100849459	5.379	**0.0000**	5.587	0.178990
EXAM3	1.167	0.103014055	11.333	**0.0000**	5.161	0.193750

Multicollinearity

- **Multicollinearity** is the situation when two or more independent variables in a multiple regression model are correlated with each other.

- Although multicollinearity does not necessarily reduce the predictability for the model as a whole, it may mislead the calculation for individual independent variables.

- To detect multicollinearity, we use VIF (Variance Inflation Factor) to quantify its severity in the model.

Variance Inflation Factor (1)

- VIF quantifies the degree of multicollinearity for each individual independent variable in the model.
- VIF calculation:
  - Assume we are building a multiple linear regression model using p predictors.

$$Y = \alpha_1 \times X_1 + \alpha_2 \times X_2 + ... + \alpha_p \times X_p + \beta$$

  - Two steps are needed to calculate VIF for X_1.
    - Step 1: Build a multiple linear regression model for X_1 by using X_2, X_3, …, X_p as predictors.

$$X_1 = a_2 \times X_2 + a_3 \times X_3 + ... + a_p \times X_p + b$$

    - Step 2: Use the R^2 generated by the linear model in step 1 to calculate the VIF for X_1.

$$VIF = \frac{1}{1 - R^2}$$

    - Apply the same methods to obtain the VIFs for other Xs. The VIF value ranges from one to positive infinity.

Variance Inflation Factor (2)

- Rules of thumb to analyze variance inflation factor (VIF):
  - If VIF = 1, there is no multicollinearity.
  - If 1 < VIF < 5, there is small multicollinearity.
  - If VIF ≥ 5, there is medium multicollinearity.
  - If VIF ≥ 10, there is large multicollinearity.

How to Deal With Multicollinearity

- Increase the sample size.
- Collect samples with a broader range for some predictors.
- Remove the variable with high multicollinearity and high p-value.
- Remove variables that are included more than once.
- Combine correlated variables to create a new one.
- In this section, we will focus on removing variables with high VIF and high p-value.

Use SigmaXL to Run a Multiple Linear Regression

- Step 5: Deal with Multicollinearity
  - Step 5.1: Identify a list of independent variables with VIF higher than 5. If no variable has VIF higher than 5, go to Step 6 directly.
  - Step 5.2: Among variables identified in Step 5.1, remove the one with the highest p-value.
  - Step 5.3: Run the model again, check the VIFs and repeat Step 5.1.
  - Note: we only remove one independent variable at a time.

Use SigmaXL to Run a Multiple Linear Regression

In this example, all three predictors have VIF higher than 5. Among them, EXAM1 has the highest p-value.

We will remove EXAM1 from the equation and run the model again.

Parameter Estimates:

Predictor Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	-4.336	3.764	-1.152	0.2623		
EXAM1	0.355938	0.121389	2.932	0.0080	7.807	0.128093
EXAM2	0.542519	0.100849459	5.379	0.0000	5.587	0.178990
EXAM3	1.167	0.103014055	11.333	0.0000	5.161	0.193750

Use SigmaXL to Run a Multiple Linear Regression

- Run the new multiple linear regression with only two predictors (i.e. EXAM2 and EXAM3).
- Check the VIFs of EXAM2 AND EXAM3 and they are both smaller than 5. Hence, there is little multicollinearity existing in the model.

Parameter Estimates:

Predictor Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	-4.338	4.366	-0.993485	0.3313		
EXAM2	0.722159	0.092917065	7.772	0.0000	3.525	0.283676
EXAM3	1.338	0.098747313	13.545	0.0000	3.525	0.283676

Use SigmaXL to Run a Multiple Linear Regression

- Step 6: Identify the statistically insignificant predictors. Remove one insignificant predictor at a time and run the model again. Repeat this step until all the predictors in the model are statistically significant.

  - Insignificant predictors are ones with p-value higher than alpha level (0.05). When p>alpha level, we fail to reject the null and the predictor isn't significant.
    - H_0: The predictor is not statistically significant.
    - H_1: The predictor is statistically significant.

  - As long as the p-value is greater than 0.05, remove the insignificant variables one at a time in the order of the highest p-value.

  - Once one insignificant variable is eliminated from the model, we need to rerun the model again to obtain new p-values for other predictors left in the new model.

Use SigmaXL to Run a Multiple Linear Regression

- In this example, both predictors' p-values are smaller than alpha level (0.05). As a result, we don't need to eliminate any variables from the model.

Parameter Estimates:

Predictor Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	-4.338	4.366	-0.993485	0.3313		
EXAM2	0.722159	0.092917065	7.772	0.0000	3.525	0.283676
EXAM3	1.338	0.098747313	13.545	0.0000	3.525	0.283676

Use SigmaXL to Run a Multiple Linear Regression

- Step 7: Display regression equation
  - The multiple linear regression equation appears automatically at the top of the spreadsheet “Multiple Regression (2)”
  - Parameter Estimates section provides the estimates of parameters in the linear regression equation.

Multiple Regression Model: FINAL = (-4.338) + (0.722159) * EXAM2 + (1.338) * EXAM3

Model Summary:	
R-Square	98.54%
R-Square Adjusted	98.41%
S (Root Mean Square Error)	3.031

Parameter Estimates:

Predictor Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	-4.338	4.366	-0.993485	0.3313		
EXAM2	0.722159	0.092917065	7.772	0.0000	3.525	0.283676
EXAM3	1.338	0.098747313	13.545	0.0000	3.525	0.283676

Interpreting the Results

- Rsquare Adj = 98.4%
  - 98% of the variation in FINAL can be explained by the predictor variables EXAM2 & EXAM3.
- P-value of the F-test = 0.000
  - We have a statistically significant model.
- Variables p-value:
  - Both are significant (less than 0.05).
- VIF
  - EXAM2 & EXAM3 are both below 5; we're in good shape!
- Equation: -4.34 + 0.722*EXAM2 + 1.34*EXAM3
  - -4.34 is the Y intercept, all equations will start with -4.34.
  - 0.722 is the EXAM2 coefficient; multiply it by EXAM2 score.
  - 1.34 is the EXAM3 coefficient; multiply it by EXAM3 score.

Interpreting the Results

- Let us say you are the professor again, and this time you want to use your prediction equation to estimate what one of your students might get on their final exam.

- Assume the following:
  - Exam 2 results were: 84
  - Exam 3 results were: 102.

- Use your equation: -4.34 + 0.722*EXAM2 + 1.34*EXAM3

- Predict your student's final exam score:
  - -4.34 + (0.722*84) + (1.34*102) =-4.34 + 60.648 + 136.68 = **192.988**

Nice work again! Now you can use your "magic" as the smart and efficient professor and allocate your time to other students because this one projects to perform much better than the average score of 162.

4.2.3 Confidence & Prediction Intervals

Prediction

- The purpose of building a regression model is not only to understand what happened in the past but more importantly to *predict* the future based on the past.

- By plugging the values of independent variables into the regression equation, we obtain the estimation/prediction of the dependent variable.

Uncertainty of Prediction

- We build the regression model using the sample data to describe as close as possible the true population relationship between dependent and independent variables.
- Due to noise in the data, the prediction will probably differ from the true response value.
- However, the true response value might fall in a range around the prediction with some certainty.
- To measure the uncertainty of the prediction, we need confidence interval and prediction interval.

Confidence Interval

- The **confidence interval** of the prediction is a range in which the population mean of the dependent variable would fall with some certainty, given specified values of the independent variables.
- The width of confidence interval is related to:
  - Sample size
  - Confidence level
  - Variation in the data.
- We build the model based on a sample set $\{y_1, y_2, \ldots, y_n\}$. The confidence interval is used to estimate the value of the population mean μ of the underlying population.
- The focus of the confidence interval are the unobservable population parameters.
- The confidence interval accounts for the uncertainty in the estimates of regression parameters.

Prediction Interval

- The prediction interval is a range in which future values of the dependent variable would fall with some certainty, given specified values of the independent variables.

- We build the model based on a sample set {y1, y2,…, yn}. The prediction interval is used to estimate the value of future observation y_{n+1}.

- The focus of the prediction interval are the future observations.

- Prediction interval is wider than confidence interval because it accounts for the uncertainty in the estimates of regression parameters and the uncertainty of the new measurement.

Use SigmaXL to Obtain Prediction

- On the right side of the newly generated spreadsheet “Multiple Regression (2)” is the table “Predicted Response Calculator”.
- By entering the values of the independent variables into the table “Predicted Response Calculator”, the predicted response would appear automatically. In this case, if we enter 85 and 75 as the values for EXAM2 and EXAM3 respectively, the predicted response is 157.359.

Predicted Response Calculator:

Predictors	Enter Settings:	Predicted Response	Lower 95% CI	Upper 95% CI	Lower 95% PI	Upper 95% PI
EXAM2						
EXAM3						

Predicted Response Calculator:

Predictors	Enter Settings:	Predicted Response	Lower 95% CI	Upper 95% CI	Lower 95% PI	Upper 95% PI
EXAM2	85	157.3592352	154.7395349	159.9789354	150.5483937	164.1700766
EXAM3	75					

Use SigmaXL to Obtain Confidence Interval

- On the right side of the newly generated spreadsheet “Multiple Regression (2)” is the table “Predicted Response Calculator”.
- By entering the values of the independent variables into the table “Predicted Response Calculator”, the upper and lower 95% confidence levels would appear automatically. In this case, if we enter 85 and 75 as the values for EXAM2 and EXAM3 respectively, the 95% confidence interval for the response is [154.74,159.98].

Predicted Response Calculator:

Predictors	Enter Settings:	Predicted Response	Lower 95% CI	Upper 95% CI	Lower 95% PI	Upper 95% PI
EXAM2	85	157.3592352	154.7395349	159.9789354	150.5483937	164.1700766
EXAM3	75					

Use SigmaXL to Obtain Prediction Interval

- On the right side of the newly generated spreadsheet “Multiple Regression (2)” is the table “Predicted Response Calculator”.
- By entering the values of the independent variables into the table “Predicted Response Calculator”, the upper and lower 95% prediction levels would appear automatically. In this case, if we enter 85 and 75 as the values for EXAM2 and EXAM3 respectively, the 95% prediction interval for the response is [150.55,164.17].

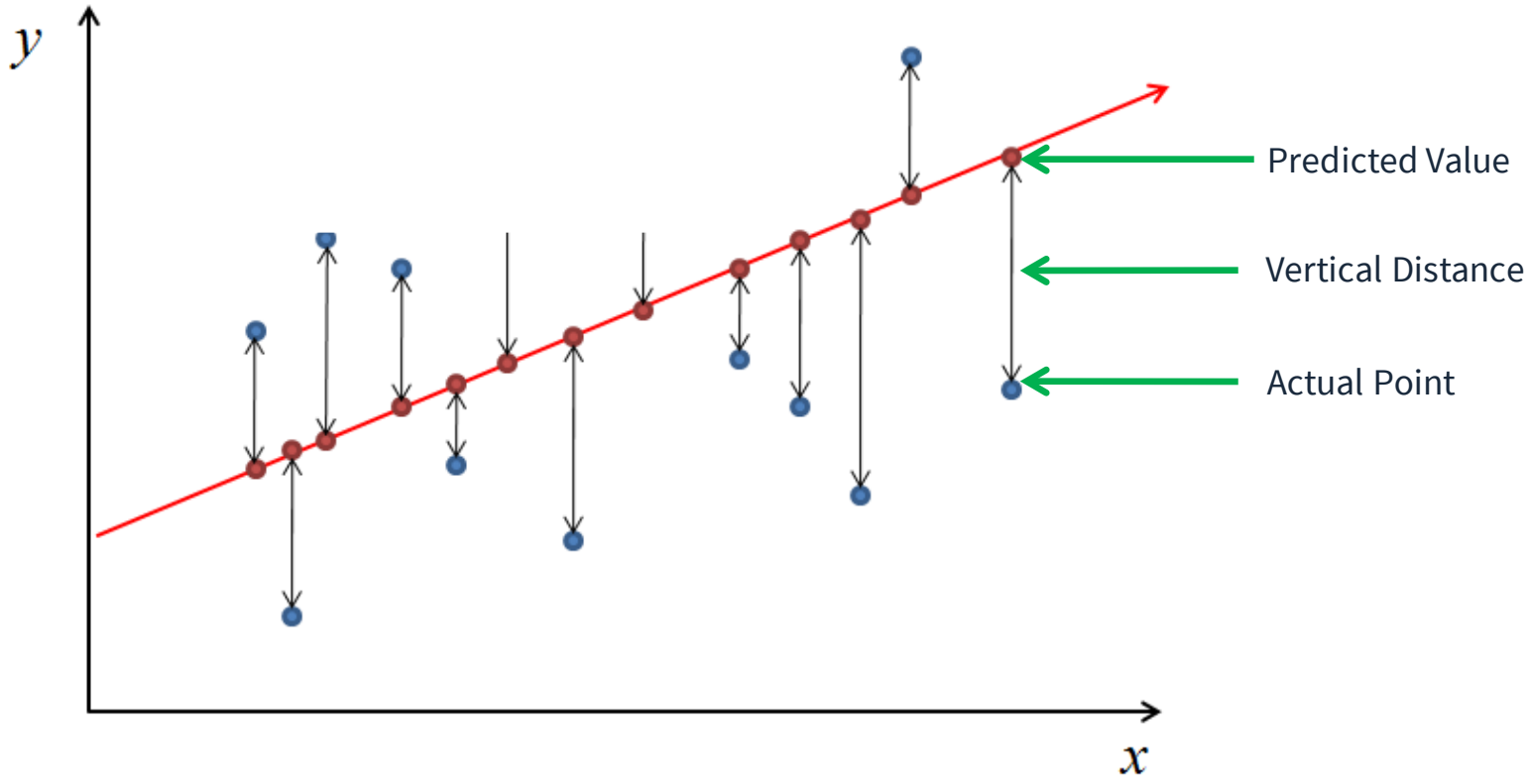
Predicted Response Calculator:

Predictors	Enter Settings:	Predicted Response	Lower 95% CI	Upper 95% CI	Lower 95% PI	Upper 95% PI
EXAM2	85	157.3592352	154.7395349	159.9789354	150.5483937	164.1700766
EXAM3	75					

4.2.4 Residuals Analysis

Remember what Residuals Are?

- **Residuals** are the vertical difference between actual values and the predicted values or the “fitted line” created by the regression model.



Why Perform Residuals Analysis?

- The *regression equation* generated based on the sample data can make accurate statistical inference only if certain assumptions are met. Residuals analysis can help to validate these assumptions. The following assumptions must be met to ensure the reliability of the linear regression model:
  - The errors are normally distributed with mean equal to zero.
  - The errors are independent.
  - The errors have a constant variance.
  - The underlying population relationship is linear.

- If the residuals performance does not meet the requirement, we will need to rebuild the model by replacing the predictors with new ones, adding new predictors, building non-linear models, and so on.

Use SigmaXL to Perform Residual Analysis

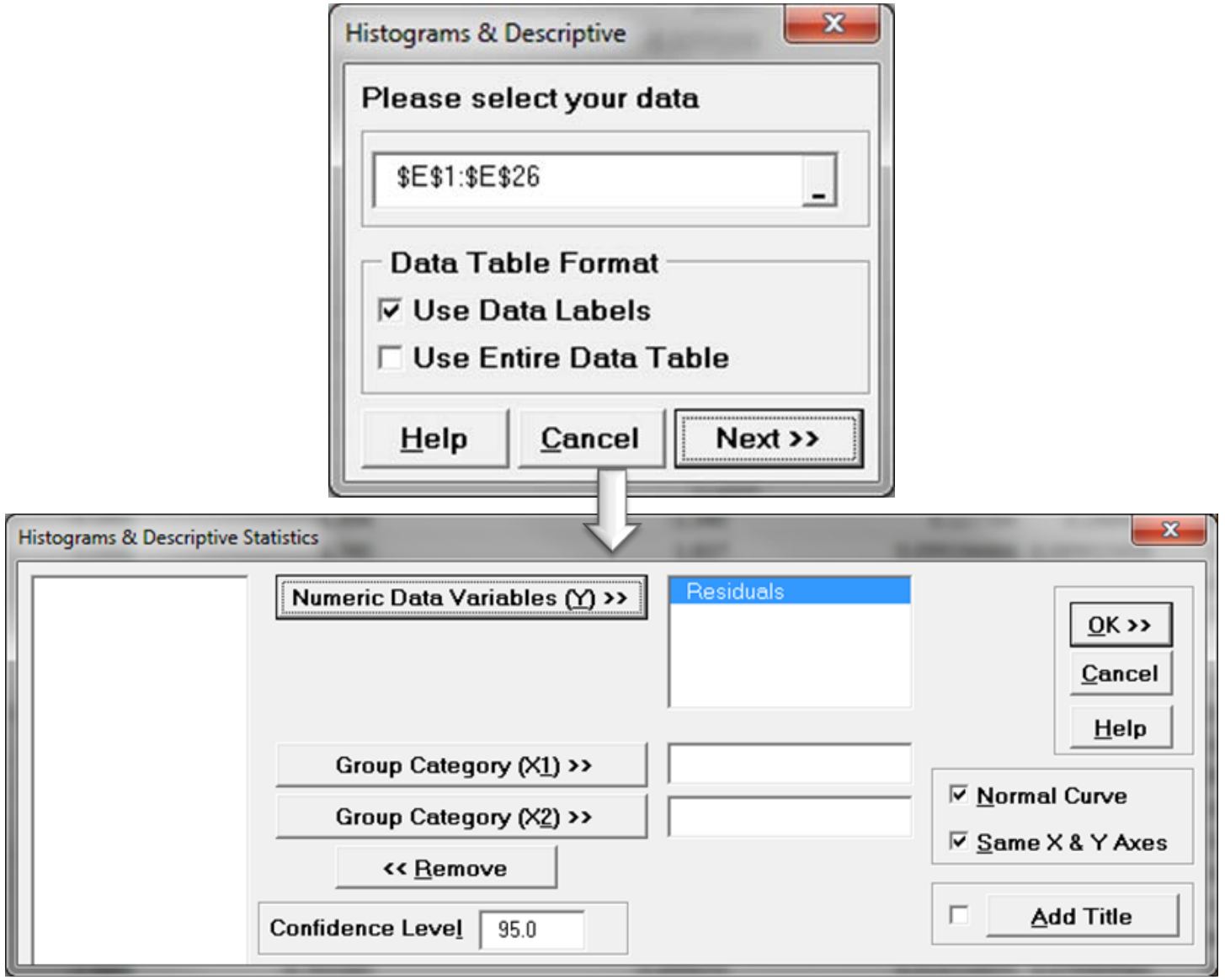
- The residuals of the model are automatically saved in column E of the spreadsheet “Mult Reg Residuals (2)”

EXAM2	EXAM3	FINAL	Predicted (Fitted) Values	Residuals	Standardized Residuals	Studentized (Deleted t) Residuals	Cook's Distance (Influence)	Leverage	DFITS
80.00	75.00	152	153.75	-1.748	-0.604782	-0.595851	0.012132265	0.090503798	-0.187962
88.00	93.00	185	183.60	1.399	0.481985	0.473409	0.007021224	0.083133126	0.142551
91.00	90.00	180	181.75	-1.755	-0.602165	-0.593229	0.009916385	0.075822698	-0.169920
98.00	100.00	196	200.19	-4.185	-1.494	-1.540	0.127764	0.146498	-0.638123
66.00	70.00	142	136.95	5.049	1.745	1.837	0.099194464	0.089015606	0.574192
46.00	55.00	101	102.44	-1.445	-0.586111	-0.577159	0.058665657	0.338767	-0.413113
74.00	77.00	149	152.09	-3.091	-1.044874737	-1.047165128	0.01835601	0.048017553	-0.235180
56.00	60.00	115	116.35	-1.354	-0.497839	-0.489156	0.020029274	0.195133	-0.240853
79.00	90.00	175	173.09	1.911	0.673598	0.665005	0.021439054	0.124152	0.250373
70.00	88.00	164	163.91	0.085482476	0.03309454	0.03233445	0.000137786	0.274000	0.01986424
70.00	73.00	141	143.85	-2.852	-0.972827	-0.971588	0.02188351	0.064869329	-0.255897
65.00	74.00	141	141.58	-0.578552	-0.202676	-0.198201	0.001749542	0.113298	-0.070847891
95.00	91.00	184	185.98	-1.981	-0.693767	-0.685354	0.020388413	0.112752	-0.244317
80.00	73.00	152	151.07	0.926582	0.326574	0.319842	0.005032631	0.124009	0.120340
73.00	78.00	148	152.71	-4.706	-1.596	-1.658	0.048656896	0.054187659	-0.396971
89.00	96.00	192	188.34	3.664	1.285	1.305	0.071697165	0.115235	0.471145
75.00	68.00	147	140.78	6.225	2.227	2.472	0.290759	0.149601	1.036767948
90.00	93.00	183	185.05	-2.045	-0.703385	-0.695073	0.014334655	0.079969561	-0.204923
92.00	86.00	177	177.13	-0.126981	-0.044483293	-0.043462507	8.42809E-05	0.113301	-0.015536136
83.00	77.00	159	158.59	0.410057	0.142551	0.139338	0.000749184	0.099588189	0.046339771
86.00	90.00	177	178.14	-1.144	-0.389899	-0.382257	0.003411602	0.063078172	-0.099184505
82.00	89.00	175	173.92	1.08206773	0.370923	0.363533	0.003661887	0.073943041	0.102724415
83.00	85.00	175	169.29	5.710	1.926	2.064	0.056718817	0.043847185	0.442011
83.00	71.00	149	150.56	-1.565	-0.588475	-0.579524	0.034582522	0.230524	-0.317200
93.00	95.00	192	189.89	2.113	0.733493	0.725556	0.019210696	0.096755911	0.237469

Use SigmaXL to Perform Residual Analysis

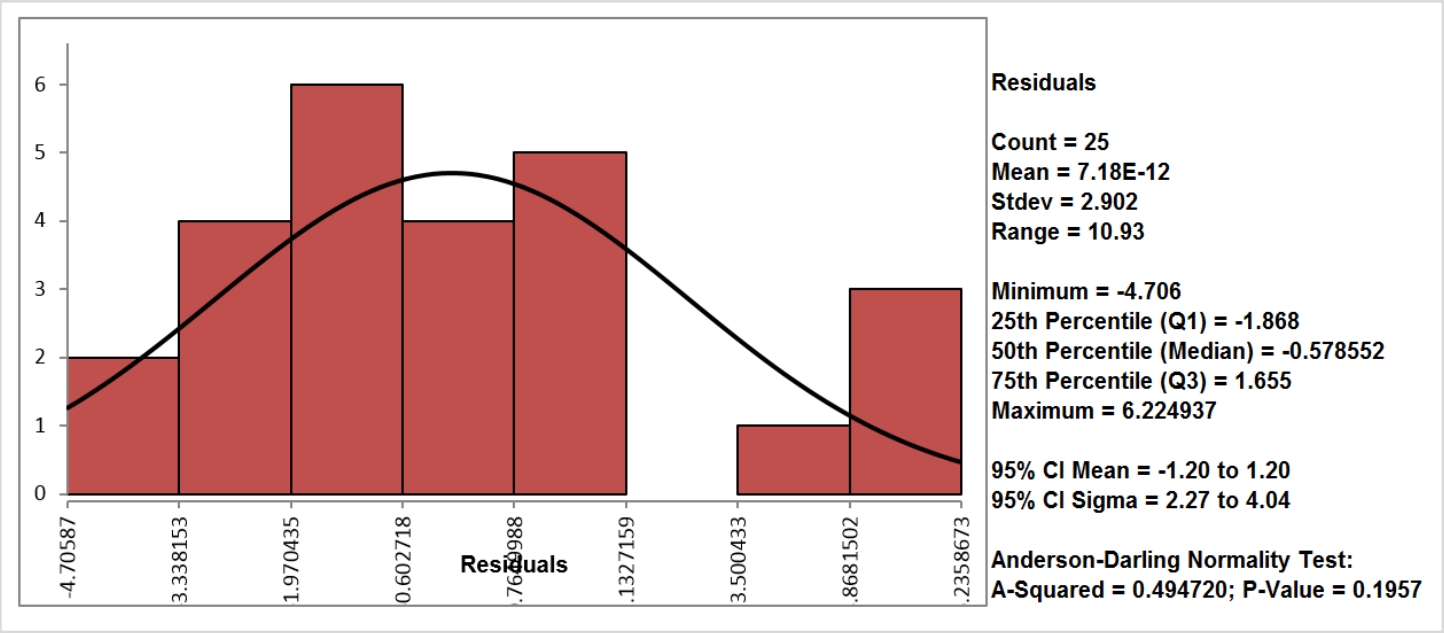
- Step 1: Check whether residuals are normally distributed around the mean of zero.
  - Select the range of residuals in spreadsheet “Mult Reg Residuals (2)”
  - Click SigmaXL -> Graphical Tool -> Histograms & Descriptive Statistics
  - A new window named “Histograms & Descriptive” pops up and the selected range of residuals automatically appears in the box below “Please select your data”.
  - Click “Next >>”
  - A new window also named “Histograms & Descriptive” pops up.
  - Select “Residuals” as “Numeric Data Variables (Y)”
  - Click “OK>>”
  - The analysis results are shown automatically in the new spreadsheet “Hist Descript(1)”

Use SigmaXL to Perform Residual Analysis



Use SigmaXL to Perform Residual Analysis

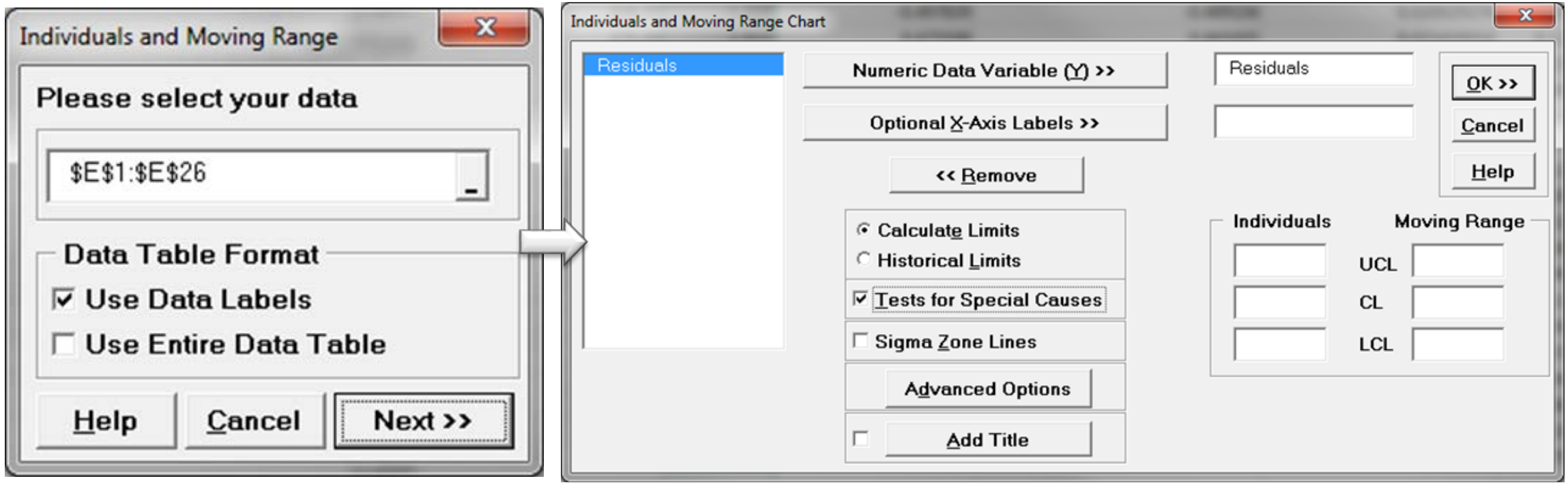
- The mean of residuals is 7.18E-12 which is approximately zero.
- Anderson-Darling Test is used to test the normality. Since the p-value (0.1957) is greater than the alpha level (0.05), we fail to reject the null and the residuals are normally distributed.
  - H_0: The residuals are normally distributed.
  - H_1: The residuals are not normally distributed.



Use SigmaXL to Perform Residual Analysis

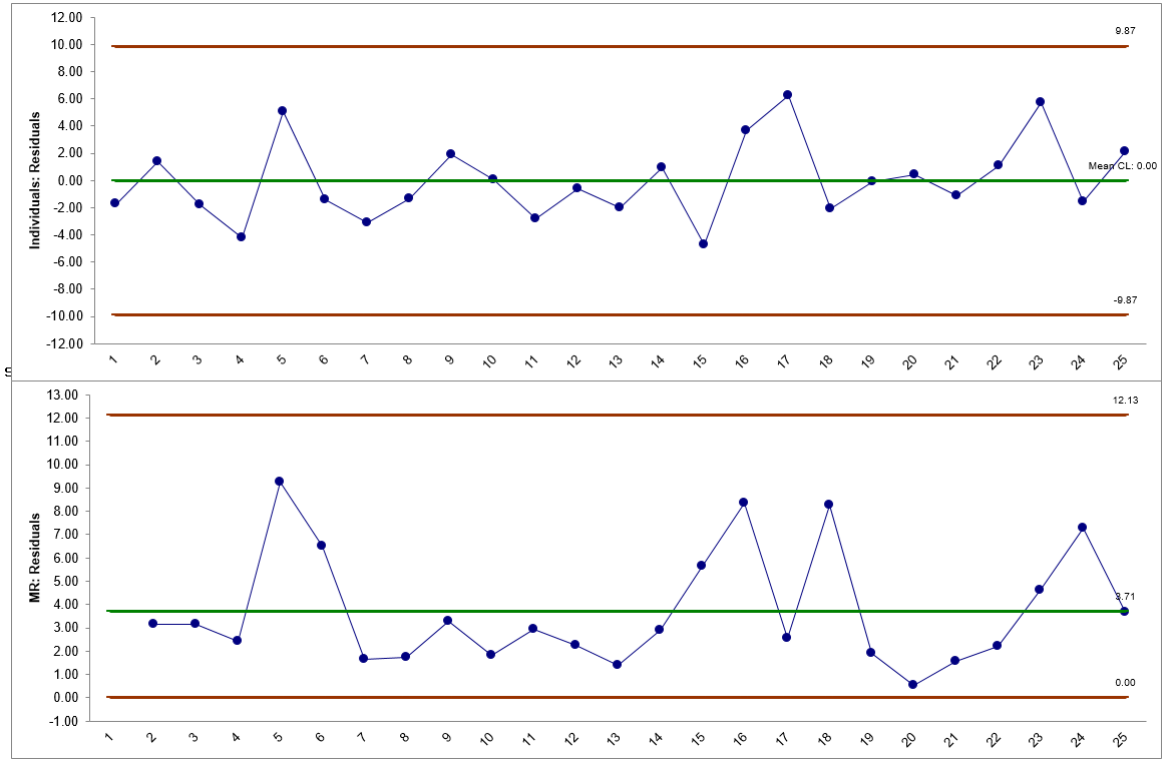
- Step 2: If the data are in time order, run the IR chart to check whether residuals are independent.
  - Select the range of residuals in spreadsheet “Mult Reg Residuals (2)”
  - Click SigmaXL -> Control Charts -> Individuals & Moving Range
  - A new window named “Individuals & Moving Range” pops up and the selected range of residuals automatically appears in the box below “Please select your data”.
  - A new window also named “Individuals & Moving Range” pops up.
  - Select “Residuals” as “Numeric Data Variables (Y)” and check the box of “Test for special causes”
  - Click “OK>>”
  - The analysis results are shown automatically in the new spreadsheet “Indiv & MR Charts (1)”

Use SigmaXL to Perform Residual Analysis



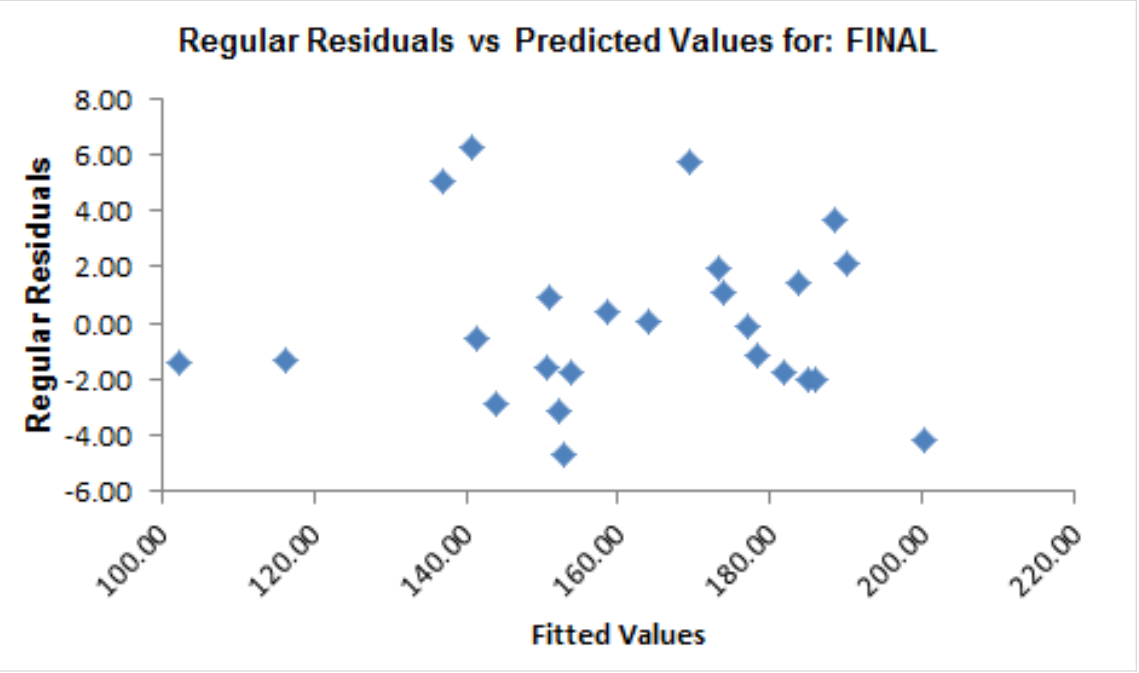
Use SigmaXL to Perform Residual Analysis

- If no data points are out of control in both the I-Chart and MR charts, the residuals are independent of each other.
- If the residuals are not independent, it is possible that some important predicators are not included in the model.
- In this example, since the IR chart is in control, residuals are independent.



Use SigmaXL to Perform Residual Analysis

- Step 3: Check whether residuals have equal variance across the predicted responses.
  - We are looking for the patterns in which residuals spread out evenly around zero from the top to the bottom.
  - The “Regular Residuals vs Predicted Values for: FINAL” chart is automatically generated on the right side of the sheet “Mult Reg Residulas (2)”.



4.2.5 Data Transformation

What is the Box-Cox Transformation?

- When a response does not fit the model well, sometimes using a transformation of the response can improve the fit.

- **Power transformation** is a class of transformation functions that raise the response to some power. For example, a square root transformation converts X to $X^{1/2}$.

- **Box-Cox transformation** is a popular power transformation method developed by George E. P. Box and David Cox.

Box-Cox Transformation Formula

- The formula of the Box-Cox transformation is:

$$\begin{cases} y = \dfrac{x^{\lambda} - 1}{\lambda} & \text{where } \lambda \neq 0 \\ y = \ln x & \text{where } \lambda = 0 \end{cases}$$

y is the transformation result.
x is the variable under transformation.
λ is the transformation parameter.

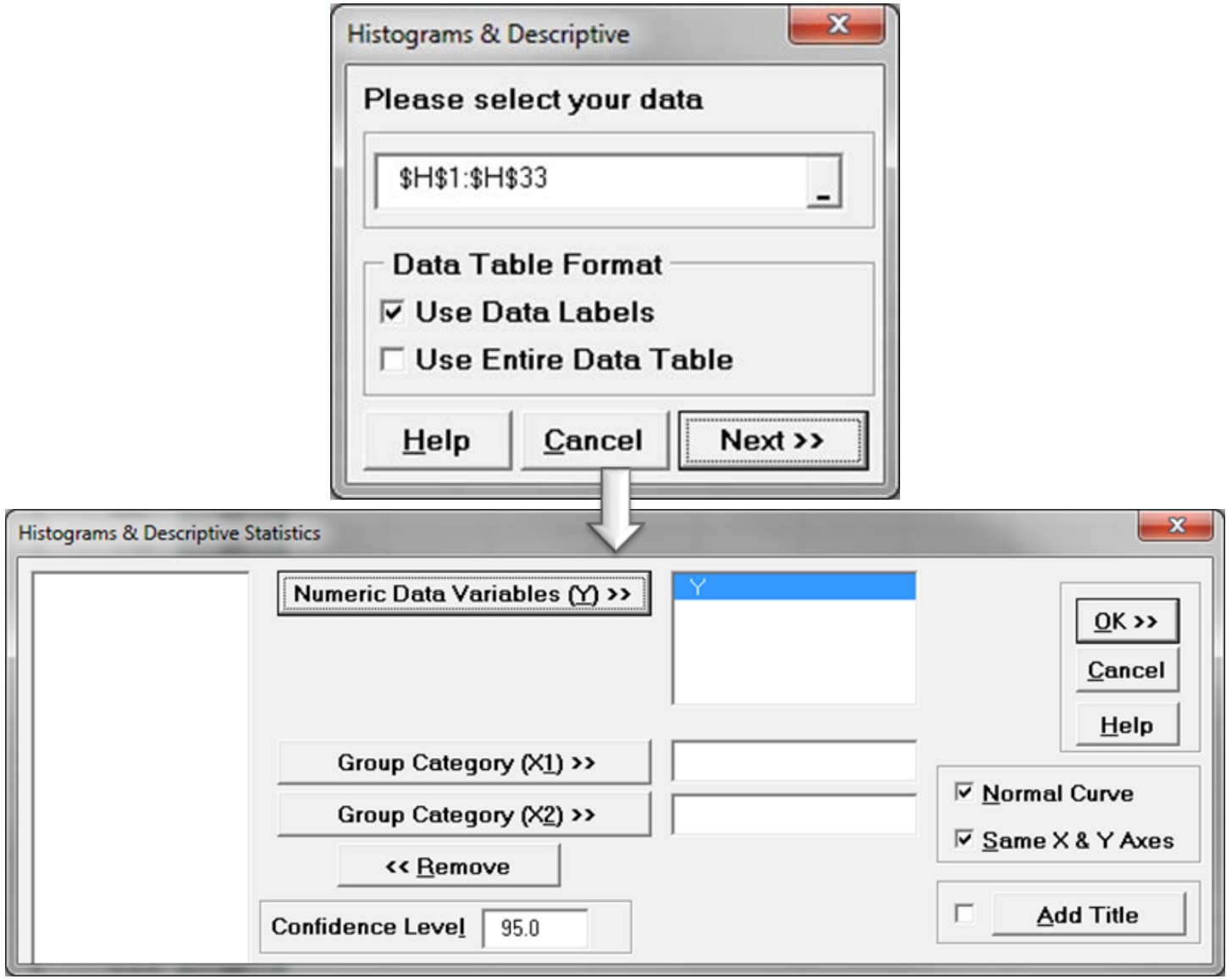
Use SigmaXL to Perform Box-Cox Transformation

- SigmaXL provides the best Box-Cox transformation with an optimal λ which minimizing the model SSE (sum of squared error).
- Here is an example of how we transform the non-normally distributed response to normal data using Box-Cox method.
- Data File: “Box-Cox” tab in “Sample Data.xlsx”

Use SigmaXL to Perform Box-Cox Transformation

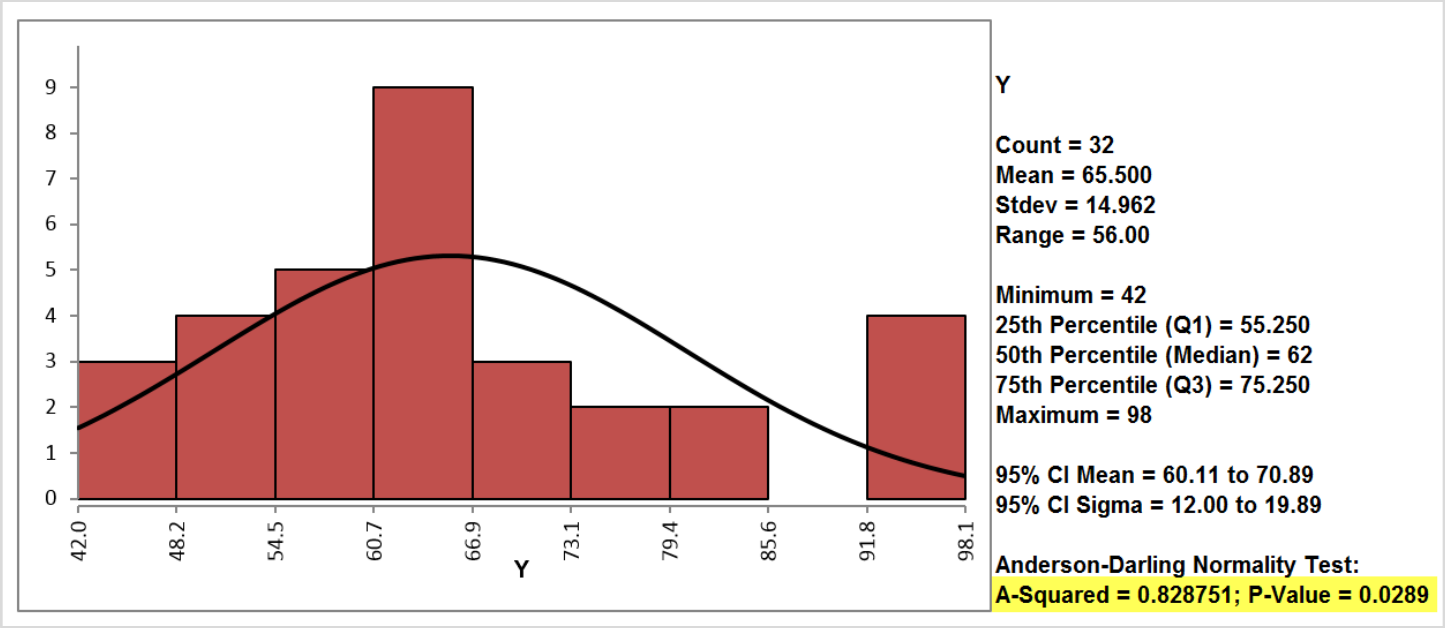
- Step 1: Test the normality of the original data set.
  - Select the entire range of “Y” in column H
  - Click SigmaXL -> Graphical Tool -> Histograms & Descriptive Statistics
  - A new window named “Histograms & Descriptive” pops up and the selected range automatically appears in the box below “Please select your data”.
  - Click “Next >>”
  - A new window named “Histograms & Descriptive Statistics” pops up.
  - Select “Y” as “Numeric Data Variables (Y)”
  - Click “OK>>”
  - The analysis results are shown automatically in the new spreadsheet “Hist Descript(1)”

Use SigmaXL to Perform Box-Cox Transformation



Use SigmaXL to Perform Box-Cox Transformation

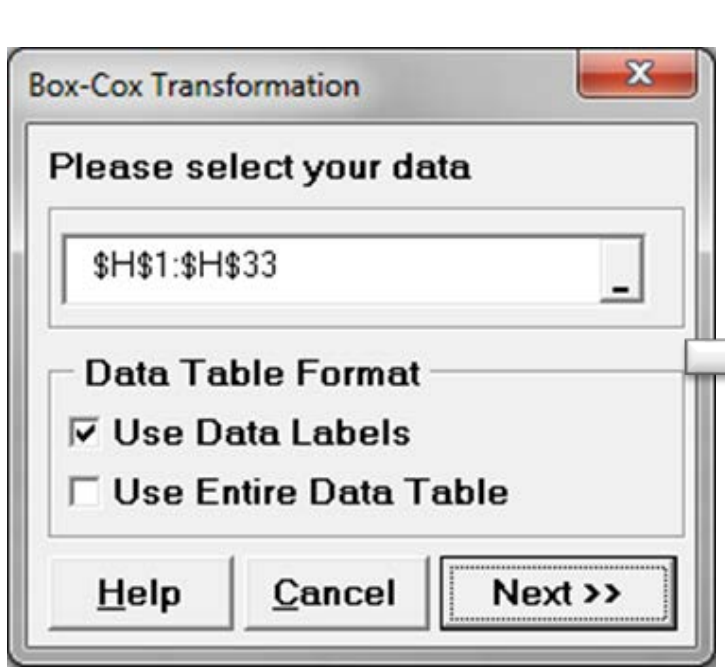
- Normality Test:
  - H0: The data is normally distributed.
  - H1: The data is not normally distributed.
- If p-value > alpha level (0.05), we fail to reject the null. Otherwise, we reject the null.
- In this example, p-value = 0.0289 < alpha level (0.05). The data is not normally distributed.

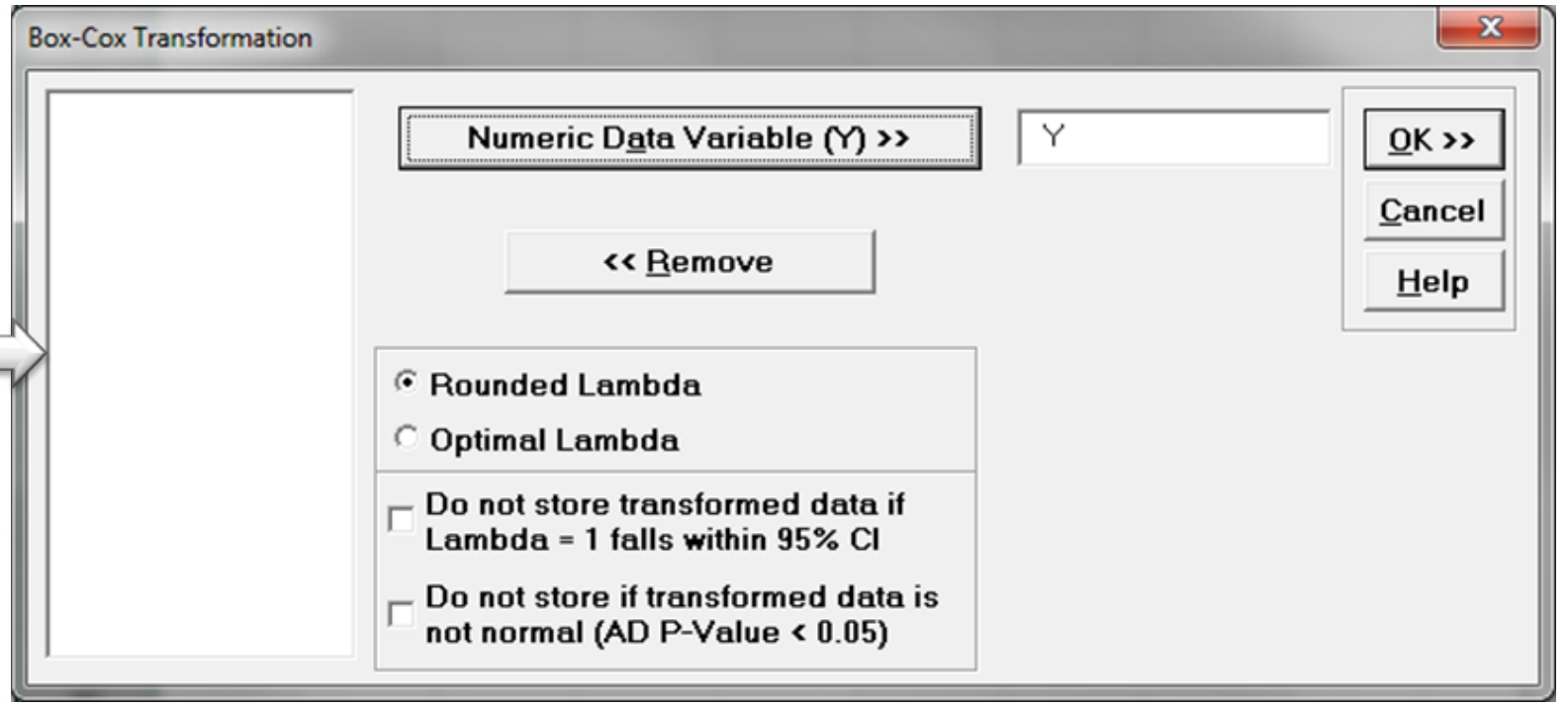


Use SigmaXL to Perform Box-Cox Transformation

- Step 2: Run the Box-Cox Transformation:
  - Select the entire range of Y in column H
  - Click SigmaXL -> Process Capability -> Nonnormal -> Box-Cox Transformation
  - A new window named “Box-Cox Transformation” pops up and the selected range appears automatically in the box under “Please select your data”
  - Click “Next >>”
  - A new window also named “Box-Cox Transformation” pops up.
  - Select “Y” as “Numeric Data Variables (Y)”
  - Click “OK>>”
  - The analysis results are shown automatically in the new spreadsheet “Box-Cox (1)”

Use SigmaXL to Perform Box-Cox Transformation

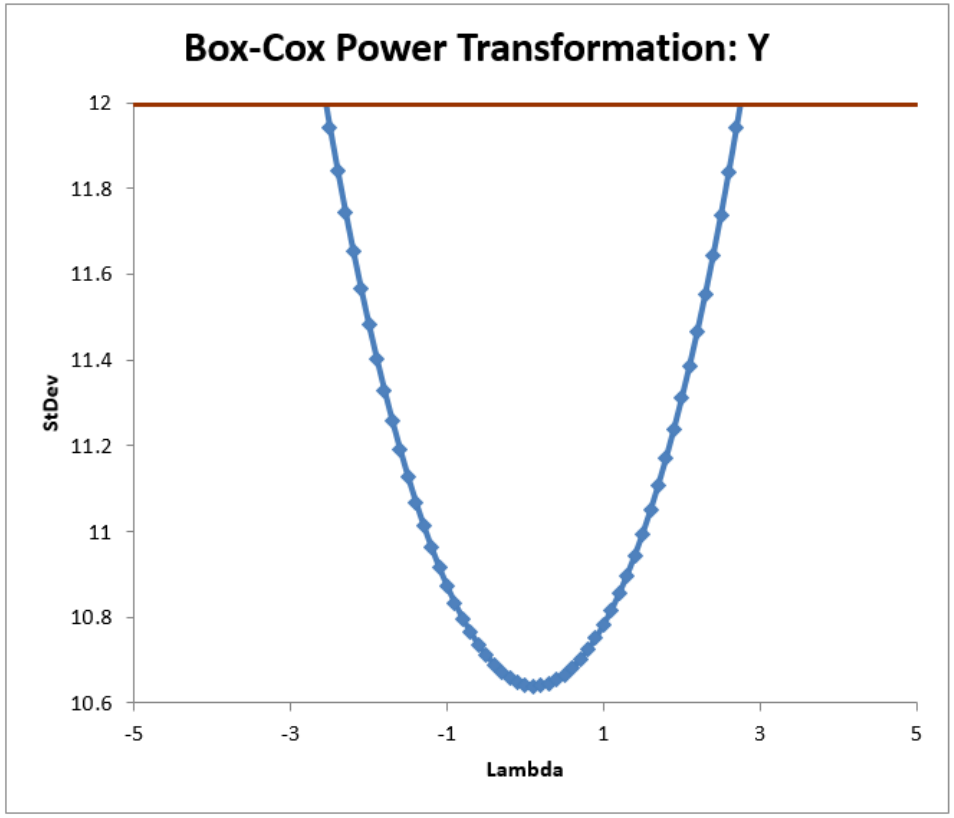




Use SigmaXL to Perform Box-Cox Transformation

Box-Cox Power Transformation: Y

Optimal Lambda	0.120000
Final (Rounded) Lambda	0
UC Lambda (95%)	2.751
LC Lamdba (95%)	-2.551
Anderson Darling Normality Test for Transformed Data:	
A-Squared	0.408443
AD P-Value	0.3273

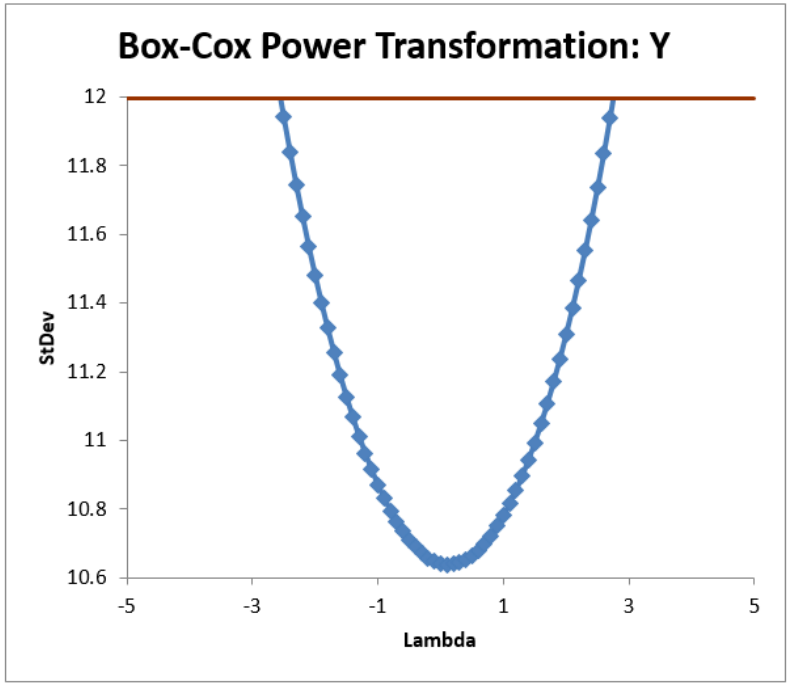


The 95% confidence interval of λ is [-2.551. 2.751] and λ=1 is outside the confidence interval which indicates that the transformation is statistically significant.

When λ is 0.12, the transformation is the best with minimum SSE.

Use SigmaXL to Perform Box-Cox Transformation

Box-Cox Power Transformation: Y	
Optimal Lambda	0.120000
Final (Rounded) Lambda	0
UC Lambda (95%)	2.751
LC Lamdba (95%)	-2.551
Anderson Darling Normality Test for Transformed Data:	
A-Squared	0.408443
AD P-Value	0.3273



Y	Transformed Data (Ln(Y))
61	4.111
53	3.970
63	4.143
61	4.111
53	3.970
56	4.025
54	3.989
61	4.111
94	4.543
93	4.533
66	4.190
60	4.094
95	4.554
98	4.585
56	4.025
63	4.143
70	4.248
65	4.174
59	4.078
55	4.007
67	4.205
65	4.174
44	3.784
45	3.807
78	4.357
77	4.344
49	3.892
42	3.738
81	4.394
82	4.407

The transformed Y is also listed in Column G in the newly generated tab “Box-Cox (1)”

Anderson Darling test is used to test the normality of the transformed data

- H0: The data is normally distributed.
- H1: The data is not normally distributed.

If p-value > alpha level (0.05), fail to reject the null. Otherwise, reject the null.

In this example, p-value = 0.3273 > alpha level (0.05). The data is normally distributed.

4.2.6 Stepwise Regression

What is Stepwise Regression?

- **Stepwise regression** is a statistical method to automatically select regression models with the best sets of predictive variables from a large set of potential variables.

- There are different statistical methods used in stepwise regression to evaluate the potential variables in the model:
  - F-test
  - T-test
  - R-square
  - AIC.

Three Approaches to Stepwise Regression

- Forward Selection
  - Bring in potential predictors one by one and keep them if they have significant impact on improving the model.
- Backward Selection
  - Try out potential predictors one by one and eliminate them if they are insignificant to improve the fit.
- Mixed Selection
  - Is a combination of both forward selection and backward selection. Add and remove variables based on pre-defined significance threshold levels.

4.2.7 Logistic Regression

What is Logistic Regression?

- **Logistic regression** is a statistical method to predict the probability of an event occurring by fitting the data to a logistic curve using logistic function.

- The dependent variable in a logistic regression can be binary (e.g., 1/0, yes/no, pass/fail), nominal (blue/yellow/green), or ordinal (satisfied/neutral/dissatisfied).

- The independent variables can be either continuous or discrete.

Logistic Function

$$f(z) = \frac{1}{1 + e^{-z}}$$

where *z* can be any value ranging from negative infinity to positive infinity.

The value of *f(z)* ranges from 0 to 1, which matches exactly the nature of probability (i.e., 0 ≤ P ≤ 1).

Logistic Regression Equation

- Based on the logistic function

$$f(z) = \frac{1}{1 + e^{-z}}$$

we define f(z) as the **probability** of an event occurring and *z* is the weighted sum of the significant predictive variables.

$$z = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

Logistic Regression Equation

- Logistic Regression Y = F(x)

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k)}}$$

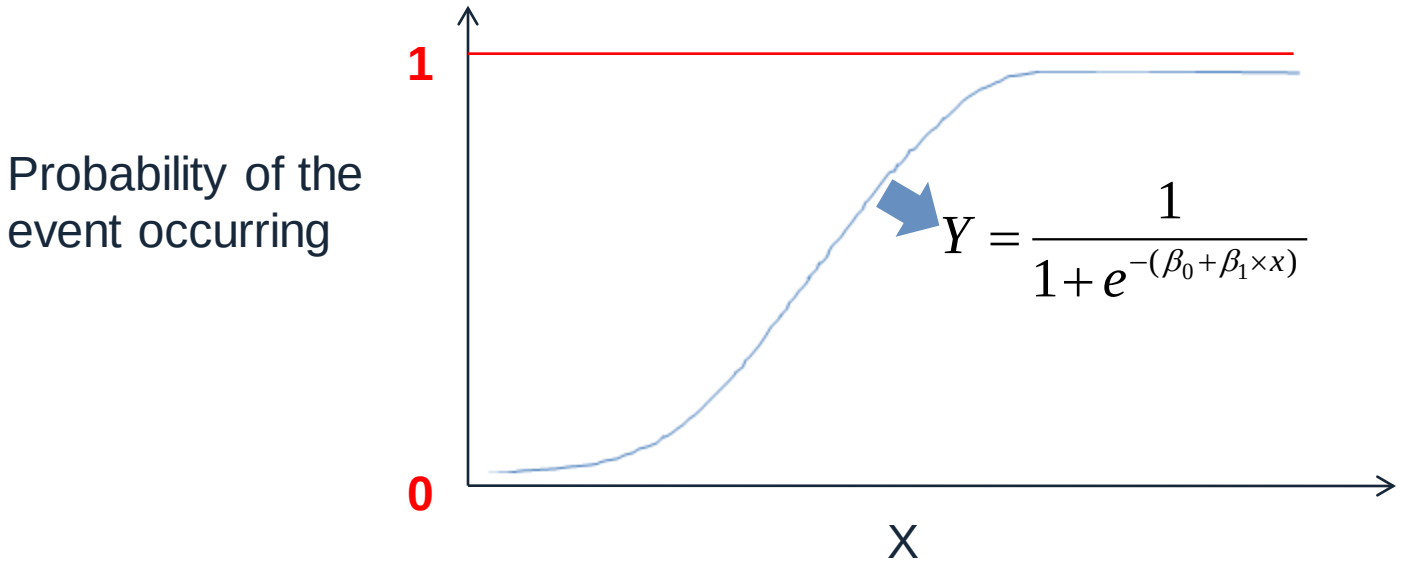
where *Y* is the probability of an event occurring and *x's* are the significant predictors.

Note:

- When building the regression model, we use the actual *Y,* which is discrete (e.g., binary, nominal, ordinal).
- After completing building the model, the fitted *Y* calculated using the logistic regression equation is the probability ranging from 0 to 1. To transfer the probability back to the discrete value, we need SMEs' inputs to select the probability cut point.

Logistic Curve

- Logistic curve for binary logistic regression with one continuous predictor:



Odds

- **Odds** is the probability of an event occurring divided by the probability of the event not occurring.

$$Odds = \frac{P}{1-P}$$

- Odds range from 0 to positive infinity.

Odds

- Probability can be calculated using odds.

$$P = \frac{odds}{1+odds} = \frac{e^{\ln(odds)}}{1+e^{\ln(odds)}} = \frac{1}{1+e^{-\ln(odds)}}$$

- Since in logistic regression model

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1\times x_1+\beta_2\times x_2+\cdots+\beta_k\times x_k)}}$$

therefore

$$\ln(odds) = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_k \times x_k$$

Three Types of Logistic Regression

- Binary Logistic Regression
  - Binary response variable
  - Example: yes/no, pass/fail, female/male
- Nominal Logistic Regression
  - Nominal response variable
  - Example: set of colors, set of countries
- Ordinal Logistic Regression
  - Ordinal response variable
  - Example: satisfied/neutral/dissatisfied
- All three logistic regression models can use multiple continuous or discrete independent variables.

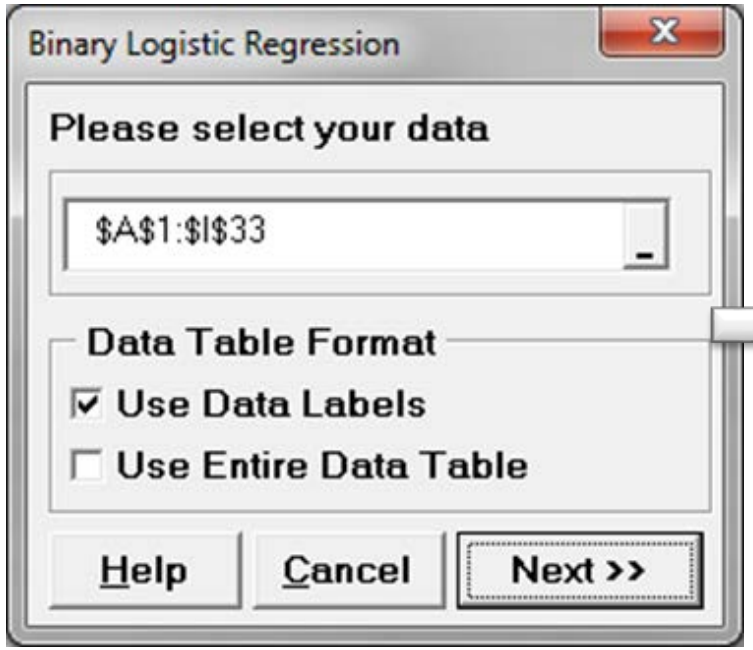
How to Run a Logistic Regression in SigmaXL

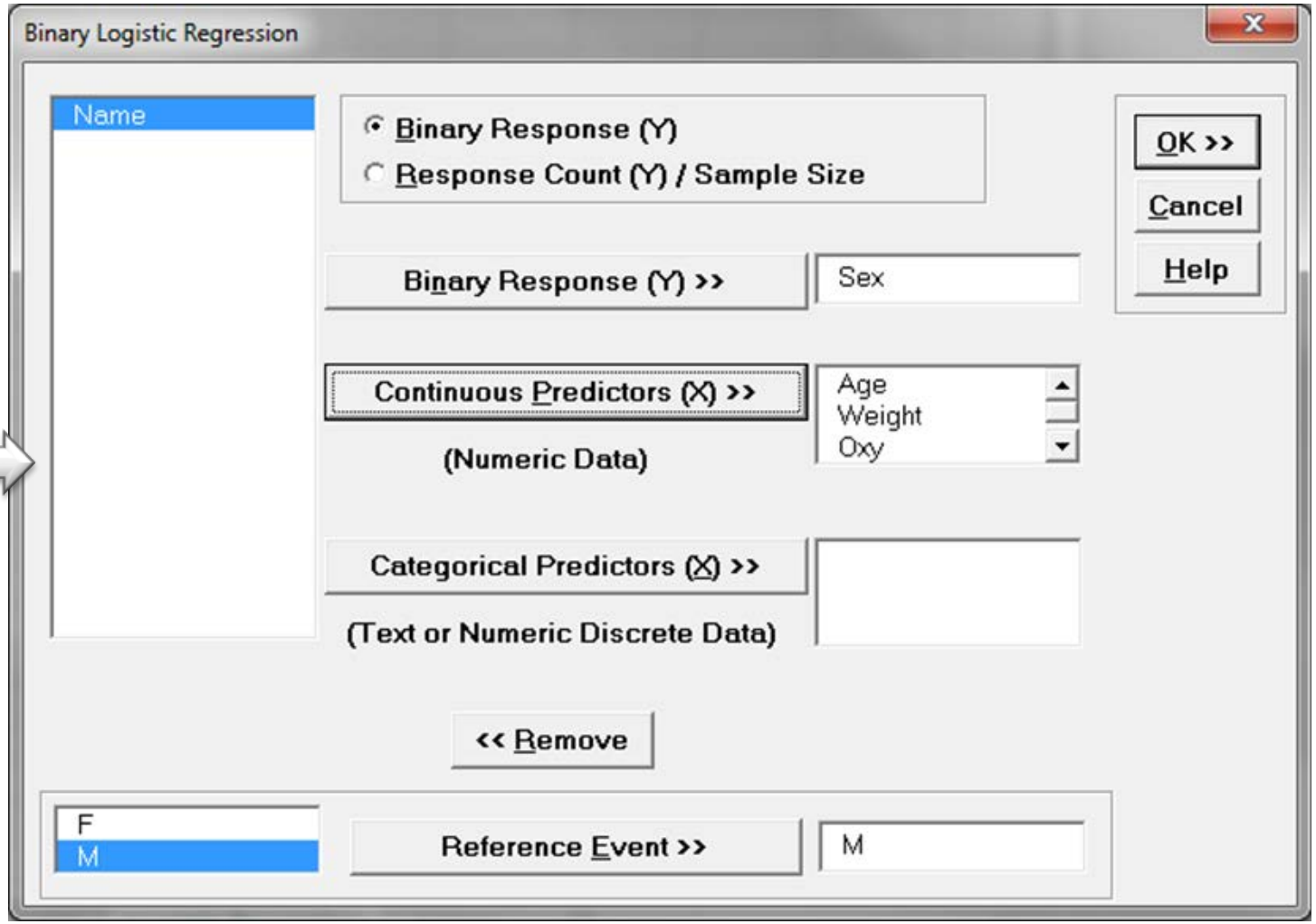
- Data File: “Logistic Regression” tab in “Sample Data.xlsx”
- Response and potential factors
  - Response (Y): Female/Male
  - Potential Factors (Xs):
    - Age
    - Weight
    - Oxy
    - Runtime
    - RunPulse
    - RstPulse
    - MaxPulse
- We want to build a logistic regression model using the potential factors to predict the probability that the person measured is female or male.

How to Run a Logistic Regression in SigmaXL

- **Step 1**:
  - Select the entire range of data (“Name”, “Sex”, “Age”, “Weight”, “Oxy”, “Runtime”, “RunPulse”, “RstPulse”, “MaxPulse” columns)
  - Click SigmaXL -> Statistical Tools -> Regression ->Binary Logistic Regression
  - A new window named “Binary Logistic Regression” appears with the selected range of data appearing in the box under “Please select your data”
  - Click “Next>>”
  - A new window also called “Binary Logistic Regression” pops up.
  - Select “Sex” as the “Binary Response (Y)”
  - Select “Age”, “Weight”, “Oxy”, “Runtime”, “RunPulse”, “RstPulse”, “MaxPulse” as the “Continuous Predictors (X)”.
  - The reference event is set as “M” by default.
  - Click “OK>>”

How to Run a Logistic Regression in SigmaXL





How to Run a Logistic Regression in SigmaXL

- Step 2:
  - Step 2.1: The results of the logistic regression model appear in the newly generated tab “Binary Logistic”.
  - Step 2.2: Check the p-values of all the independent variables in the model
  - Step 2.3: Remove the insignificant independent variable one at a time from the model and rerun the model
  - Step 2.4: Repeat step 2.1 until all of the independent variables in the model are statistically significant.

How to Run a Logistic Regression in SigmaXL

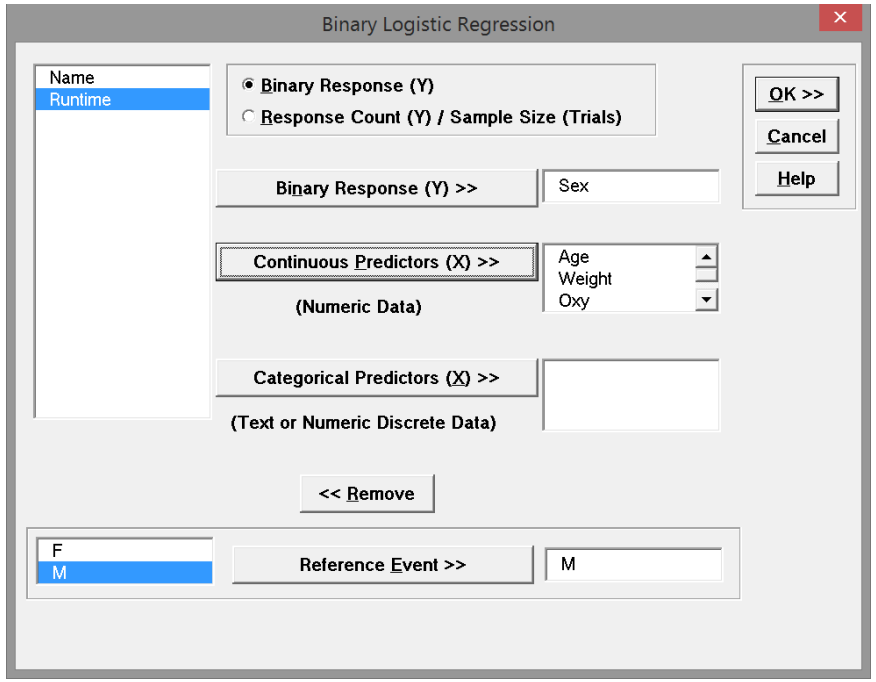
Since the p-values of all the independent variables are higher than the alpha level (0.05), we need to remove the insignificant independent variables one at a time from the model starting from the one with the highest p-value.

Runtime has the highest p-value (0.9897), so it would be removed from the model first.

Parameter Estimates:

Term	Coefficient	SE Coefficient	Z	P	Odds Ratio	Lower 95% Odds Ratio	Upper 95% Odds Ratio
Constant	-4.653	22.173	-0.209859	0.8338			
Age	0.003032381	0.105841262	0.028650276	0.9771	1.003036984	0.815123	1.234
Weight	0.064324651	0.058134689	1.106476211	0.2685	1.066438562	0.951592	1.195
Oxy	-0.229541	0.197793	-1.161	0.2458	0.794898	0.539444	1.171
Runtime	0.008537188	0.658453	0.012965516	0.9897	1.008573734	0.277473	3.666
RunPulse	-0.171215	0.139803	-1.225	0.2207	0.842640	0.640678	1.108267449
RstPulse	0.017738716	0.075834618	0.233913	0.8151	1.017896981	0.877308	1.181
MaxPulse	0.220916	0.148945	1.483	0.1380	1.247	0.931446	1.670

How to Run a Logistic Regression in SigmaXL



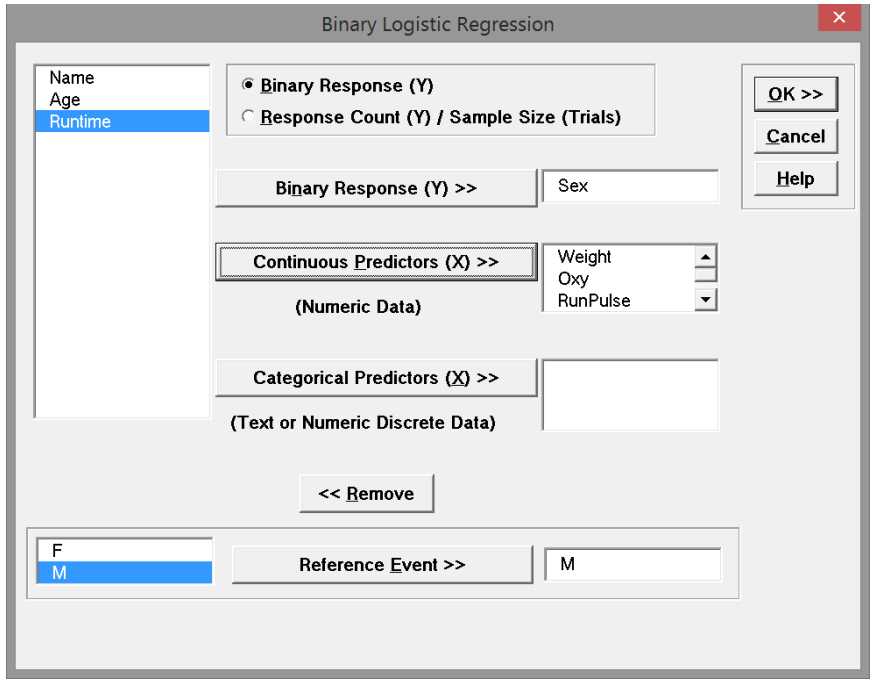
After removing Runtime from the model, the p-values of all the independent variables are still higher than the alpha level (0.05), we need to continue removing the insignificant independent variables one at a time from the model starting from the one with the highest p-value.

Age has the next highest p-value (0.9773), so it would be removed from the model next.

Parameter Estimates:

Term	Coefficient	SE Coefficient	Z	P	Odds Ratio	Lower 95% Odds Ratio	Upper 95% Odds Ratio
Constant	-4.480	17.722	-0.252805	0.8004			
Age	0.003010292	0.105858458	0.028436951	0.9773	1.003014827	0.815077	1.234
Weight	0.064309677	0.058123705	1.106427687	0.2685	1.066422594	0.951598	1.195
Oxy	-0.231330	0.141727	-1.632	0.1026	0.793477	0.601026	1.047551417
RunPulse	-0.171763	0.133301	-1.289	0.1976	0.842179	0.648540	1.093633848
RstPulse	0.018052886	0.071835081	0.251310	0.8016	1.018216824	0.884490	1.172
MaxPulse	0.221374	0.144702	1.530	0.1261	1.248	0.939655	1.657

How to Run a Logistic Regression in SigmaXL



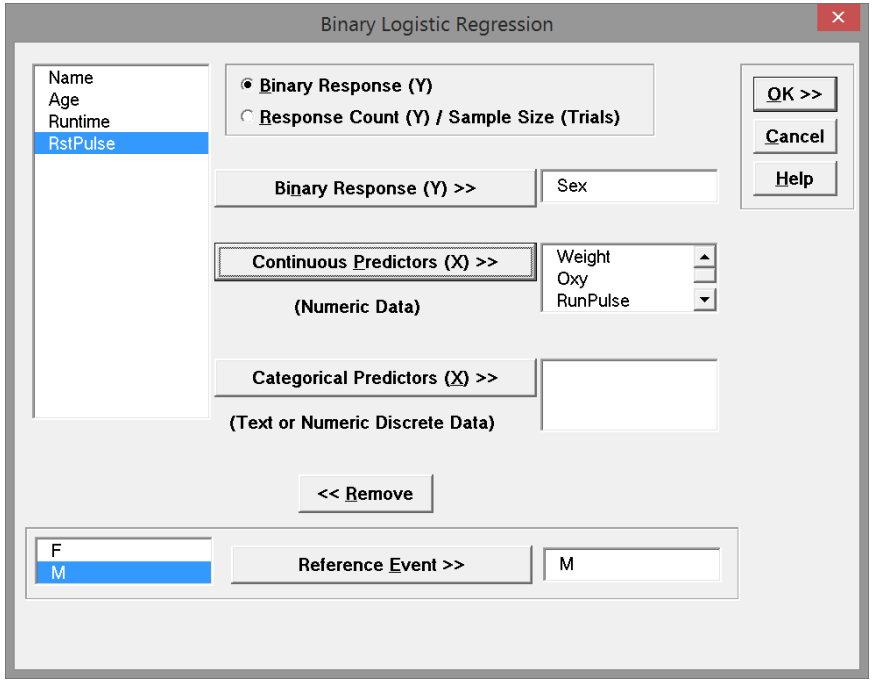
After removing Age from the model, the p-values of all the independent variables are still higher than the alpha level (0.05), we need to continue removing the insignificant independent variables one at a time from the model starting from the one with the highest p-value.

RstPulse has the next highest p-value (0.8017), so it would be removed from the model next.

Parameter Estimates:

Term	Coefficient	SE Coefficient	Z	P	Odds Ratio	Lower 95% Odds Ratio	Upper 95% Odds Ratio
Constant	-4.090	11.217	-0.364627	0.7154			
Weight	0.063902267	0.056348179	1.134	0.2568	1.065988211	0.954526	1.190
Oxy	-0.233057	0.128318	-1.816	0.0693	0.792109	0.615968	1.018617112
RunPulse	-0.171849	0.133279	-1.289	0.1973	0.842106	0.648512	1.093492905
RstPulse	0.017600868	0.070080466	0.251152	0.8017	1.017756676	0.887136	1.168
MaxPulse	0.220833	0.143495	1.539	0.1238	1.247	0.941371	1.652

How to Run a Logistic Regression in SigmaXL



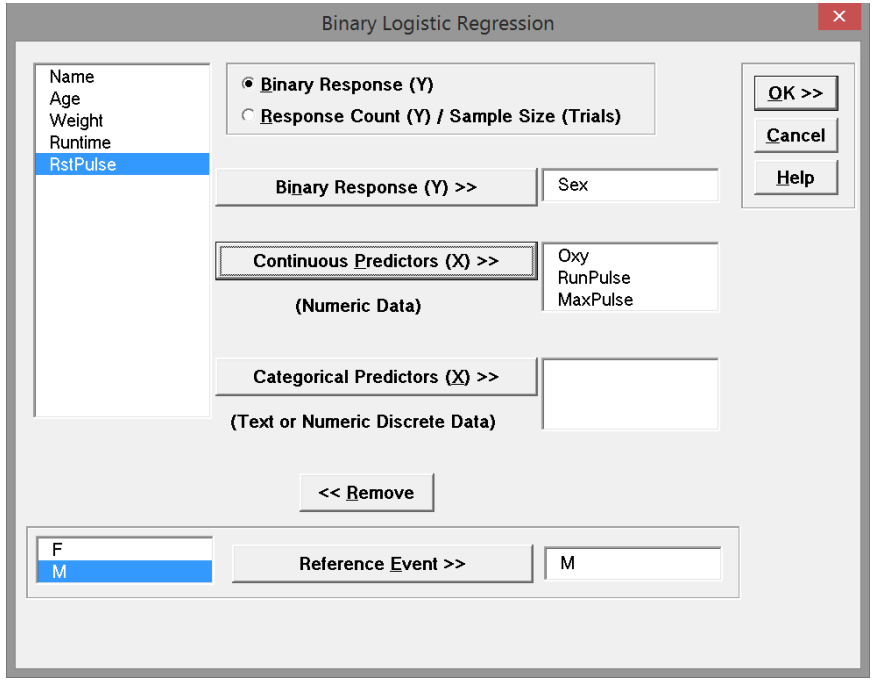
After removing RstPulse from the model, the p-values, Oxy shows significance with the other independent variables still higher than the alpha level (0.05), we need to continue removing them one at a time from the model.

Now, Weight is the next highest p-value (0.242), so we will remove from the model next.

Parameter Estimates:

Term	Coefficient	SE Coefficient	Z	P	Odds Ratio	Lower 95% Odds Ratio	Upper 95% Odds Ratio
Constant	-3.378	10.817	-0.312274	0.7548			
Weight	0.065405499	0.055900255	1.170	0.2420	1.067591844	0.956802	1.191
Oxy	-0.244530	0.122720	-1.993	**0.0463**	0.783072	0.615659	0.996010
RunPulse	-0.173619	0.134068	-1.295	0.1953	0.840617	0.646363	1.0932501
MaxPulse	0.226328	0.143883	1.573	0.1157	1.254	0.945840	1.663

How to Run a Logistic Regression in SigmaXL



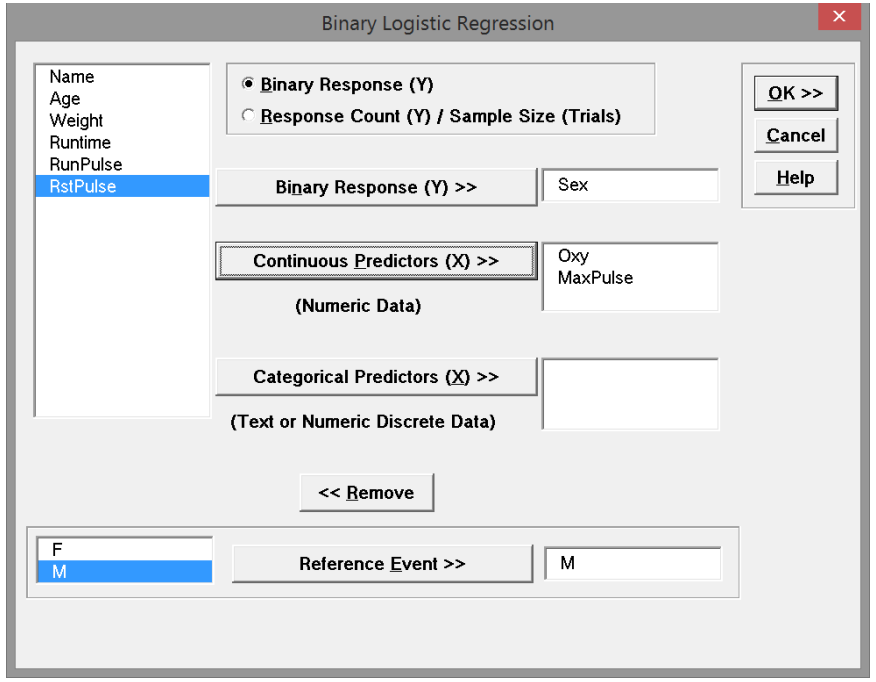
After removing Weight from the model, we continue to have p-values higher than the alpha level (0.05), we need to continue removing them.

At this point, RunPulse has the highest p-value (0.1604), so it need to be removed next.

Parameter Estimates:

Term	Coefficient	SE Coefficient	Z	P	Odds Ratio	Lower 95% Odds Ratio	Upper 95% Odds Ratio
Constant	1.945	9.364	0.207648	0.8355			
Oxy	-0.255986	0.122018	-2.098	0.0359	0.774153	0.609484	0.983311
RunPulse	-0.186392	0.132791	-1.404	0.1604	0.829948	0.639760	1.076674936
MaxPulse	0.240825	0.141608	1.701	0.0890	1.272	0.963940	1.679

How to Run a Logistic Regression in SigmaXL



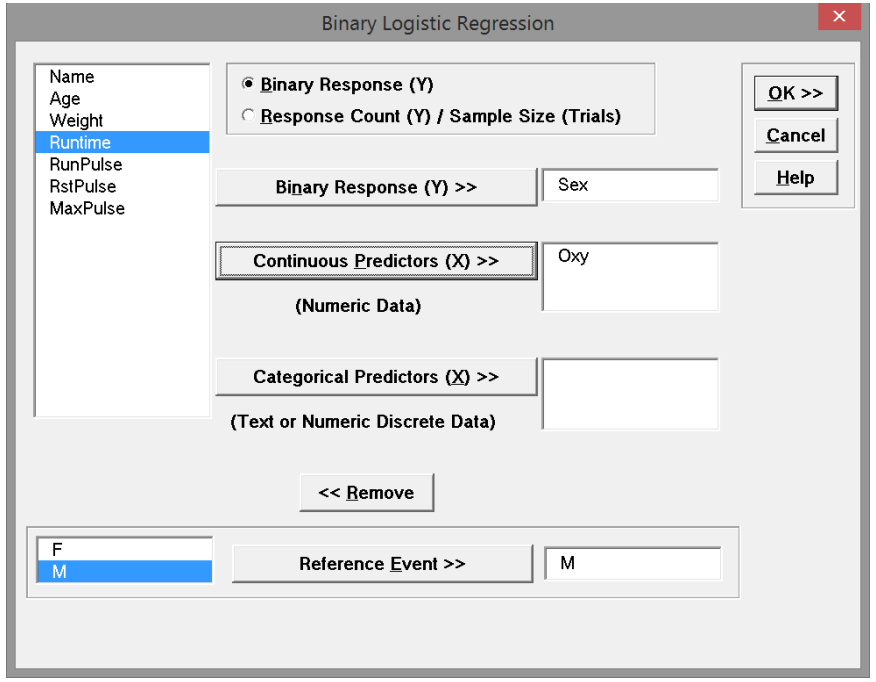
After removing RunPulse from the model, the p-values of all the independent variables are still higher than the alpha level (0.05), we need to continue removing the insignificant independent variables.

Now, MaxPulse has the highest p-value (0..229), so it would be removed from the model next.

Parameter Estimates:

Term	Coefficient	SE Coefficient	Z	P	Odds Ratio	Lower 95% Odds Ratio	Upper 95% Odds Ratio
Constant	-2.803	8.902	-0.314907	0.7528			
Oxy	-0.145686	0.078620423	-1.853	0.0639	0.864429	0.740979	1.008445067
MaxPulse	0.056051349	0.046593695	1.203	0.2290	1.057651992	0.965343	1.159

How to Run a Logistic Regression in SigmaXL



After removing Weight from the model, the p-value of the only remaining independent variable "Oxy" is at the alpha level (0.05). There is no need to remove "Oxy" from the model, we will accept the minute risk of rejecting the null at this p-value (0.0556). But before we do that, let's check the validity of the model as a whole..

Parameter Estimates:

Term	Coefficient	SE Coefficient	Z	P	Odds Ratio	Lower 95% Odds Ratio	Upper 95% Odds Ratio
Constant	-7.372	3.858	-1.911	0.0560			
Oxy	0.154725	0.080909694	1.912	**0.0558**	1.167	0.996149	1.368

How to Run a Logistic Regression in SigmaXL

Model Summary and Goodness-of-Fit Statistics:	
Log-Likelihood	-19.840
Test that all slope coefficients are equal to zero:	
Likelihood Ratio Chi-Square (G)	4.681
DF	1
P-Value	0.0305
McFadden's Pseudo R-Square	10.55%
Goodness-of-Fit Tests (P-Value < .05 indicates Lack-of-Fit):	
Pearson Residuals Chi-Square	28.268
DF	29
P-Value	0.5036
Deviance Residuals Chi-Square	36.908
DF	29
P-Value	0.1486
Hosmer-Lemeshow Chi-Square	6.217
DF	6
P-Value	0.3994
Measures of Association	
Concordant	180
Discordant	75
Ties	1
Total	256
Concordant Percent	70.31
Discordant Percent	29.30
Ties Percent	0.39
Goodman-Kruskal Gamma	0.411765
Somers' D	0.410156
Kendall's Tau-a	0.211694

- Step 3:
  - Analyze the binary logistic report and check the performance of the logistic regression model
  - The p-value here is smaller than alpha level (0.05). We conclude that at least one of the slope coefficients is not equal to zero.
  - The pseudo R-squared is 10.55%. The R-squared of logistic regression is in general lower than the R-squared of the traditional multiple linear regression model.
  - The p-value of lack of fit test is higher than alpha level (0.05). We conclude that the model fits the data

How to Run a Logistic Regression in SigmaXL

Observed and Predicted Outcomes:

Observed Outcome	Predicted Outcome		Row Total
	Y = F	Y = M	
Y = F	9	7	16
Y = M	5	11	16
Column Total	14	18	32
Percent Correctly Predicted:	62.50%		

62.50% of the predicted outcomes match the observed outcomes.

How to Run a Logistic Regression in SigmaXL

- Step 4:
  - Enter the setting of the Oxy into the cell highlighted in yellow and the predicted event probability would appear automatically. In this case, if we set the oxy value to 50, the probability that the person measured being male is 59%.

Response Event Probability:

Predictors	Enter Settings:	Predicted Event Probability
Oxy		

Response Event Probability:

Predictors	Enter Settings:	Predicted Event Probability
Oxy	50	0.590112421

4.3 Designed Experiments

Black Belt Training: Improve Phase

4.1 Simple Linear Regression
- 4.1.1 Correlation
- 4.1.2 X-Y Diagram
- 4.1.3 Regression Equations
- 4.1.4 Residuals Analysis

4.2 Multiple Regression Analysis
- 4.2.1 Non-Linear Regression
- 4.2.2 Multiple Linear Regression
- 4.2.3 Confidence Intervals
- 4.2.4 Residuals Analysis
- 4.2.5 Data Transformation, Box Cox
- 4.2.6 Stepwise Regression
- 4.2.7 Logistic Regression

4.3 Designed Experiments
- 4.3.1 Experiment Objectives
- 4.3.2 Experimental Methods
- 4.3.3 DOE Design Considerations

4.4 Full Factorial Experiments
- 4.4.1 2k Full Factorial Designs
- 4.4.2 Linear and Quadratic Models
- 4.4.3 Balanced and Orthogonal Designs
- 4.4.4 Fit, Model, and Center Points

4.5 Fractional Factorial Experiments
- 4.5.1 Designs
- 4.5.2 Confounding Effects
- 4.5.3 Experimental Resolution

4.3.1 Experiment Objectives

What is an Experiment?

- An **experiment** is a scientific exercise to gather data to test a hypothesis, theory, or previous results.

- Experiments are planned studies in which data are collected actively and purposefully.

- A typical experiment follows this sequence:
  - A problem arises.
  - A hypothesis is stated.
  - An experiment is designed and implemented to collect data.
  - The analysis is performed to test the hypothesis.
  - Conclusions are drawn based on the analysis results.

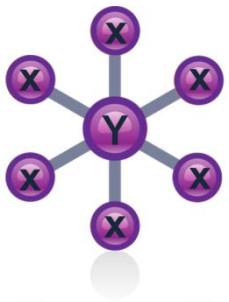
Why Use Experiments?

- Resolve Problems
  - Eliminate defects or defectives.
  - Shift performance means to meet customer expectations.
  - Squeeze variation, making process more predictable.
- Optimize Performance
  - Use statistical methods to get desired process response.
  - Minimize undesirable conditions, waste, or costs.
  - Reduce variation to eliminate out-of-spec conditions.
- Intelligent Design
  - Design processes around variables that have significant impact on process outputs.
  - Design processes around variables that are easier or more cost effective to manage.

Traditional Methods of Learning

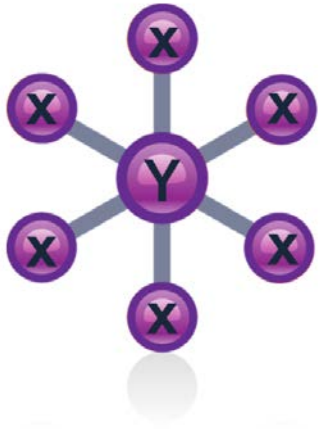
- Passive Learning
  - Study or observe events while they occur.
  - Study or analyze events after they occur.

- OFAT Experiment (One Factor at a Time)
  - An experimental style limited to controlling one factor independently while others remain either static or are allowed to vary.



- Design of Experiments
  - Purposefully and proactively manipulate variables so their effect on the dependent variable can be studied.
  - Encourages an informative event to occur within specific planned parameters.

What is the Design of Experiment (DOE)?



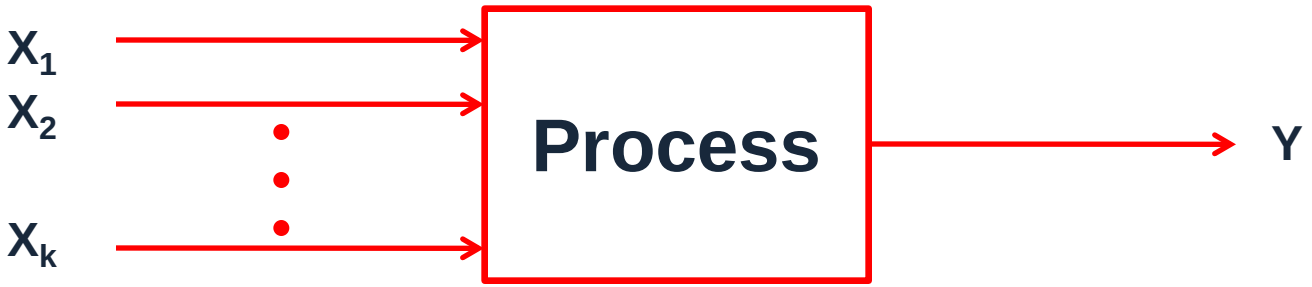
- **Design of experiment** (DOE) is a systematic and cost effective approach to plan the exercises necessary to collect data and quantify the cause-and-effect relationship between the response and different factors.

- In DOE, we observe the changes occurring in the output (i.e., response, dependent variable) of a process by changing one or multiple inputs (i.e., factors, independent variables).
- Using DOE we are able to manipulate several factors with every run and determine their effect on the “Y” or output.

Common DOE Terminology

- **Response**: Y, dependent variable, process output measurement
- **Effect**: The change in the average response across two levels of a factor or between experimental conditions.
- **Factor:** X's, inputs, independent variables
- **Fixed Factor**: Factor that can be controlled during the study
- **Random Factor**: Factor that cannot be controlled during the study
- **Level**: factor settings, usually high and low, + and -, 1 and -1
- **Treatment Combination**: setting of all factors to obtain one response measurement (also referred to as a "run")
- **Replication**: Running the same treatment combination more than once (sequence of runs is typically randomized)
- **Repeat**: Non-randomized replicate of all treatment combinations
- **Inference Space**: Operating range of factors under study

Y = f (Xs)



- In Design of Experiments (DOEs)
  - X_1, X_2, …, X_k are the inputs of the process. They can be either continuous or discrete variables.
  - Y is the output of the process. Y is a continuous variable.
  - A single X or a group of Xs potentially have statistically significant impact on the Y with different levels of influence.

Note: DOE for discrete Y is not covered in this module.

Objectives of Experiments

- Active Learning:
  - Learn as much as possible with as little resources as possible. Efficiency is the name of the game.
- Identify Critical X's:
  - Identify the critical factors that drive the output or dependent variable ("Y").
- Quantify Relationships:
  - Generate an equation that characterizes the relationship between the X's and the Y.
- Optimize:
  - Determine the required settings of all the X's to achieve the optimal output or response.
- Validate:
  - Prove results through confirmation or pilot implementation before fully implementing solutions.

Principle of Parsimony

- The ultimate goal of DOE is to use the minimum amount of data to discover the maximum amount of information.
- The experiment is designed to obtain data as parsimoniously as possible to draw as much information about the relationship between the response and factors as possible.
- DOE is often used when the resources to collect the data and build the model are limited.

Tradeoffs

- The objective of a specific DOE study should be determined by the team.
- The more complex the objective is, the more data are required. As a result, more time and money are needed to collect the data.
- By prioritizing the objectives and filtering factors which obviously do not have any effect on the response, a considerable amount of cost of running the DOE would be reduced.

4.3.2 Experimental Methods

Planning a DOE

1. **Problem Statement**: Quantifiably define the problem
2. **Objective**: State the objective of the experiment
3. **Primary Metric**: Determine the response variable or “Y”
4. **Key Process Input Variables (KPIV)**: Select the input variable(s) or X’s
5. **Factor Levels**: Determine level settings for all factors
6. **Design**: Select the experimental design for your DOE
7. **Project Plan**: Prepare a project plan for your experiment – human and capital resource allocation, timing, duration etc.
8. **Gain Buy-in**: Your experiment will need support and buy-in from several places, get it!
9. **Run the experiment**
10. **Analyze, Interpret, and Share Results:**

DOE Problem Statement

- Much like the overall problem statement discussed in the Define phase, a DOE problem statement should specific and have quantifiable elements.

- Make sure your problem statement ties to a business goal or business performance indicator.

- Problem statements should *not* include:
  - Conclusions
  - Solutions
  - Causes.

DOE Objectives

State the Objective

- DOE Objective Examples:
  - “To determine the effects of tire pressure on gas mileage”
  - “To determine which factors have the most impact on gas mileage”
  - “To determine the effects of a pick-up truck’s tailgate (up vs. down) on the truck’s gas mileage”
  - “To determine which HVAC and Lighting factors are contributing most to Kwh consumption”
- DOE objective statements should:
  - Clearly define what you want to learn from the experiment
  - Declare study parameters or scope.

DOE Primary Metric

- Recall our discussion of primary metrics in the Define phase?
  - We said something to this effect: “the primary metric is the most important measure of success.”
- This statement still holds true when discussing the primary metric for a DOE.
- The primary metric for a DOE is your experiment’s output or response under study.
- It is the “Y” in your transfer function $Y = f(x)$.
- It is critical to understand the following things about your metric:
  - Measurement accuracy, R&R, and/or bias
  - Measurement frequency
  - Measurement resolution.

DOE Primary Metric

- Before embarking on a design of experiment you should have validated your measurement system.
- If you have not, now would be a wise time to back track and do so.

- DOEs can be costly on a firm's time and resources.
- DOEs not planned and run properly will create waste and defective products because treatment combinations should "test" boundaries.
- Do not allow your entire study to be wasted because of a poor measurement system or one that can not respond frequently enough for your test.
- Measurement resolution is another key factor to consider for your DOE. Make sure your primary metric can decipher the size of the main effects and interactions you are looking for.

DOE Primary Metric

- There is still even more to understand about your primary metric before jumping into a DOE.
- Be sure to answer these questions before proceeding:

1. Is it discrete or continuous?
2. Do you want to shift its mean?
3. Do you want to squeeze variation?
4. What is the baseline?
5. Is it in control?
6. Is there seasonality, trending, or any cycle?
7. How much change do you want to detect?
8. Is your metric normally distributed?

DOE Input Variables – KPIVs (key process input variables)

- Factor selection is an important element in planning for a design of experiment.
- Factors are the “things” or metrics that your study responds to.
- Factors can be discrete or continuous.
- Factors should have largely been determined through the use of tools and analytics used throughout the DMAIC roadmap.
  - Failure modes and effects analysis
  - Cause and effect diagram (fishbone)
  - X-Y matrix
  - Process mapping
  - Planned studies
  - Passive analytics
  - etc.

DOE Factor Settings

- **Factor settings** are the range of values selected for each factor.
- Factor settings need to be predetermined before every experiment.
  - Let us use this paper airplane as an example. Our factors under study for max distance as a response might be:
    1. Number of folds
    2. Paper weight
    3. Paperclip (yes or no)
    4. Thrust
    5. Launch height
    6. Launch angle etc.

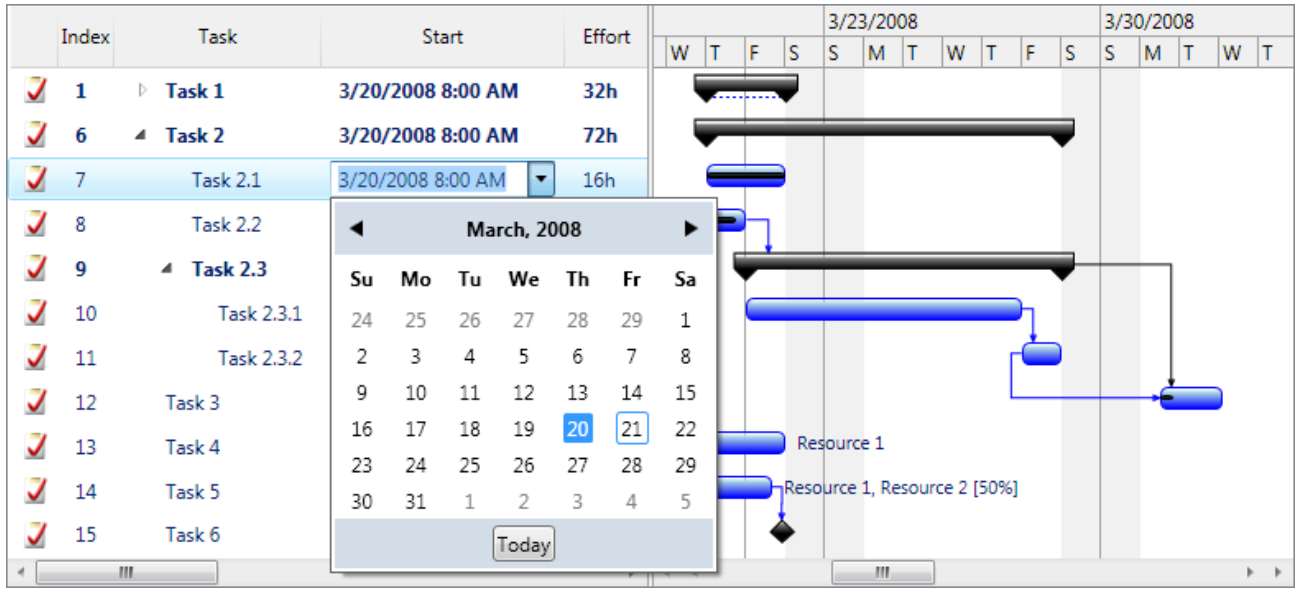
  - Factor settings are important because each setting selection should provide value to the study.
    - Understanding the “margin” where that value can be found is where an experienced Black Belt will earn their keep.
    - For example, what would we learn if we chose the test two levels for “number of folds” and we tested zero folds and 80 folds as our two settings for that factor? Right, nothing!

DOE – Design Considerations

- **Design selection** is an aspect of properly planned design of experiments.
- High-level overview of design types (you will learn more about these in later modules)
  - Screening Designs:
    - Fractional factorials
    - Intended to narrow down the many factors to the vital few
    - Screening designs enable many factors to be evaluated at a lower cost because of the nature of the design (main effects)
  - Characterization Designs:
    - Higher resolution fractional factorials
    - Full factorials with fewer variables
    - Main effect and interactions
  - Optimization Designs:
    - Full factorials
    - Response surface designs.

DOE Project Plan

- DOEs can require a fair amount of coordination between human and capital resources.
- Preparing a DOE project plan is a wise step.
- Make sure you identify critical dependencies like interfering with production or operating hours etc.
- Use project management software or, at a minimum, a Gantt chart to help you identify *critical path items*.



DOE – Getting Support

- By running a DOE you are about to "discover" something new and better or validate an existing belief or assumption. *(If you do not believe this you should not be running your DOE)*
- You are also about to cost your firm money or disrupt someone' s routine…Get everyone on the same page before moving forward!

- Hopefully, you have already established the framework for the following items. If not, follow these steps:
  1. **Business Case and Justification**
  2. **Stakeholder Analysis** – figure out who is with you and your project and who is not
  3. **Communication Plan** – build and execute your communication plan based on the first two items.

Run Your Experiment

Analyze, Interpret, and Share Your DOE Results

DOE Iterations

- One single DOE might not be enough to completely answer the questions due to the following reasons:
  - None of the potential factors identified are statistically significant to the response.
  - More factors need to be included in the model.
  - Residuals of the model are not performing well.
  - The objectives of DOE changes due to business reasons.

4.3.3 Experiment Design Considerations

Considerations of Experiments

- Objectives of the experiments
- Resource availability to run the experiments
- Potential costs to implement the experiments
- Stakeholders' support for successful experiments
- Accuracy and precision of the measurement system
- Statistical stability of the process
- Unexpected plans or changes
- Sequence of running the experiments
- Simplicity of the model
- Inclusion of potential significant factors
- Settings of factors
- Well-behaved residuals

4.4 Full Factorial Experiments

Black Belt Training: Improve Phase

4.1 Simple Linear Regression
- 4.1.1 Correlation
- 4.1.2 X-Y Diagram
- 4.1.3 Regression Equations
- 4.1.4 Residuals Analysis

4.2 Multiple Regression Analysis
- 4.2.1 Non-Linear Regression
- 4.2.2 Multiple Linear Regression
- 4.2.3 Confidence Intervals
- 4.2.4 Residuals Analysis
- 4.2.5 Data Transformation, Box Cox
- 4.2.6 Stepwise Regression
- 4.2.7 Logistic Regression

4.3 Designed Experiments
- 4.3.1 Experiment Objectives
- 4.3.2 Experimental Methods
- 4.3.3 DOE Design Considerations

4.4 Full Factorial Experiments
- 4.4.1 2k Full Factorial Designs
- 4.4.2 Linear and Quadratic Models
- 4.4.3 Balanced and Orthogonal Designs
- 4.4.4 Fit, Model, and Center Points

4.5 Fractional Factorial Experiments
- 4.5.1 Designs
- 4.5.2 Confounding Effects
- 4.5.3 Experimental Resolution

4.4.1 2k Full Factorial Designs

DOE Key Terms

- **Response**: Y, dependent variable, process output measurement
- **Effect**: The change in the average response across two levels of a factor or between experimental conditions
- **Factor:** X's, inputs, independent variables
- **Fixed Factor**: Factor that can be controlled during the study
- **Random Factor**: Factor that cannot be controlled during the study
- **Factor Levels**: Factor settings, usually high and low, + and -, 1 and -1
- **Treatment Combination**: setting of all factors to obtain one response measurement (also referred to as a "run")
- **Replication**: Running the same treatment combination more than once (sequence of runs is typically randomized)
- **Repeat**: Non-randomized replicate of all treatment combinations
- **Inference Space**: Operating range of factors under study
- **Main Effect**: The average change from one level setting to another for a single factor
- **Interaction**: The combined effect of two factors independent of the main effect of each factor

More About Factors

- **Factors** in DOE are the potential causes that might have significant impact on the outcome of a process or a product.
- Factors are the inputs of a process or system and the *response* is the output of it.
- Factors can be continuous or discrete variables.
- Most experiments use two or more factors. Ask subject matter experts' opinions for main factors selection. The more factors in DOE, the more costs triggered.

More About Factors

- The goal of a DOE is to measure the effects of factors and interactions of factors on the outcome.

- In DOE:
  - First, we create a set of different combinations of factors each at different levels.
  - Second, we run the experiment and collect the response values in different scenarios.
  - Third, we analyze the results and test how the response changes accordingly when factors change.

Factor Levels

- Factor levels are the selected settings of a factor we are testing in the experiment.
- The more levels a factor has, the more scenarios we have to create and test. As a result, more resources are required to collect samples and the model is more complicated.
- The most popular DOE is two-level design, meaning only two levels for each factor.

Factor Levels

- Code of factor levels
  - High vs. Low
  - (+1) vs. (-1)
  - (+) vs. (-)
- Example of factor levels
  - A study on how two factors affect the taste of cakes.

	Factor	Settings	Code
Factor 1 (A)	Temperature of the oven	375 degrees 350 degrees	+1 -1
Factor 2 (B)	Time length of baking	30 minutes 25 minutes	+1 -1

Treatment Combination

- A **treatment** is a combination of different factors at different levels.
- It is a unique scenario setting in an experiment.
- Coding treatments:
  - If there are two factors in an experiment, we name one factor A and the other factor B.
  - A single treatment in an experiment is named for each factor:
    - --, +-, -+ and ++
    - l, a, b and ab

Treatment Combination

- Example of treatment
  - A study on how two factors affect the taste of cakes.

Treatment	Factors	
	Temperature of the oven	Time length of baking
I	350	25
a	375	25
b	350	30
ab	375	30

Treatment	Factors	
	Factor A	Factor B
I	-1	-1
a	+1	-1
b	-1	+1
ab	+1	+1

Response

- A **response** is the output of a process, a system, or a product.
- It is the outcome we are interested to improve by optimizing the settings of factors.
- In an experiment, we observe, record, and analyze the corresponding changes occurring in the response after changing the settings of different factors.

Response

- Example of treatment
  - A study on how two factors affect the taste of cakes. We use a predefined metric to measure the cake's tastiness.
  - After running each treatment, we obtain and record the resulting response value.

Treatment	Factors		Response
	Temperature of the oven	Time length of baking	
I	350	25	20
a	375	25	18
b	350	30	22
ab	375	30	12

Treatment	Factors		Response
	Factor A	Factor B	
I	-1	-1	20
a	+1	-1	18
b	-1	+1	22
ab	+1	+1	12

Main Effect

- **Main effect** is the average change in the response resulting from the change in the levels of a factor.

- It captures the effect of a factor alone on the response without taking into consideration other factors.

- In the baking cake example, the main effects of temperature and time length are

$$MainEffect_A = \frac{18+12}{2} - \frac{20+22}{2} = -6$$

$$MainEffect_B = \frac{22+12}{2} - \frac{20+18}{2} = -2$$

Main Effect

- Interpretation of main effect in the example:
  - By changing the temperature level of the oven from high to low, the response (i.e., tastiness of the cake) increases by 6 measurement units.
  - By changing the time length of baking from high to low, the response (i.e., tastiness of the cake) increases by 2 measurement units.

Interaction Effect

- **Interaction effect** is the average change in the response resulting from the change in the interaction of multiple factors.

- It occurs when the change in the response triggered by one factor depends on the change in other factor(s).

Interaction Effect

- Example of interaction effect

Treatment	Factors		Interaction (A*B)	Response
	Factor A	Factor B		
I	-1	-1	+1	20
a	+1	-1	-1	18
b	-1	+1	-1	22
ab	+1	+1	+1	12

One interaction

$$InteractionEffect = \frac{20+12}{2} - \frac{22+18}{2} = -4$$

- By changing the interaction from high to low, the response (i.e., tastiness of the cake) increases by 4 measurement units.

2^k Full Factorial DOE

- In a **full factorial experiment**, all of the possible combinations of factors and levels are created and tested.

- For example, for two-level design (i.e., each factor has two levels) with k factors, there are 2^k possible scenarios or treatments.
  - 2 factors, each with 2 levels, we have $2^2 = 4$ treatments
  - 3 factors, each with 2 levels, we have $2^3 = 8$ treatments
  - k factors, each with 2 levels, we have 2^k treatments

2^k Full Factorial DOE

- **Full factorial DOE** is used to discover the cause-and-effect relationship between the response and both individual factors and the interaction of factors.

- Generate an equation to describe the relationship between Y and the important Xs:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_1 X_2 \cdots X_k + \varepsilon$$

where
Y is the response and $X_1, X_2, \ldots, X_k$ are the factors.
α_0 is the intercept and $\alpha_1, \alpha_2, \ldots, \alpha_p$ are the coefficients of the factors and interactions.
ε is the error of the model.

Two-Level Two-Factor Full Factorial

- Below is a design pattern of a two-level **two-factor** full factorial experiment:

Run	Treatment	Factors	
		A	B
1	I	-1	-1
2	a	+1	-1
3	b	-1	+1
4	ab	+1	+1

Two-Level Three-Factor Full Factorial

- Below is a design pattern of a two-level **three-factor** full factorial experiment

Run	Treatment	Factors		
		A	B	C
1	I	-1	-1	-1
2	a	+1	-1	-1
3	b	-1	+1	-1
4	ab	+1	+1	-1
5	c	-1	-1	+1
6	ac	+1	-1	+1
7	bc	-1	+1	+1
8	abc	+1	+1	+1

Two-Level Four-Factor Full Factorial

- Below is a design pattern of a two-level **four-factor** full factorial experiment

Run	Treatment	Factors			
		A	B	C	D
1	I	-1	-1	-1	-1
2	a	+1	-1	-1	-1
3	b	-1	+1	-1	-1
4	ab	+1	+1	-1	-1
5	c	-1	-1	+1	-1
6	ac	+1	-1	+1	-1
7	bc	-1	+1	+1	-1
8	abc	+1	+1	+1	-1
9	d	-1	-1	-1	+1
10	ad	+1	-1	-1	+1
11	bd	-1	+1	-1	+1
12	abd	+1	+1	-1	+1
13	cd	-1	-1	+1	+1
14	acd	+1	-1	+1	+1
15	bcd	-1	+1	+1	+1
16	abcd	+1	+1	+1	+1

Two-Level Five-Factor Full Factorial

Run	Treatment	Factors				
		A	B	C	D	E
1	l	-1	-1	-1	-1	-1
2	a	+1	-1	-1	-1	-1
3	b	-1	+1	-1	-1	-1
4	ab	+1	+1	-1	-1	-1
5	c	-1	-1	+1	-1	-1
6	ac	+1	-1	+1	-1	-1
7	bc	-1	+1	+1	-1	-1
8	abc	+1	+1	+1	-1	-1
9	d	-1	-1	-1	+1	-1
10	ad	+1	-1	-1	+1	-1
11	bd	-1	+1	-1	+1	-1
12	abd	+1	+1	-1	+1	-1
13	cd	-1	-1	+1	+1	-1
14	acd	+1	-1	+1	+1	-1
15	bcd	-1	+1	+1	+1	-1
16	abcd	+1	+1	+1	+1	-1
17	e	-1	-1	-1	-1	+1
18	ae	+1	-1	-1	-1	+1
19	be	-1	+1	-1	-1	+1
20	abe	+1	+1	-1	-1	+1
21	ce	-1	-1	+1	-1	+1
22	ace	+1	-1	+1	-1	+1
23	bce	-1	+1	+1	-1	+1
24	abce	+1	+1	+1	-1	+1
25	de	-1	-1	-1	+1	+1
26	ade	+1	-1	-1	+1	+1
27	bde	-1	+1	-1	+1	+1
28	abde	+1	+1	-1	+1	+1
29	cde	-1	-1	+1	+1	+1
30	acde	+1	-1	+1	+1	+1
31	bcde	-1	+1	+1	+1	+1
32	abcde	+1	+1	+1	+1	+1

- At left is a design pattern of a two-level five-factor full factorial experiment

Order to Run Experiments

- The four design patterns shown earlier are listed in the standard order.
- *Standard order* is used to design the combinations/treatments before experiments start.
- When actually running the experiments, *randomizing* the standard order is recommended to minimize the noise.

Replication in Experiments

- Each treatment can be tested multiple times in an experiment in order to increase the degrees of freedom and improve the capability of analysis. We call this method **replication**.

- **Replicates** are the number of repetitions of running an individual treatment.

- The order to run the treatments in an experiment should be randomized to minimize the noise.

2^2 Full Factorial DOE

- *Case study:* we are running a 2^2 full factorial DOE to discover the cause-and-effect relationship between the cake tastiness and two factors: temperature of the oven and time length of baking.
- Each factor has two levels and there are four treatments in total.
- We decide to run each treatment twice so that we have enough degrees of freedom to measure the impact of two factors and the interaction between two factors. Therefore, there are eight observations in response eventually.

	Factor	Settings	Code
Factor 1 (A)	Temperature of the oven	375 degrees 350 degrees	+1 -1
Factor 2 (B)	Time length of baking	30 minutes 25 minutes	+1 -1

2^2 Full Factorial DOE

- After running the four treatments twice in a random order, we obtain the following results

Treatment	Factors		Interaction (A*B)	Response
	Factor A	Factor B		
I	-1	-1	+1	20
a	+1	-1	-1	18
b	-1	+1	-1	22
ab	+1	+1	+1	12
I	-1	-1	+1	21
a	+1	-1	-1	17
b	-1	+1	-1	22
ab	+1	+1	+1	13

2^2 Full Factorial DOE

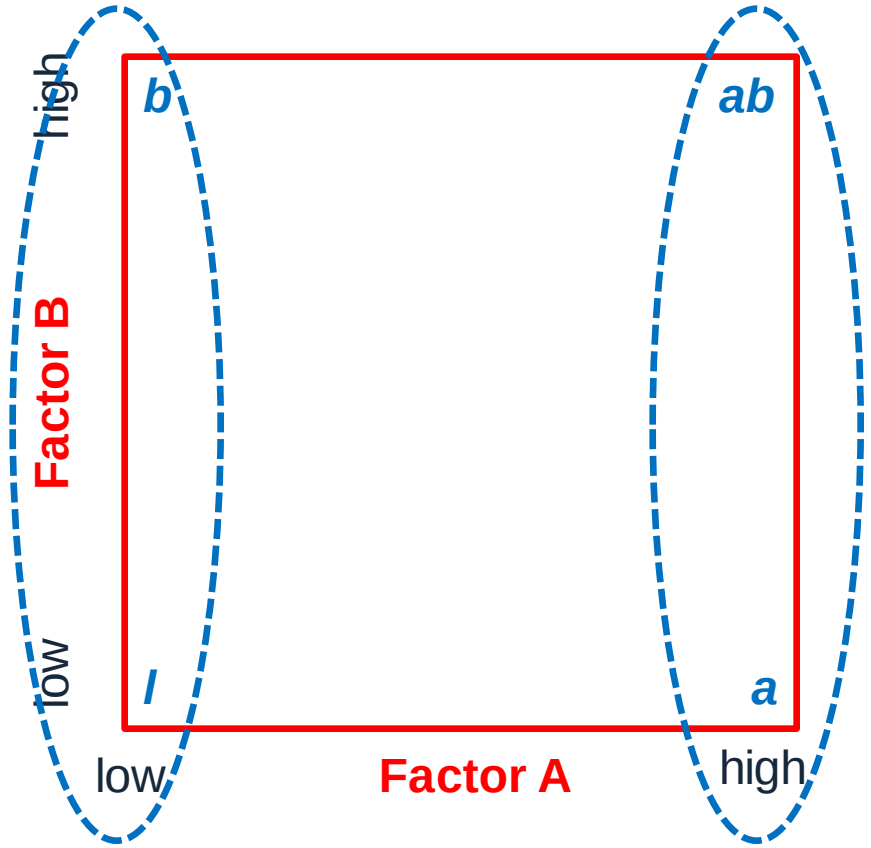
- The experiment results are consolidated into the following table

Treatment	Factors		Interaction	Response		
	Factor A	Factor B	(A*B)	Run 1	Run 2	Total
I	-1	-1	+1	20	21	41
a	+1	-1	-1	18	17	35
b	-1	+1	-1	22	22	44
ab	+1	+1	+1	12	13	25

2^2 Full Factorial DOE

- Main effect of factor A (temperature of the oven):

$$MainEffect_A = \frac{(a+ab)-(b+l)}{2^{k-1}r}$$



$$= \frac{(35+25)-(44+41)}{2^{2-1}\times 2}$$

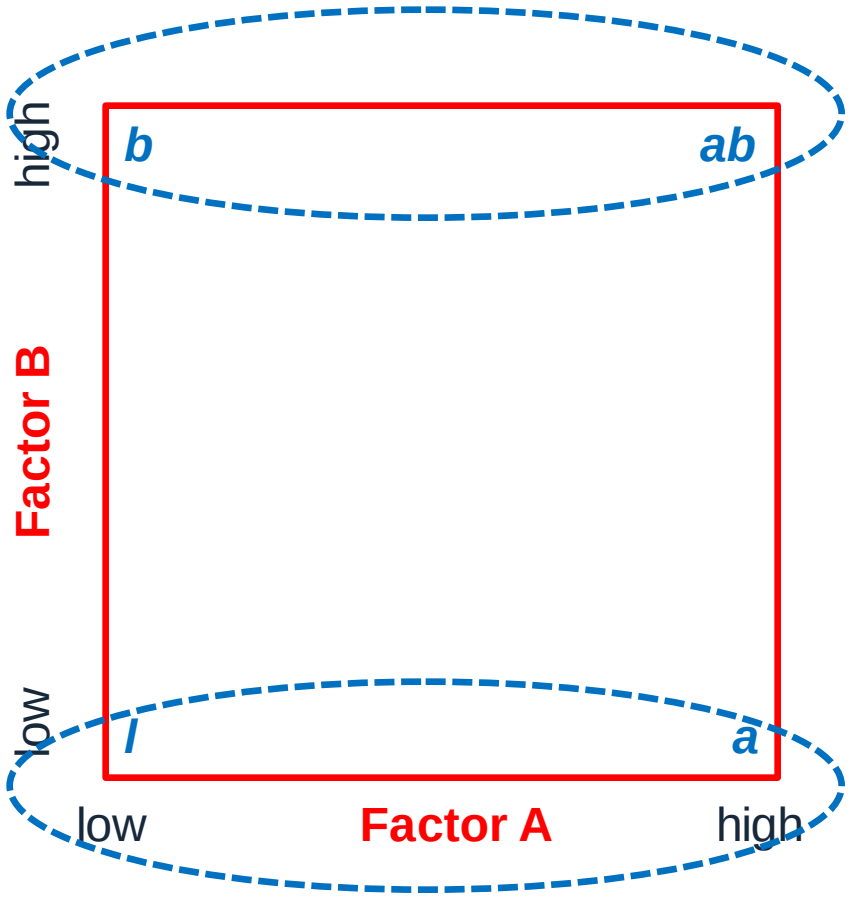
$$= -6.25$$

where
k is the number of factors,
r is the number of times individual treatments are being run.

2^2 Full Factorial DOE

- Main effect of factor B (time length of baking):

$$MainEffect_B = \frac{(b + ab) - (a + l)}{2^{k-1} r}$$



$$= \frac{(44 + 25) - (35 + 41)}{2^{2-1} \times 2}$$

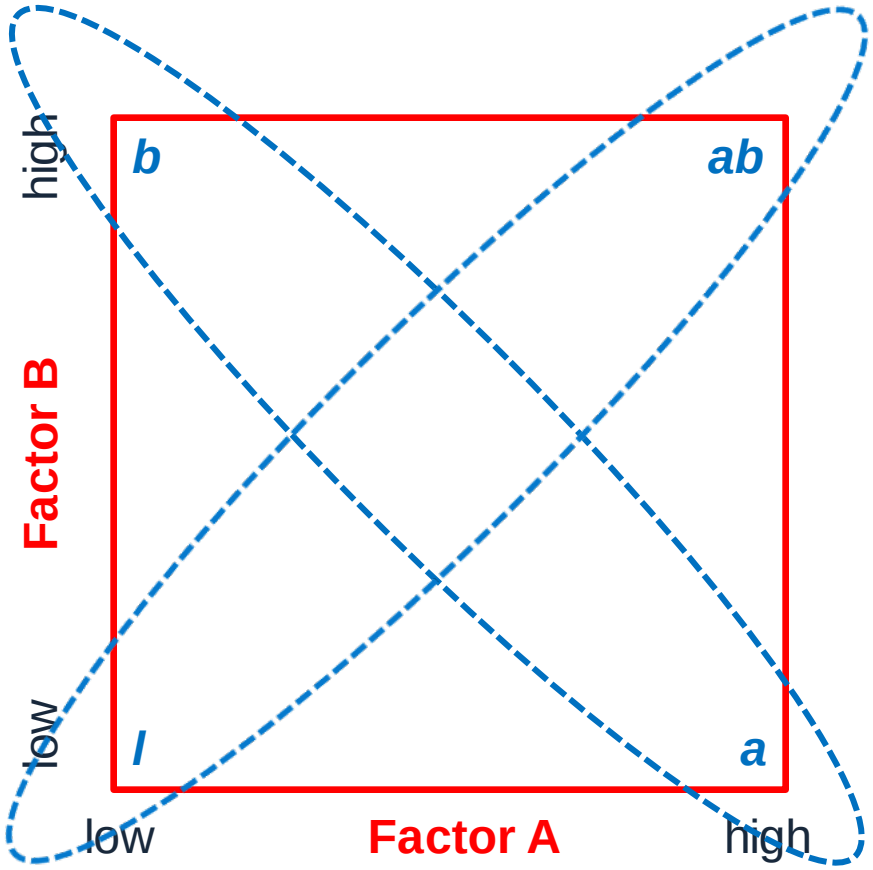
$$= -1.75$$

where
k is the number of factors,
r is the number of times individual treatments are being run.

2^2 Full Factorial DOE

- Interaction (i.e., A*B) effect

$$InteractionEffect = \frac{(l + ab) - (a + b)}{2^{k-1} r}$$



$$= \frac{(41 + 25) - (35 + 44)}{2^{2-1} \times 2}$$

$$= -3.25$$

where
k is the number of factors,
r is the number of times individual treatments are being run.

2^2 Full Factorial DOE

- Sum of squares of factors and interaction

$$SS_A = \frac{(a + ab - b - l)^2}{2^k r} = \frac{(35 + 25 - 44 - 41)^2}{2^2 \times 2} = 78.125$$

$$SS_B = \frac{(b + ab - a - l)^2}{2^k r} = \frac{(44 + 25 - 35 - 41)^2}{2^2 \times 2} = 6.125$$

$$SS_{Interaction} = \frac{(l + ab - a - b)^2}{2^k r} = \frac{(41 + 25 - 35 - 44)^2}{2^2 \times 2} = 21.125$$

where
k is the number of factors,
r is the number of times individual treatments are being run.

2^2 Full Factorial DOE

- **Degrees of freedom** of factors and interaction

$$df_A = 1$$

$$df_B = 1$$

$$df_{Interaction} = 1$$

$$df_{error} = 1$$

- Four degrees of freedom are required in the model because there are three independent variables (a, b, ab interaction) in the equation and one degree of freedom is required for the error:

$$Y = \alpha_0 + \alpha_1 * A + \alpha_2 * B + \alpha_3 * A * B + Error$$

2^2 Full Factorial DOE

- Mean squares of factors and interaction

$$MS_A = \frac{SS_A}{df_A} = \frac{78.125}{1} = 78.125$$

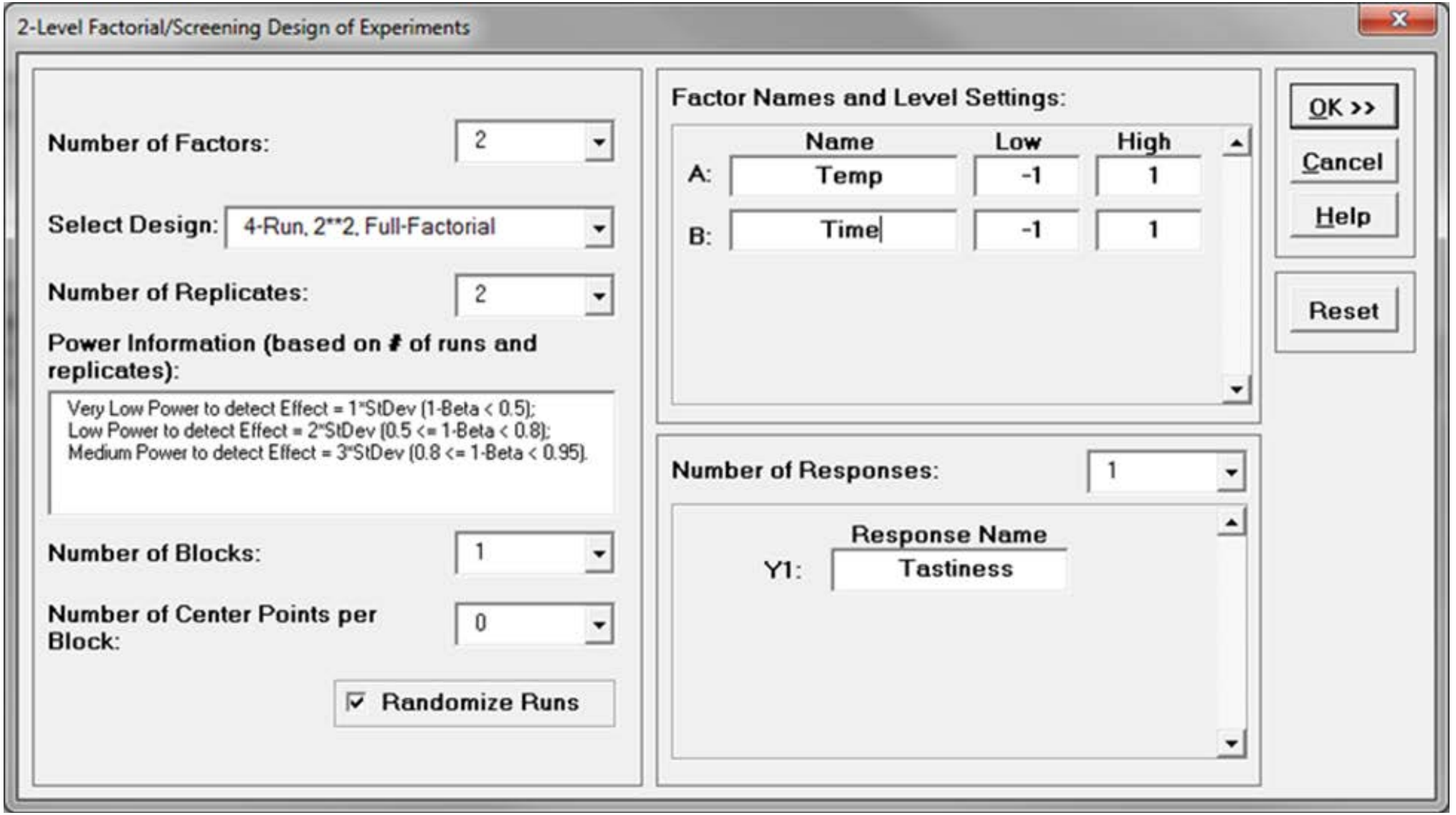
$$MS_B = \frac{SS_B}{df_B} = \frac{6.125}{1} = 6.125$$

$$MS_{Interaction} = \frac{SS_{Interaction}}{df_{Interaction}} = \frac{21.125}{1} = 21.125$$

Use SigmaXL to Run a 2^k Full Factorial DOE

- Step 1: Initiate the experiment design
  - Click SigmaXL -> Design of Experiments -> 2-Level Factorial/Screening -> 2-Level Factorial/Screening Designs
  - A new window named “2-Level Factorial/Screening Design of Experiments” pops up.
- Step 2: Enter the response.
  - Select “1” as the Number of Response.
  - Enter “Tastiness” into the “Response Name” box.
- Step 3: Enter the factors and make the design
  - Select “2” as the Number of Factors
  - Select “4-Run, 2**2, Full-Factorial” as the design
  - Select “2” as the Number of Replicates
  - Enter “Temperature” as the name for factor A
  - Enter “Time Length” as the name for factor B
  - Click “OK>>”
  - The 2^2 full factorial DOE template appears in the newly generated tab “2 Factor DOE”.

Use SigmaXL to Run a 2^k Full Factorial DOE



Use SigmaXL to Run a 2^k Full Factorial DOE

Design of Experiments Worksheet

Title:	
Date:	
Name of Experimenter:	
Notes:	

Design Type:	2 Factor, 4-Run, 2**2, Full-Factorial
Number of Replicates:	2
Number of Blocks:	1
Number of Center Points per Block:	0
Number of Responses:	1

Run Order	Std. Order	Center Points	Blocks	A: Temperature	B: Time Length	Tastiness
1	5	1	1	-1	-1	
2	8	1	1	1	1	
3	4	1	1	1	1	
4	2	1	1	1	-1	
5	6	1	1	1	-1	
6	3	1	1	-1	1	
7	7	1	1	-1	1	
8	1	1	1	-1	-1	

Use SigmaXL to Run a 2^k Full Factorial DOE

- Step 4: Run the experiment and record the response in the table created by SigmaXL

Design of Experiments Worksheet

Title:	
Date:	
Name of Experimenter:	
Notes:	

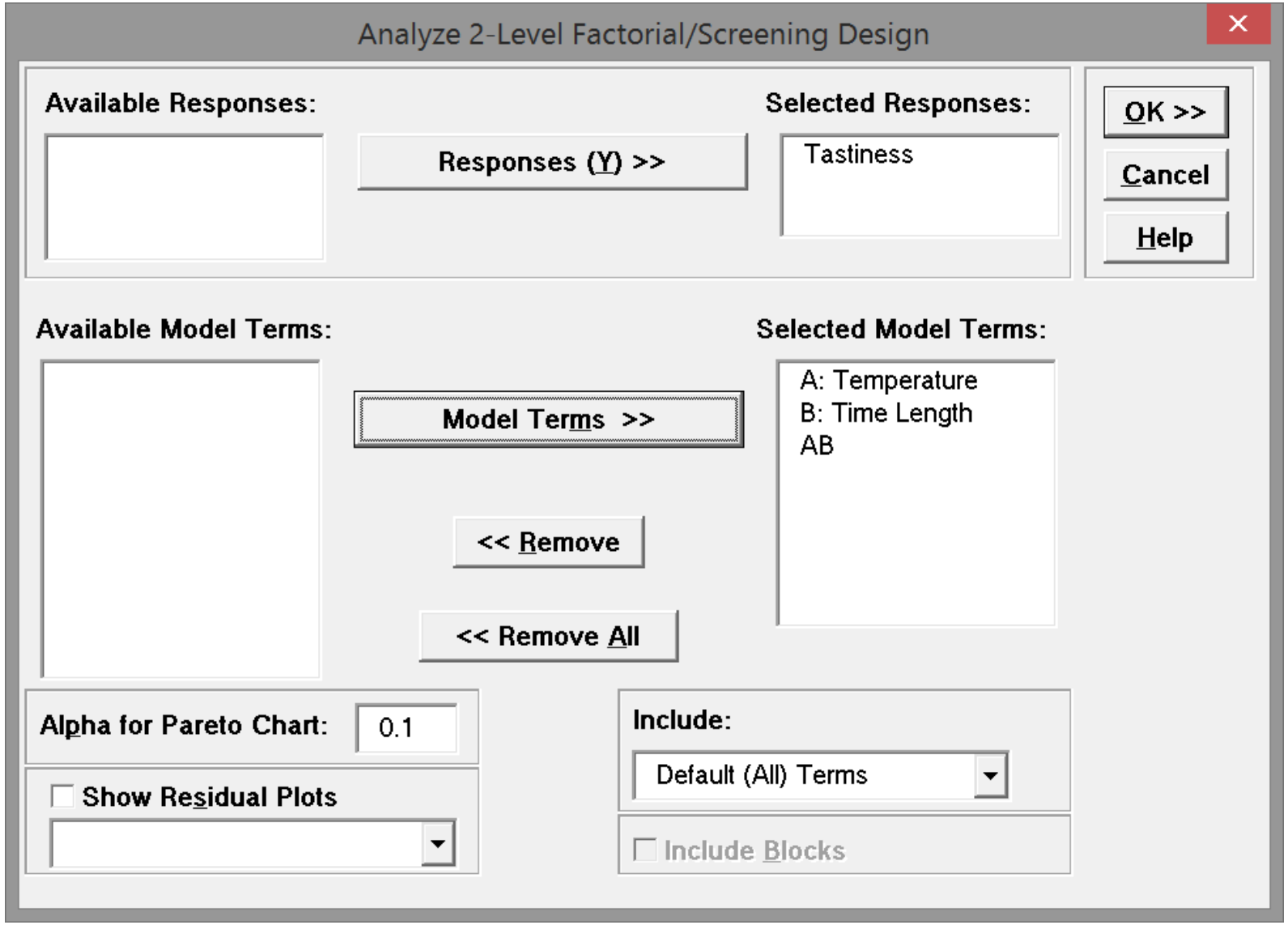
Design Type:	2 Factor, 4-Run, 2**2, Full-Factorial
Number of Replicates:	2
Number of Blocks:	1
Number of Center Points per Block:	0
Number of Responses:	1

Run Order	Std. Order	Center Points	Blocks	A: Temperature	B: Time Length	Tastiness
1	5	1	1	-1	-1	20
2	8	1	1	1	1	12
3	4	1	1	1	1	13
4	2	1	1	1	-1	22
5	6	1	1	1	-1	22
6	3	1	1	-1	1	17
7	7	1	1	-1	1	18
8	1	1	1	-1	-1	21

Use SigmaXL to Run a 2^k Full Factorial DOE

- Step 5: Analyze the experiment results
  - Click SigmaXL -> Design of Experiments -> 2-Level Factorial/Screening -> Analyze 2-Level Factorial/Screening Design
  - A new window named “Analyze 2-Level Factorial/Screening Design” appears in which “Tastiness” is automatically selected as the response variable and three factors including the interaction term are automatically selected as the independent variables.
  - Click “OK>>”
  - The DOE analysis results appear in the newly generated tab “Analyze – 2 Factor DOE”.

Use SigmaXL to Run a 2^k Full Factorial DOE



Use SigmaXL to Run a 2^k Full Factorial DOE

DOE Multiple Regression Model: Tastiness = (18.125) + (-0.875) * A: Temperature + (-3.125) * B: Time Length + (-1.625) * AB

Title:	
Date:	
Name of Experimenter:	
Notes:	

Design Type:	2 Factor, 4-Run, 2**2, Full-Factorial
Number of Replicates:	2
Number of Blocks:	1
Number of Center Points per Block:	0
Response:	Tastiness

Model Summary:

R-Square	98.60%
R-Square Adjusted	97.54%
S (Root Mean Square Error)	0.612372

High R^2 value shows around 98% of the variation in the response can be explained by the model.

Parameter Estimates (Coded Units):

Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	18.125	0.216506351	83.716	**0.0000**		
A: Temperature	-0.875	0.216506351	-4.041	**0.0156**	1	1
B: Time Length	-3.125	0.216506351	-14.434	**0.0001**	1	1
AB	-1.625	0.216506351	-7.506	**0.0017**	1	1

Analysis of Variance for Model:

Source	DF	SS	MS	F	P
Model	3	105.38	35.125	93.667	**0.0004**
Error	4	1.500	0.375000		
Pure Error	4	1.500	0.375000		
Total (Model + Error)	7	106.88	15.268		

Durbin-Watson Test for Autocorrelation in Residuals:

DW Statistic	1.667
P-Value Positive Autocorrelation	0.1268
P-Value Negative Autocorrelation	0.3484

P value smaller than alpha level (0.05) indicates that the model is statistically significant (i.e. at least one of the independent variables in the model has coefficient statistically different from zero).

Use SigmaXL to Run a 2^k Full Factorial DOE

DOE Multiple Regression Model: Tastiness = (18.125) + (-0.875) * A: Temperature + (-3.125) * B: Time Length + (-1.625) * AB

Title:	
Date:	
Name of Experimenter:	
Notes:	

Design Type:	2 Factor, 4-Run, 2**2, Full-Factorial
Number of Replicates:	2
Number of Blocks:	1
Number of Center Points per Block:	0
Response:	Tastiness

Model Summary:

R-Square	98.60%
R-Square Adjusted	97.54%
S (Root Mean Square Error)	0.612372

Parameter Estimates (Coded Units):

Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	18.125	0.216506351	83.716	0.0000		
A: Temperature	-0.875	0.216506351	-4.041	0.0156	1	1
B: Time Length	-3.125	0.216506351	-14.434	0.0001	1	1
AB	-1.625	0.216506351	-7.506	0.0017	1	1

Analysis of Variance for Model:

Source	DF	SS	MS	F	P
Model	3	105.38	35.125	93.667	0.0004
Error	4	1.500	0.375000		
Pure Error	4	1.500	0.375000		
Total (Model + Error)	7	106.88	15.268		

Durbin-Watson Test for Autocorrelation in Residuals:

DW Statistic	1.667
P-Value Positive Autocorrelation	0.1268
P-Value Negative Autocorrelation	0.3484

Since p values of all the independent variables in the mode are smaller than alpha level (0.05), both factors and their interaction have statistically significant impact on the response.

Use SigmaXL to Run a 2^k Full Factorial DOE

Enter the actual settings of the independent variables into the yellow cells on the "Analyze – 2 Factor DOE" tab. The predicted response will be calculated automatically.

Predicted Response Calculator:

Predictors	Enter Actual Settings:	Coded Settings	Predicted Response	Lower 95% CI	Upper 95% CI	Lower 95% PI	Upper 95% PI
A:							
B:							

Note:
1. Enter settings for predictors. If the predictors are block terms, specify a 0 or 1 for each level.
2. Do not insert or delete rows or columns in this worksheet.

Predicted Response Calculator:

Predictors	Enter Actual Settings:	Coded Settings	Predicted Response	Lower 95% CI	Upper 95% CI	Lower 95% PI	Upper 95% PI
A:	1	1	22	20.797764	23.202236	19.91766617	24.08233383
B:	-1	-1					

Note:
1. Enter settings for predictors. If the predictors are block terms, specify a 0 or 1 for each level.
2. Do not insert or delete rows or columns in this worksheet.

4.4.2 Linear & Quadratic Models

Linear DOE

- If we assume the relationship between the response and the factors is linear, we use a two-level design experiment in which there are two settings for individual factors.
- The linear model of a two-level design with three factors:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_{12} X_1 X_2 + \alpha_{13} X_1 X_3 + \alpha_{23} X_2 X_3 + \alpha_{123} X_1 X_2 X_3 + \varepsilon$$

where
Y is the response.
X_1, X_2, X_3 are the three two factors.
α_0 is the intercept.
α_1, α_2, …, α_{123} are the coefficients of the factors and interactions.
ε is the error of the model.

Quadratic DOE

- It is possible that the relationship between the response and the factors is non-linear.
- To test the non-linear relationship between the response and the factors, we need at least a three-level design, i.e., quadratic design of experiment.
- The quadratic DOE adds the X_1^2, X_2^2,…, X_k^2 and corresponding interactions into the model as potential significant factor of the response.

4.4.3 Balanced & Orthogonal Designs

Balanced Design

- In DOE, an experiment design is *balanced* if individual treatments have the same number of observations.
- In other words, if an experiment is balanced, each level of each factor is run the same number of times.

- The following design is balanced since the two factors are both run twice at each level.

Run	Treatment	Factors	
		A	B
1	I	-1	-1
2	a	+1	-1
3	b	-1	+1
4	ab	+1	+1

Orthogonal Design

- In DOE, an experiment design is *orthogonal* if the main effect of one factor can be evaluated independently of other factors.

- In other words, for each level (high setting or low setting) of a factor, the number of high setting and low setting in any other factors must be the same.

Orthogonal Design

- In the following design, for the higher level (+1) of factor A, the number of "+1" in factor B is 2 but the number of "-1" in factor B is 0. Therefore, this is *not* an orthogonal design.

Run	Treatment	Factors	
		A	B
1	l	-1	-1
2	a	-1	-1
3	b	+1	+1
4	ab	+1	+1

Orthogonal Design

- In the following design, for the higher level (+1) of factor A, the number of "+1" in factor B is 1 and the number of "-1" in factor B is 1 as well. Therefore, this is an orthogonal design.

Run	Treatment	Factors	
		A	B
1	I	-1	-1
2	a	+1	-1
3	b	-1	+1
4	ab	+1	+1

Checking for an Orthogonal Design

- We can check to see if our design is orthogonal with some very simple math.
- In an orthogonal design, the sum of products for each run should equal zero.
- Let us check the earlier design that we said was orthogonal:

  - Run 1: -1 * -1 = 1
  - Run 2: +1 * -1 = -1
  - Run 3: -1 * +1 = -1
  - Run 4: +1 * +1 = 1
  - Sum = **0**
  - **Design is orthogonal**

Run	Treatment	Factors	
		A	B
1	I	-1	-1
2	a	+1	-1
3	b	-1	+1
4	ab	+1	+1

Checking for an Orthogonal Design

- Now let us check the earlier design that we said was not orthogonal:

  - Run 1: -1 * -1 = 1
  - Run 2: -1 * -1 = 1
  - Run 3: +1 * +1 = 1
  - Run 4: +1 * +1 = 1
  - Sum = **4**
  - **Design is not orthogonal**

Run	Treatment	Factors	
		A	B
1	I	-1	-1
2	a	-1	-1
3	b	+1	+1
4	ab	+1	+1

4.4.4 Fit, Diagnosis and Center Points

Use SigmaXL to Fit the Model of DOE

- Recall in the example introduced in last module, we used "SigmaXL -> Design of Experiments -> 2-Level Factorial/Screening -> Analyze 2-Level Factorial/Screening Design" to generate a linear model as follows for the response, the tastiness of the cake.

DOE Multiple Regression Model: Tastiness = (18.125) + (-0.875) * A: Temperature + (-3.125) * B: Time Length + (-1.625) * AB

Title:	
Date:	
Name of Experimenter:	
Notes:	

Design Type:	2 Factor, 4-Run, 2**2, Full-Factorial
Number of Replicates:	2
Number of Blocks:	1
Number of Center Points per Block:	0
Response:	Tastiness

Model Summary:

R-Square	98.60%
R-Square Adjusted	97.54%
S (Root Mean Square Error)	0.612372

Since p values of all the independent variables in the mode are smaller than alpha level (0.05), both factors and their interaction have statistically significant impact on the response.

Parameter Estimates (Coded Units):

Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	18.125	0.216506351	83.716	0.0000		
A: Temperature	-0.875	0.216506351	-4.041	0.0156	1	1
B: Time Length	-3.125	0.216506351	-14.434	0.0001	1	1
AB	-1.625	0.216506351	-7.506	0.0017	1	1

Analysis of Variance for Model:

Source	DF	SS	MS	F	P
Model	3	105.38	35.125	93.667	0.0004
Error	4	1.500	0.375000		
Pure Error	4	1.500	0.375000		
Total (Model + Error)	7	106.88	15.268		

Durbin-Watson Test for Autocorrelation in Residuals:

DW Statistic	1.667
P-Value Positive Autocorrelation	0.1268
P-Value Negative Autocorrelation	0.3484

Use SigmaXL to Fit the Model of DOE

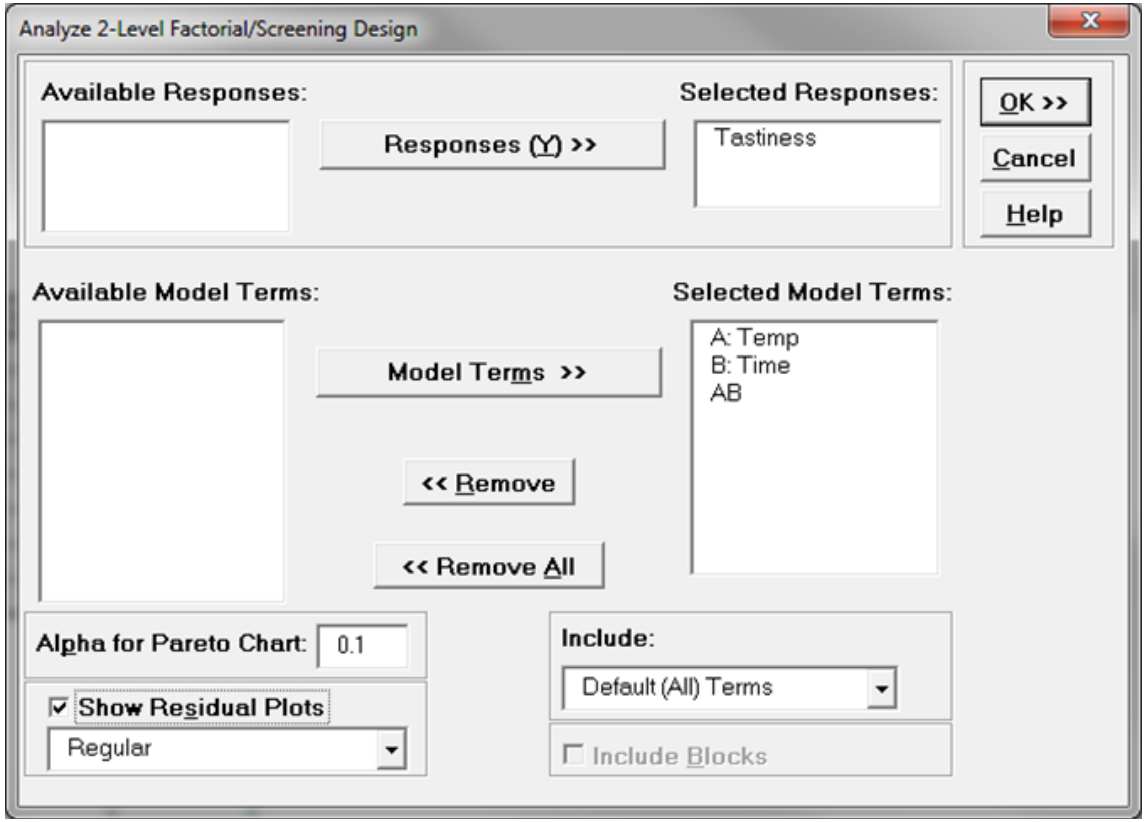
- If one of the factors or interactions has p-value larger than alpha level (0.05), it indicates that the particular factor or interaction does not have statistically significant impact on the response.
- To fit the model better, we need to remove the insignificant independent variables one at a time and run the model again until all the independent variables in the model are statistically significant (i.e. p-value smaller than alpha level).

Use SigmaXL to Diagnose the Model

- To ensure that the quality of the model, we need to conduct the residual analysis.
  - Residual is the difference between the actual response value and the fitted response value
- Well-performing residuals satisfy
  - Normally distributed
  - Mean is equal to zero
  - Independent
  - Equal variance across the fitted values

Use SigmaXL to Diagnose the Model

- Step 1: Save the residuals and residual plots
  - Click SigmaXL -> Design of Experiments -> 2-Level Factorial/Screening -> Analyze 2-Level Factorial/Screening Design
  - Both the dependent and independent variables are populated automatically.
  - Check the checkbox “show residual plots” and select the regular residuals in the drop down box right below “Show Residual Plots”
  - Click “OK>>”
  - The residuals and residual plots appear at the bottom of the tab “Analyze – 2 Factor DOE (1)”

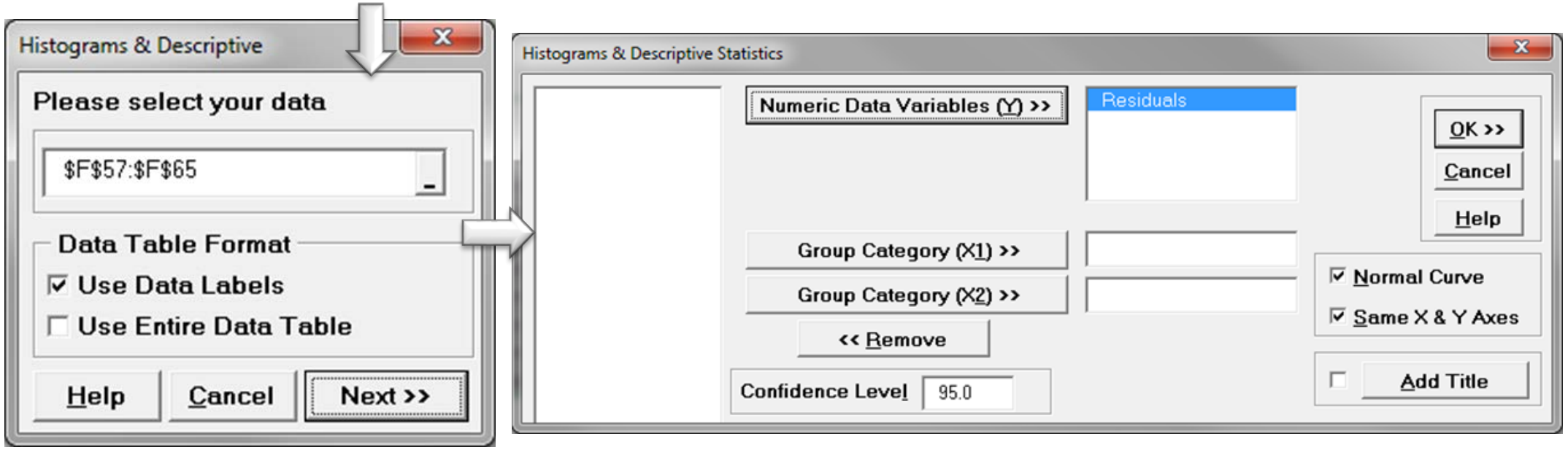


Use SigmaXL to Diagnose the Model

- Step 2: Check whether the mean of residuals is zero and residuals are normally distributed.
  - Select the entire range of the residuals data listed in the residual report
  - Click SigmaXL -> Graphical Tools -> Histograms & Descriptive Statistics
  - A new window named “Histograms & Descriptive” appears with the selected residuals automatically populated in the box below “Please select your data”
  - Click “Next>>”
  - A new window named “Histograms & Descriptive Statistics” appears
  - Select “Residuals” as the “Numeric Data Variables (Y)”
  - Click “OK>>”
  - The histogram and normality test results appear in the newly generated tab “Hist Descript (1)”

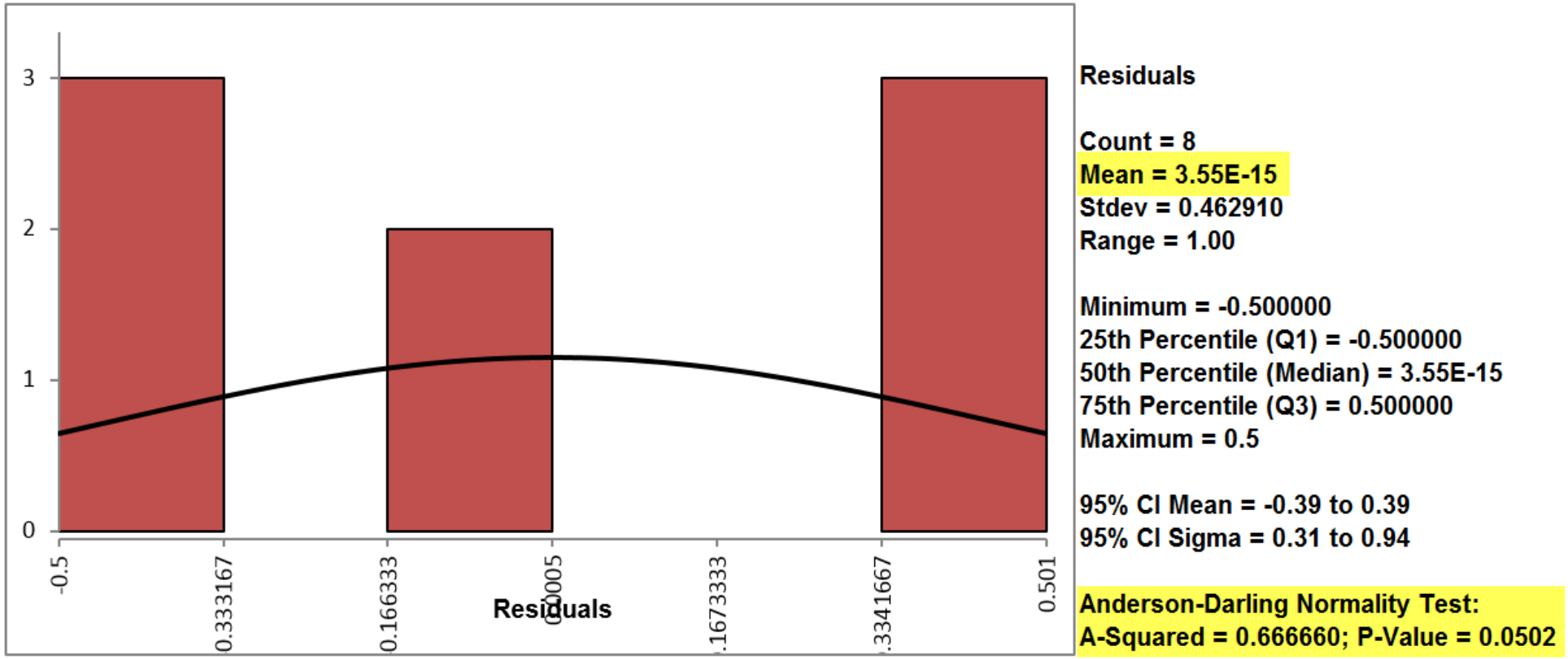
Use SigmaXL to Diagnose the Model

A: Temp	B: Time	Tastiness	Predicted (Fitted) Values	Residuals	Standardized Residuals	Studentized (Deleted t)	Cook's Distance (Influence)	Leverage	DFITS
-1	1	22	22.000	0.000000000000004	0.000000000000008	0.000000000000007	0	0.500000	0.000000000000007
1	-1	17	17.500	-0.499999999999996	-1.154700538379240	-1.224744871391580	0.333333	0.500000	-1.224744871391580
-1	-1	21	20.500	0.500000000000004	1.154700538379260	1.224744871391600	0.333333	0.500000	1.224744871391600
-1	1	22	22.000	0.000000000000004	0.000000000000008	0.000000000000007	0	0.500000	0.000000000000007
-1	-1	20	20.500	-0.499999999999996	-1.154700538379240	-1.224744871391580	0.333333	0.500000	-1.224744871391580
1	-1	18	17.500	0.500000000000004	1.154700538379260	1.224744871391600	0.333333	0.500000	1.224744871391600
1	1	12	12.500	-0.499999999999996	-1.154700538379240	-1.224744871391580	0.333333	0.500000	-1.224744871391580
1	1	13	12.500	0.500000000000004	1.154700538379260	1.224744871391600	0.333333	0.500000	1.224744871391600



Use SigmaXL to Diagnose the Model

The mean of the residuals are approximately zero.

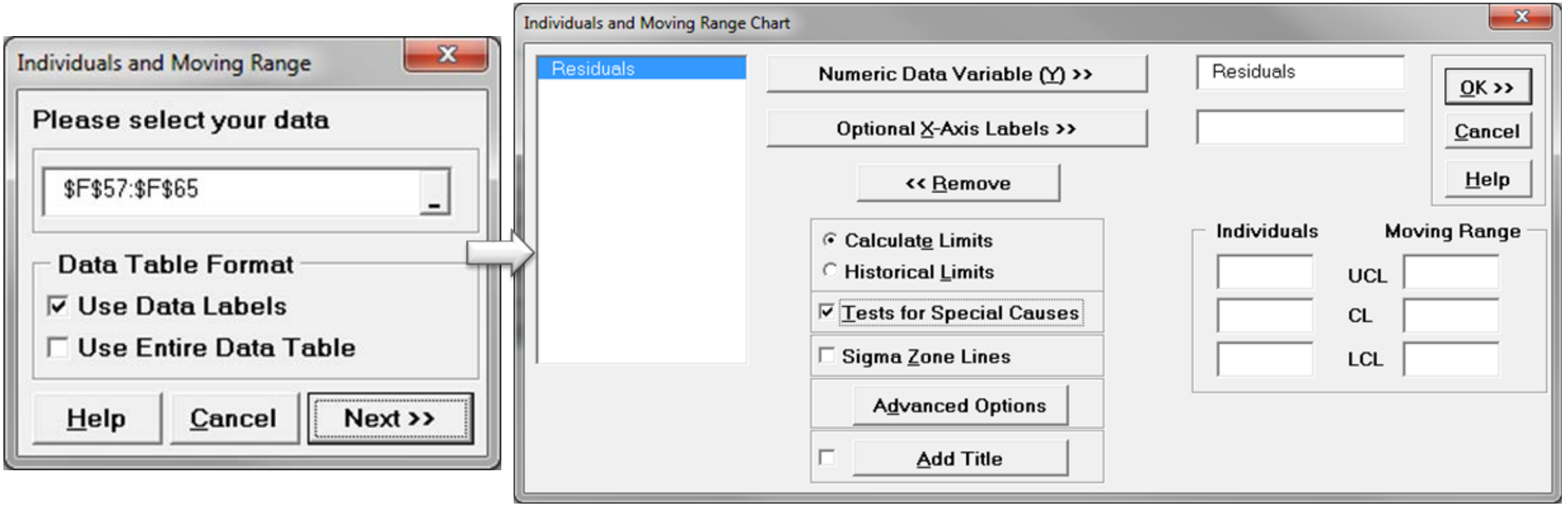


Since the p-value of the normality test is greater than alpha level, the residuals are normally distributed.

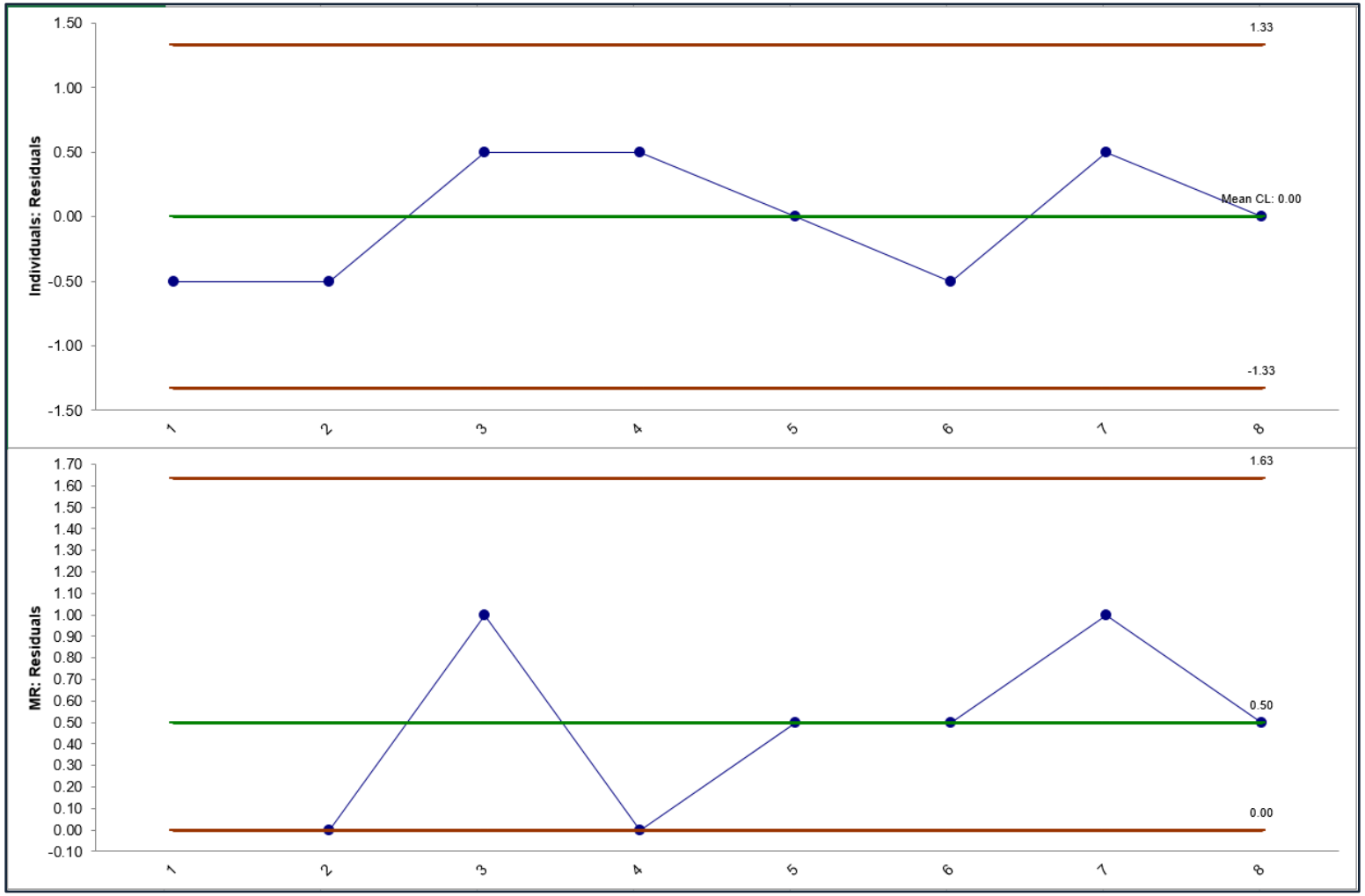
Use SigmaXL to Diagnose the Model

- Step 3: Check whether the residuals are independent.
  - Select the entire range of the residuals in the residual report
  - Click SigmaXL -> Control Charts -> Individuals & Moving Range
  - A new window named “Individuals & Moving Range” appears with the selected range automatically populated into the box under “Please select your data”
  - Click “Next>>”
  - A new window named “Individuals & Moving Range Chart” pops up
  - Select “Residuals” as the “Numeric Data Variable (Y)”
  - Check the checkbox “Test for Special Causes”
  - Click “OK”
  - If no data points on the control charts fail any tests, the residuals are in control and independent of each other.
  - Note: The prerequisite of plotting IR chart for residuals is that the residuals are in the time order.

Use SigmaXL to Diagnose the Model

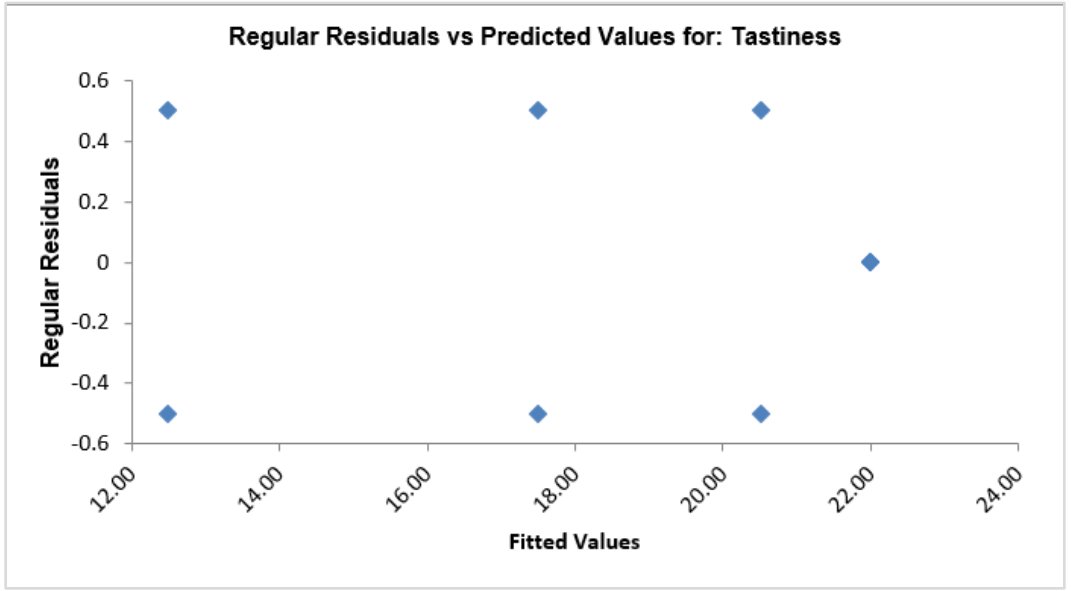


Use SigmaXL to Diagnose the Model



Use SigmaXL to Diagnose the Model

- Step 4: Check whether the residuals have equal variance across the predicted response values.
  - Close to the bottom of the tab “Analyze – 2 Factor DOE (1)” is the residual by predicted plot.
  - We look for patterns in which the residuals tend to have even variation across the entire range of the fitted response values.



Center Points in DOE

- The two-level design of experiment assumes that the relationship between the response and the factors is linear.
- We can use center points to check whether the assumption is true by adding center point runs to the experiment.

- In each center point run, the factors are all set to be in the center setting (zero) between high (+1) and low (-1) settings.
  - Center points do not change the model to quadratic.
  - They allow a check for adequacy of linear model.
  - If a line connecting factor settings passes through the center of the design, the model is adequate to predict within the inference space.

4.5 Fractional Factorial Experiments

Black Belt Training: Improve Phase

4.1 Simple Linear Regression
- 4.1.1 Correlation
- 4.1.2 X-Y Diagram
- 4.1.3 Regression Equations
- 4.1.4 Residuals Analysis

4.2 Multiple Regression Analysis
- 4.2.1 Non-Linear Regression
- 4.2.2 Multiple Linear Regression
- 4.2.3 Confidence Intervals
- 4.2.4 Residuals Analysis
- 4.2.5 Data Transformation, Box Cox
- 4.2.6 Stepwise Regression
- 4.2.7 Logistic Regression

4.3 Designed Experiments
- 4.3.1 Experiment Objectives
- 4.3.2 Experimental Methods
- 4.3.3 DOE Design Considerations

4.4 Full Factorial Experiments
- 4.4.1 2k Full Factorial Designs
- 4.4.2 Linear and Quadratic Models
- 4.4.3 Balanced and Orthogonal Designs
- 4.4.4 Fit, Model, and Center Points

4.5 Fractional Factorial Experiments
- 4.5.1 Designs
- 4.5.2 Confounding Effects
- 4.5.3 Experimental Resolution

4.5.1 Fractional Designs

What Are Fractional Factorial Experiments?

- In simple terms, a fractional factorial experiment is a subset of a full factorial experiment.
- Fractional factorials use fewer treatment combinations and runs.

- Fractional factorials are less able to determine effects because of fewer degrees of freedom available to evaluate higher order interactions.
- Fractional factorials can be used to screen a larger number of factors.
- Fractional factorials can also be used for optimization.

Why Fractional Factorial Experiments?

- To run a full factorial experiment for k factors, we need 2^k unique treatments. In other words, we need resources that can afford at least 2^k runs.

- With k increasing, the number of runs required in full factorial experiments rises dramatically even without any replications, and the percentage of degrees of freedom spent on the main effects decreases.

- The main effects and two-way interaction are the key effects we need to evaluate. The higher order the interaction is, the more we can ignore it.

Why Fractional Factorial Experiments?

Number of Factors	Number of Treatments
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024

How Does a Fractional Factorial Work?

- We are trying to find the cause-and-effect relationship between a response (Y) and three factors (factor A, B, and C) and their interactions (AB, BC, AC, and ABC).
  - As follows is the 2^3 full factorial design (2 level 3 factor).
  - There are eight treatment combinations (2 * 2 * 2).

Run	Treatment	Factor			2-Way Interaction			3-Way Interaction
		A	B	C	AB	BC	AC	ABC
1	I	-1	-1	-1	+1	+1	+1	-1
2	a	+1	-1	-1	-1	+1	-1	+1
3	b	-1	+1	-1	-1	-1	+1	+1
4	ab	+1	+1	-1	+1	-1	-1	-1
5	c	-1	-1	+1	+1	-1	-1	+1
6	ac	+1	-1	+1	-1	-1	+1	-1
7	bc	-1	+1	+1	-1	+1	-1	-1
8	abc	+1	+1	+1	+1	+1	+1	+1

How Does a Fractional Factorial Work?

- To perform a 2^3 full factorial experiment, we need to run at least eight unique treatments (2 * 2 * 2).
- What if we only have enough resources to run four treatments?
- As a result, we need to carefully select a subset from the eight treatments so that all of our main effects can be evaluated and the design can be kept balanced and orthogonal.

How Does a Fractional Factorial Work?

- Example of an invalid design

Factor		
A	B	C
-1	-1	-1
+1	-1	-1
-1	+1	-1
+1	+1	-1

- This design is invalid because only the low setting of factor C is tested.
- We cannot evaluate the main effect of factor C using this design.
- Remember orthogonality?

How Does a Fractional Factorial Work?

- Example of an invalid design

Factor		
A	B	C
+1	+1	-1
+1	-1	-1
-1	+1	+1
+1	+1	+1

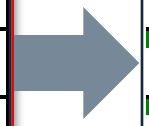
- This design is also invalid because it is neither balanced nor orthogonal.

- Checking orthogonality: the sum of AC interaction signs should equal zero (0).
  - Run 1 (-)
  - Run 2 (-)
  - Run 3 (-)
  - Run 4 (+)
  - **Sum (-1)**

How Does Fractional Factorial Work?

- To select the four treatments run in the 2^{3-1} fractional factorial experiment, we start from the 2^2 full factorial design of experiment.
- If we replace the two-way interaction (AB) column with the factor C column, the design will be valid.

A	B	AB
-1	-1	+1
+1	-1	-1
-1	+1	-1
+1	+1	+1



A	B	C
-1	-1	+1
+1	-1	-1
-1	+1	-1
+1	+1	+1

How Does a Fractional Factorial Work?

- 2^{3-1} Fractional Factorial Design Pattern
  - Three factors and four treatments

Run	Treatment	Factor			2-Way Interaction			3-Way Interaction
		A	B	C	AB	BC	AC	ABC
1	c	-1	-1	+1	+1	-1	-1	+1
2	a	+1	-1	-1	-1	+1	-1	+1
3	b	-1	+1	-1	-1	-1	+1	+1
4	abc	+1	+1	+1	+1	+1	+1	+1

- Note: We also call this kind of design a *half-factorial design* since we only have half of the treatments that we would have in a full factorial design.
- In 2^{3-1} fractional factorial design of experiment, the effect of three-way interaction (ABC) is not measurable since it only has “+1”.

How Does a Fractional Factorial Work?

- In the 2^{3-1} fractional factorial design, we notice that the column of each main effect has identical “+1” and “-1” values with one two-way interaction column.
  - A and BC
  - B and AC
  - C and AB

- In this situation, we say that A is *aliased* with BC or A is the *alias* of BC.

How Does a Fractional Factorial Work?

- By multiplying any column with itself, we obtain the *identity (I)*.
  - A * A = I
  - The product of any column and the identity is the column itself.
  - A * I = A

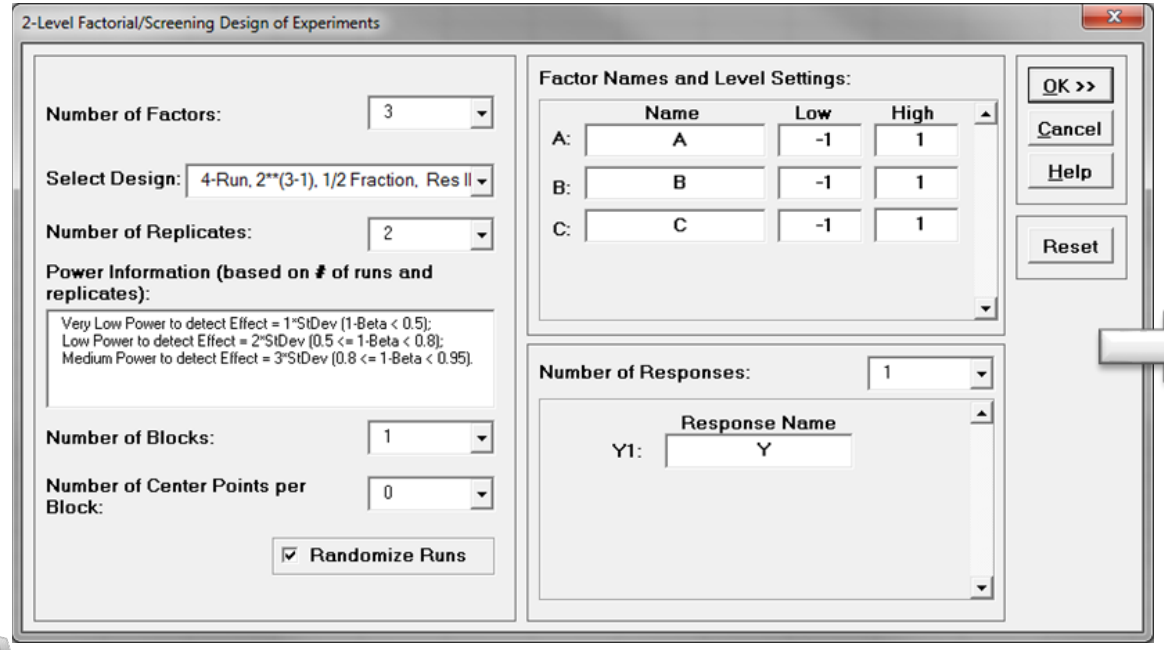
- Column ABC is called the *generator*.
  - By multiplying any column with the generator, we obtain its *alias*.
  - A * ABC = (A*A) * BC = I * BC = BC

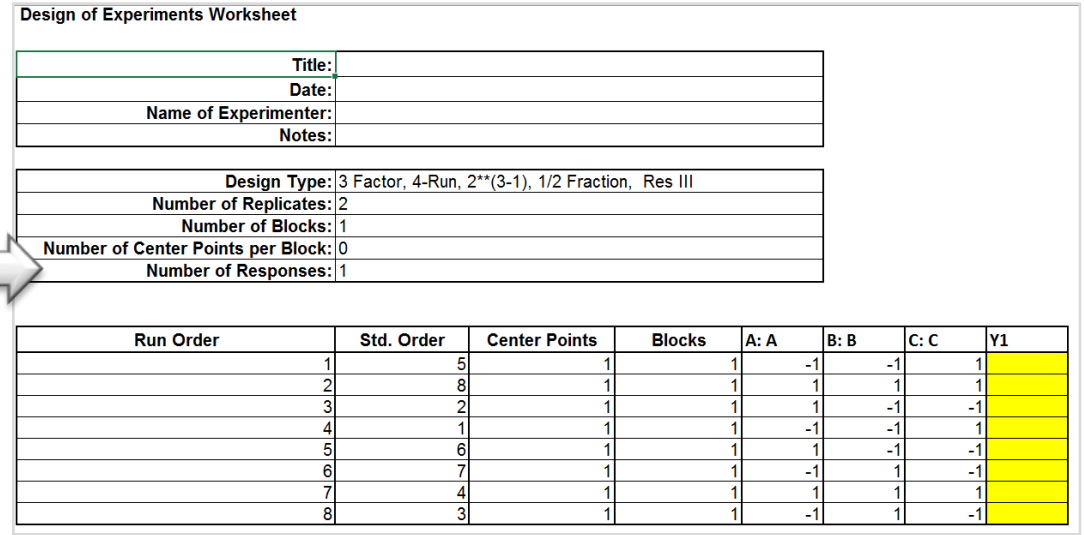
Use SigmaXL to Run a Fractional Factorial Experiment

- Step 1: Initiate the experiment design
  - Click SigmaXL -> Designs of Experiments -> 2-Level Factorial/Screening -> 2-Level Factorial/Screening Designs
  - A window named “2-Level Factorial/Screening Designs of Experiments” pops up.
- Step 2: Enter the response and factors information in the window “2-Level Factorial/Screening Design of Experiments”
  - One response: Y
  - Three factors: A, B and C
  - Two-level design: each factor has two settings
- Step 3: Select a design from the pre-populated design list.
  - In this case, we select “4-Run, 2**(3-1), ½ Fraction, Res III”.

Use SigmaXL to Run a Fractional Factorial Experiment

- Step 4: Select the number of replications and make design
  - In this case, we assume the sufficient resources allow each treatment to be run twice.
  - Enter "2" into the box of "Number of Replicates"
  - Click "OK>>"
  - The DOE template appears in the newly generated tab "3 Factor DOE".





Design of Experiments Worksheet

Title:	
Date:	
Name of Experimenter:	
Notes:	

Design Type:	3 Factor, 4-Run, 2**(3-1), 1/2 Fraction, Res III
Number of Replicates:	2
Number of Blocks:	1
Number of Center Points per Block:	0
Number of Responses:	1

Run Order	Std. Order	Center Points	Blocks	A: A	B: B	C: C	Y1
1	5	1	1	-1	-1	1	
2	8	1	1	1	1	1	
3	2	1	1	1	-1	-1	
4	1	1	1	-1	-1	1	
5	6	1	1	1	-1	-1	
6	7	1	1	-1	1	-1	
7	4	1	1	1	1	1	
8	3	1	1	-1	1	-1	

Use SigmaXL to Run a Fractional Factorial Experiment

- Step 5: Implement the experiment and record the results in yellow cells in the DOE table being sure to enter the data by Std. Order.

Design of Experiments Worksheet

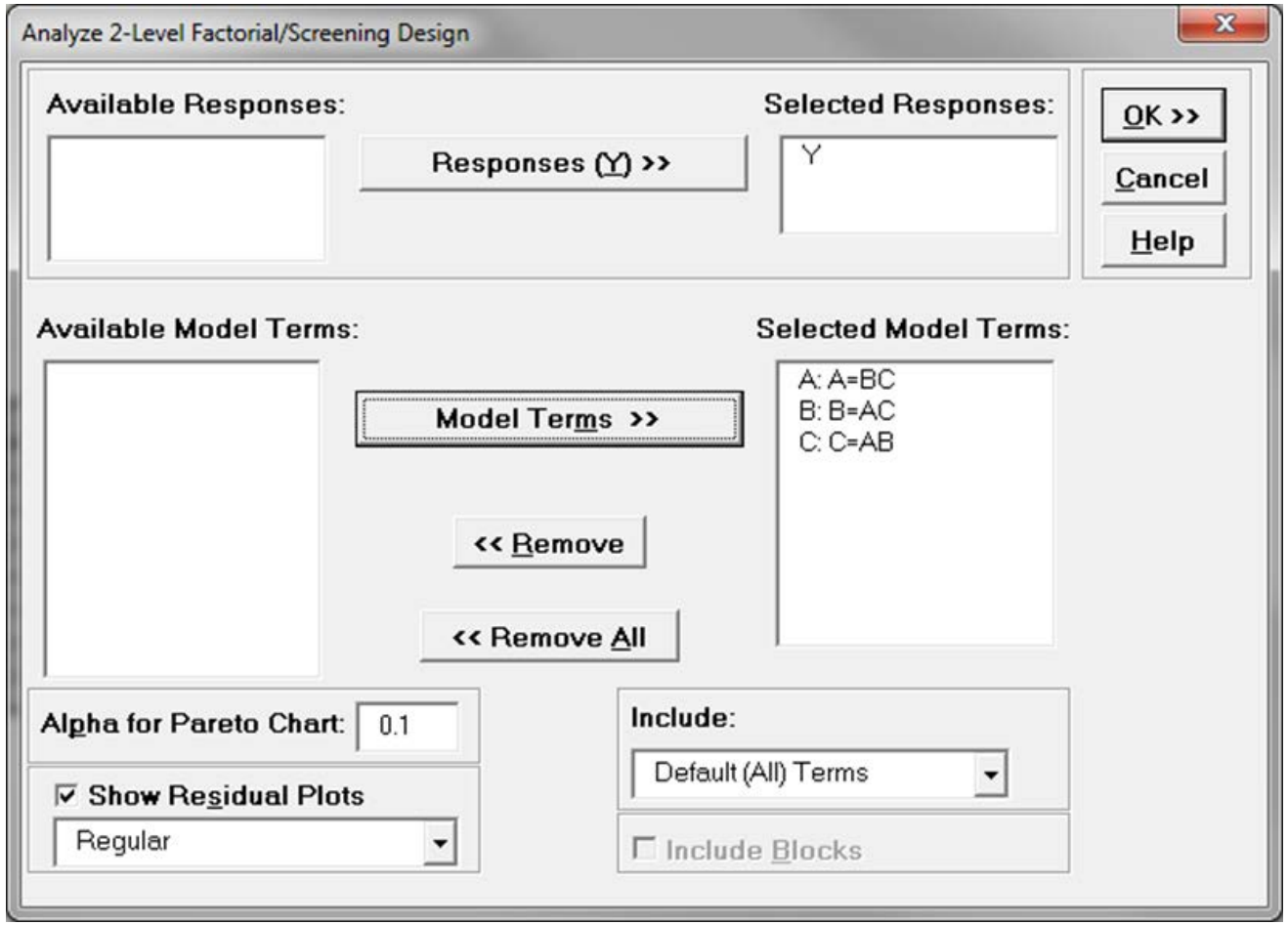
Title:	
Date:	
Name of Experimenter:	
Notes:	

Design Type:	2 Factor, 4-Run, 2**2, Full-Factorial
Number of Replicates:	2
Number of Blocks:	1
Number of Center Points per Block:	0
Number of Responses:	1

Run Order	Std. Order	Center Points	Blocks	A: A	B: B	Y1
1	1	1	1	-1	-1	21
2	8	1	1	1	1	15
3	3	1	1	-1	1	22
4	6	1	1	1	-1	17
5	5	1	1	-1	-1	20
6	2	1	1	1	-1	18
7	4	1	1	1	1	15
8	7	1	1	-1	1	23

Use SigmaXL to Run a Fractional Factorial Experiment

- Step 6: Fit the model using the experiment results
  - Click SigmaXL -> Design of Experiments -> 2-Level Factorial/Screening -> Analyze 2-Level Factorial/Screening Design
  - A new window named "Analyze 2-Level Factorial/Screening Design" appears with the response and factors pre-populated.
  - Check the checkbox "Show Residual Plots"
  - Click "OK>>"



Use SigmaXL to Run a Fractional Factorial Experiment

- Step 7: Analyze the model results
  - Check whether the model is statistically significant.
  - Check which factors are insignificant.
  - If any independent variables are not significant, remove them one at a time and rerun the model until all the independent variables in the model are significant.

Use SigmaXL to Run a Fractional Factorial Experiment

Model Summary:

R-Square	97.76%
R-Square Adjusted	96.07%
S (Root Mean Square Error)	0.612372

Parameter Estimates (Coded Units):

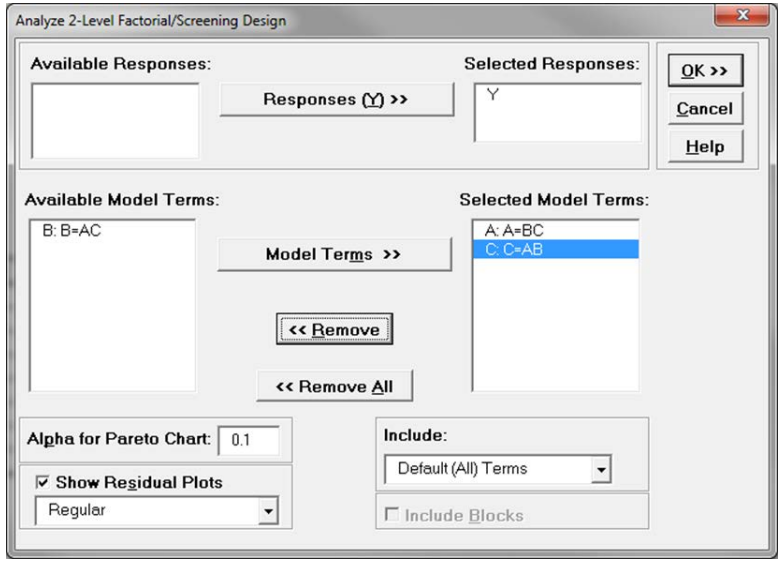
Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	18.875	0.216506351	87.180	0.0000		
A: A=BC	-2.625	0.216506351	-12.124	0.0003	1	1
B: B=AC	-0.125	0.216506351	-0.577350	0.5946	1	1
C: C=AB	-1.125	0.216506351	-5.196	0.0065	1	1

Analysis of Variance for Model:

Source	DF	SS	MS	F	P
Model	3	65.375	21.792	58.111	0.0009
Error	4	1.500	0.375000		
Pure Error	4	1.500	0.375000		
Total (Model + Error)	7	66.875	9.554		

- The p-value of factor B is greater than the alpha level (0.05), so it is not statistically significant.
- Next we need to remove factor B and re-run the model.

Use SigmaXL to Run a Fractional Factorial Experiment



The p-values of all the independent variables are smaller than 0.05. There is no need to remove any independent variables from the model.

Model Summary:

R-Square	97.57%
R-Square Adjusted	96.60%
S (Root Mean Square Error)	0.570088

Parameter Estimates (Coded Units):

Term	Coefficient	SE Coefficient	T	P	VIF	Tolerance
Constant	18.875	0.201556444	93.646	**0.0000**		
A: A=BC	-2.625	0.201556444	-13.024	**0.0000**	1	1
C: C=AB	-1.125	0.201556444	-5.582	**0.0025**	1	1

Analysis of Variance for Model:

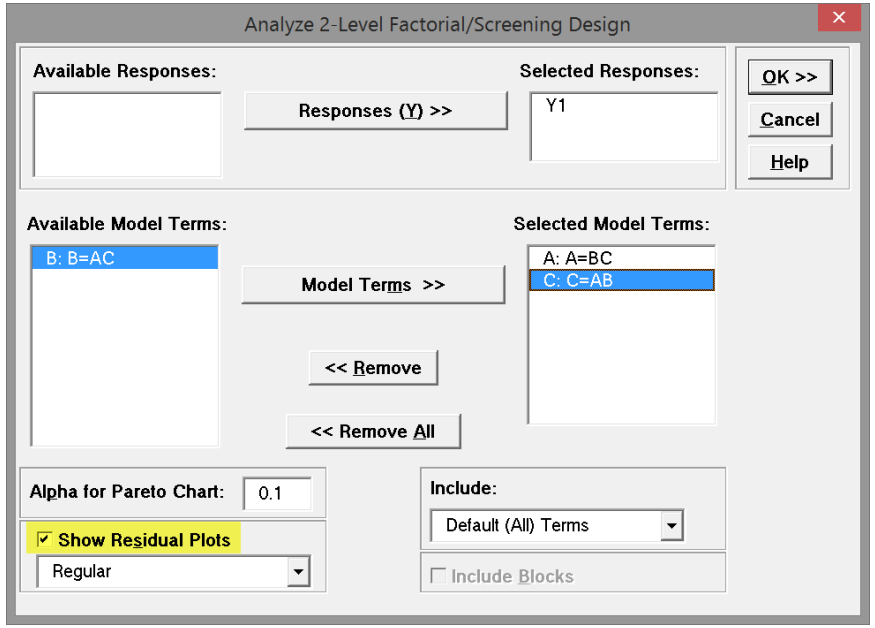
Source	DF	SS	MS	F	P
Model	2	65.250	32.625	100.38	**0.0001**
Error	5	1.625	0.325000		
Lack of Fit	1	0.125000	0.125000	0.333333	0.5946
Pure Error	4	1.500	0.375000		
Total (Model + Error)	7	66.875	9.554		

Use SigmaXL to Run a Fractional Factorial Experiment

- Step 8: Conduct residual analysis to ensure that the residuals of the model satisfy the following criteria.
  - Mean is equal to zero.
  - Normally distributed
  - Independent
  - Equal variance across the fitted response values

Use SigmaXL to Run a Fractional Factorial Experiment

- Step 8.1: Save residuals generated by the model
  - When running the “Analyze 2-Level Factorial/Screening Design”, make sure to check the checkbox “Show Residual Plots” before clicking “OK>>”
  - The residual plots and the residuals appear at the bottom of the tab “Analyze – 3 Factor DOE (1)”.



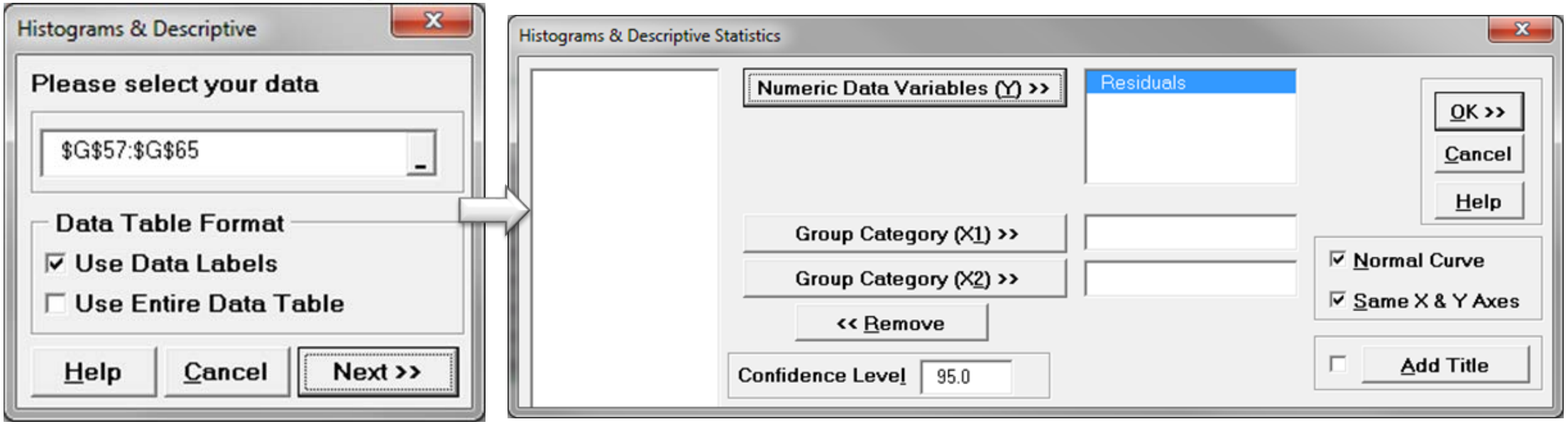
Use SigmaXL to Run a Fractional Factorial Experiment

- Step 8.2: Check whether residuals are normally distributed with mean equal to zero.
  - Select the entire range of the residuals in the residual report
  - Click SigmaXL -> Graphical Tools -> Histograms & Descriptive Statistics
  - A new window named “Histograms & Descriptive” appears and the selected range is automatically populated into the box below “Please select your data”
  - Click “Next>>”
  - A new window named “Histograms & Descriptive Statistics” pops up.
  - Select “Residuals” as the “Numeric Data Variables (Y)”
  - Click “OK>>”
  - The histogram and the normality test of the residuals appear in the newly generated tab “Hist Descript (1)”.
  - If the p-value of the normality test is greater than the alpha level (0.05), the residuals are normally distributed.

Use SigmaXL to Run a Fractional Factorial Experiment

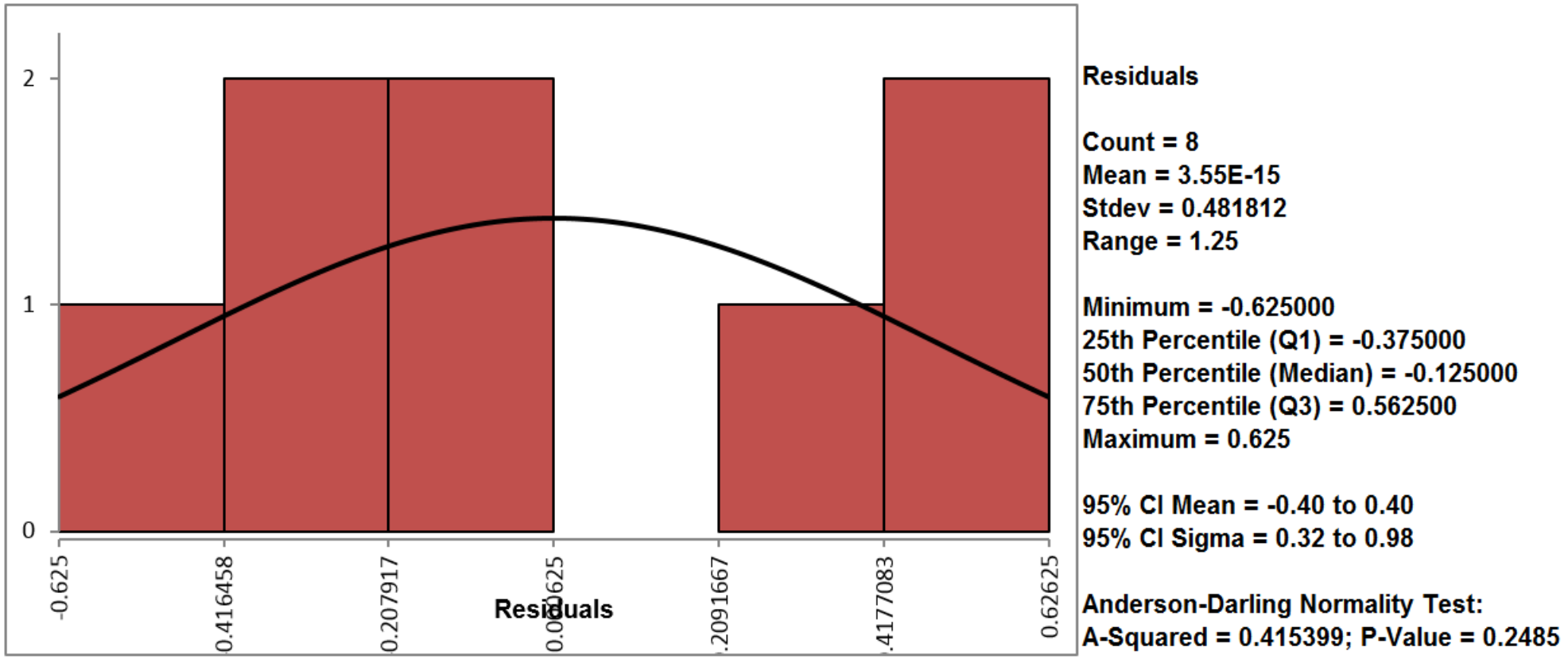
Residuals Report

A: A	B: B	C: C	Y1	Predicted (Fitted) Values	Residuals	Standardized Residuals
1	-1	-1	17	17.375	-0.375000	-0.832050
1	1	1	15	15.125	-0.125000	-0.277350
-1	1	-1	23	22.625	0.375000	0.832050
1	-1	-1	18	17.375	0.625000	1.387
-1	-1	1	21	20.375	0.625000	1.387
-1	-1	1	20	20.375	-0.375000	-0.832050
1	1	1	15	15.125	-0.125000	-0.277350
-1	1	-1	22	22.625	-0.625000	-1.387



Use SigmaXL to Run a Fractional Factorial Experiment

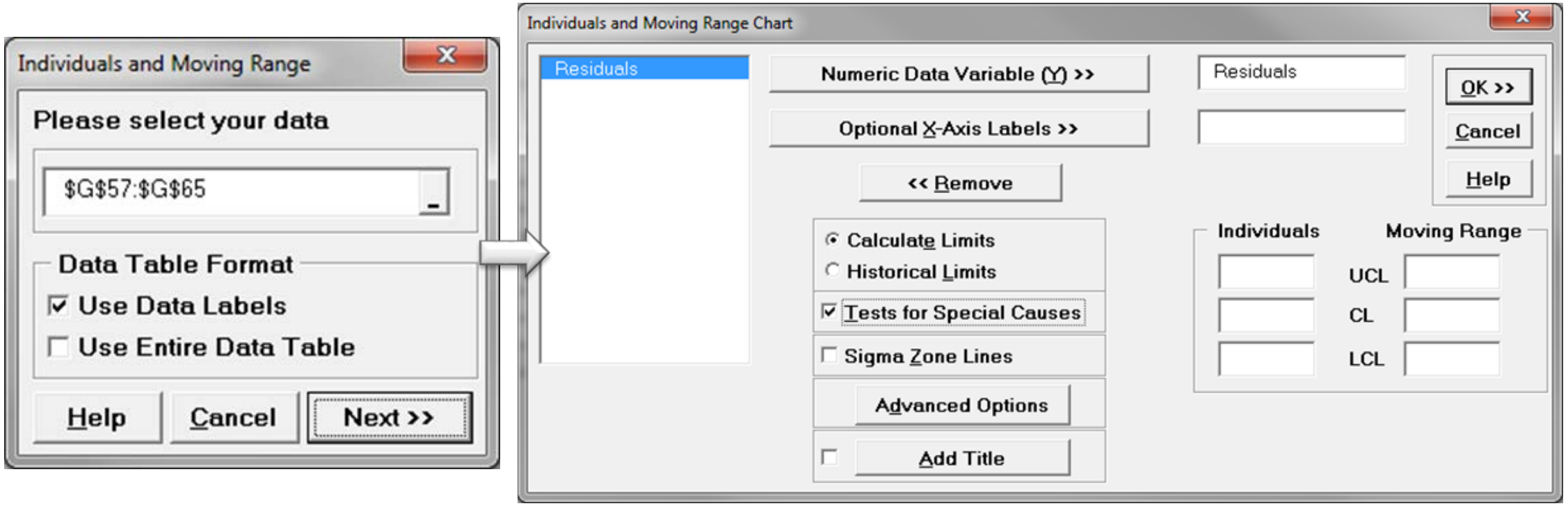
The p-value of the normality test is larger than alpha level (0.05). The residuals are normally distributed.



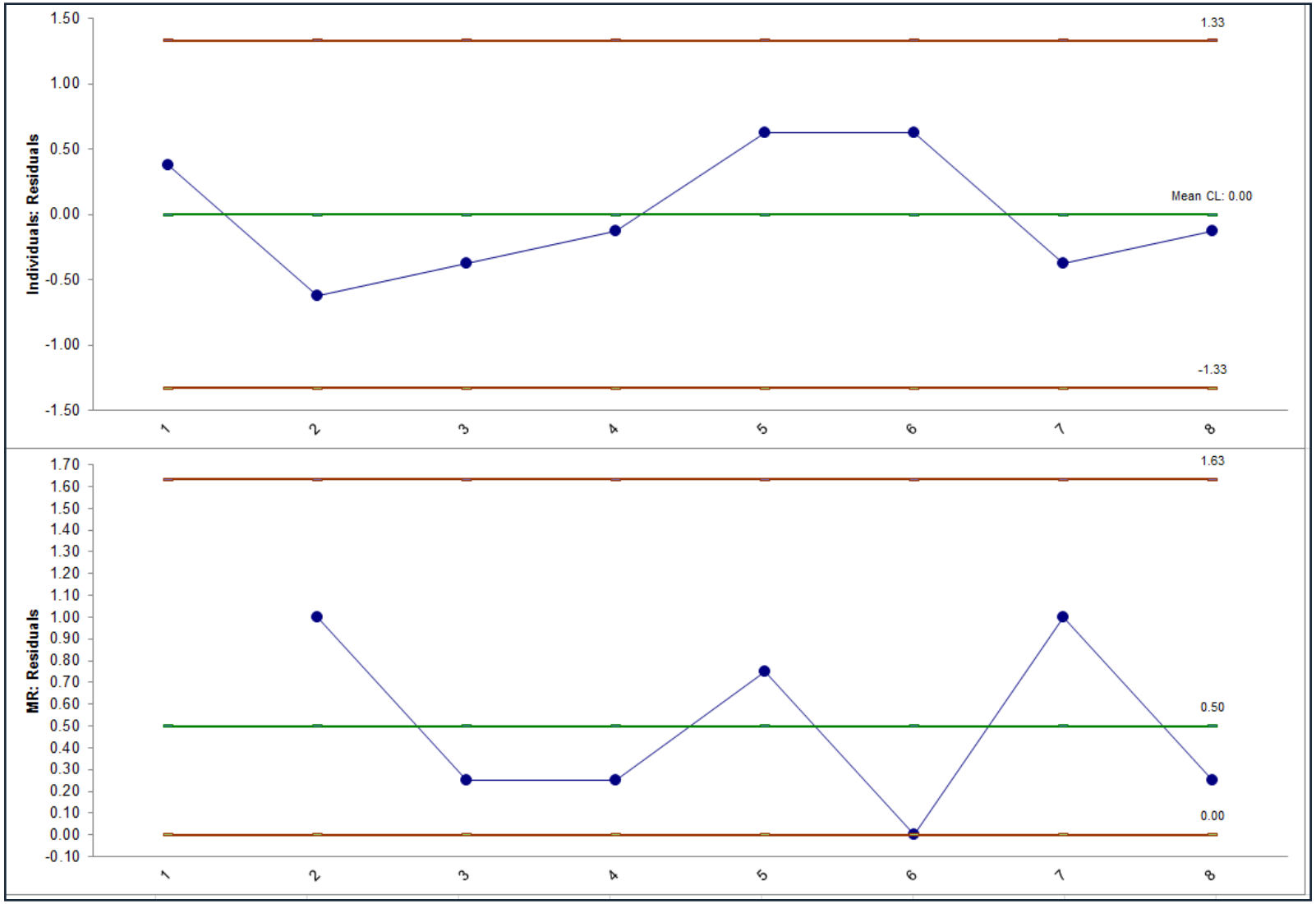
Use SigmaXL to Run a Fractional Factorial Experiment

- Step 8.3: Check whether the residuals are independent.
  - Select the entire range of the residuals in the residual report.
  - Click SigmaXL -> Control Charts -> Individuals & Moving Range
  - A new window named “Individuals and Moving Range” appears with the selected range automatically populated into the box below “Please select your data”.
  - Click “Next>>”
  - A new window named “Individuals and Moving Range Chart” pops up.
  - Select the “Residuals” as the “Numeric Data Variables (Y)”
  - Click “OK>>”
  - The control charts appear in the newly generated tab “Indiv & MR Charts (1)”.
  - If no data points on the control charts fail any tests, the residuals are in control and independent of each other.
  - Note: The prerequisite of plotting IR chart for residuals is that the residuals are in the time order.

Use SigmaXL to Run a Fractional Factorial Experiment

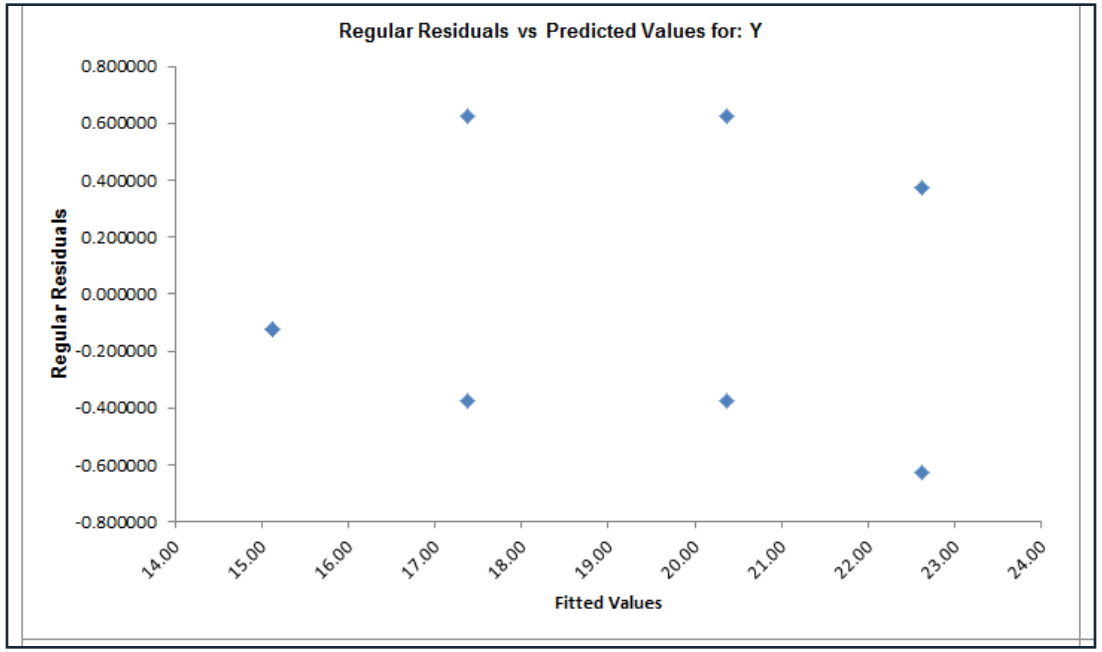


Use SigmaXL to Run a Fractional Factorial Experiment



Use SigmaXL to Run a Fractional Factorial Experiment

- Step 8.4: Check whether the residuals have equal variance across the predicted response values.
  - Close to the bottom of the tab “Analyze – 3 Factor DOE (1)” is the residual by predicted plot.
  - We look for patterns in which the residuals tend to have even variation across the entire range of the fitted response values.



4.5.2 Confounding Effects

Confounding Effects

- In full factorial experiments the effects of every factor and their interactions can be calculated because there is a degree of freedom for every treatment combination as well as the error term.
- Fractional factorial experiments, however, are designed with fewer runs or treatment combinations but will have the same number of input factors.
- As a result, the experimenter must understand the impact of these consequences.
- **Confounding** (or aliased factors) is that consequence.
- **Resolution** is a quantification or degree of confounding.

Confounding Effects

When two input factors are aliases of each other, the effects they each have on the response cannot be separated.

In the 2^{3-1} fractional factorial design, A and BC are aliases.

The main effect that A has on the response would be exactly the same as the interaction effect of BC. We cannot tell which one is truly significant.

- The effects of aliased factors are what is referred to as **confounded**.
- Due to the confounding nature of aliased factors, the effect we observe is the mixed effect of aliased factors.
  - A + BC
  - B + AC
  - C + AB

Resolution

- **Resolution** is the measure or degree of confounding.
- Higher resolution means less confounding or merely confounding main effects with higher order interactions.
- The resolution of a fractional factorial experiment will always be one number higher than the order of interactions that are confounded with main effects.
- **Resolution III**: Main effects are confounded with two-way interactions. If you are only interested in main effects, then resolution III is acceptable. Otherwise, resolution III is typically undesirable.
- **Resolution IV**: Main effect are confounded with three-way interactions.
- **Resolution V**: Main effects are confounded with four-way interactions.

Consequences of Fractional Factorials

- Benefits:
  - The primary benefits of properly designed fractional factorial experiments are achieved because of fewer runs.
  - You may have limited time, resources, or capital and cannot run a full factorial.
- Disadvantages:
  - Confounding (a.k.a. aliasing)

- Mitigation Option
  - **Replications**: by adding replicates you can introduce more study runs and increase the resolution of your experiment.
  - Replications, however, will also increase the number of trials or runs sometimes, defeating the purpose of running a fractional factorial experiment.

4.5.3 Experimental Resolution

Alias in 2^{3-1} Fractional Factorial Experiment

- In the 2^{3-1} fractional factorial design of experiment, the main factor is *aliased* or *confounded* with two-way interactions.
  - A is the alias of BC.
  - B is the alias of AC.
  - C is the alias of AB.

Alias in 2^{4-1} Fractional Factorial Experiment

- In the 2^{4-1} fractional factorial design of experiment, the main factor is aliased with three-way interaction and two-way interaction is aliased with two-way interaction.

  - A is the alias of BCD.
  - B is the alias of ACD.
  - C is the alias of ABD.
  - D is the alias of ABC.
  - AB is the alias of CD.
  - AC is the alias of BD.
  - AD is the alias of BC.

Design Resolution

- Design resolution (i.e., experimental resolution) refers to the confounding patterns indicating how the effects are confounded with others.
- If two factors are aliases, they have confounded effect on the response. There is no need to consider both factors in the model due to the redundancy. Only one factor needs to be included in the model for simplicity.
- Most popular resolutions are resolution III, IV, and V.
  - **Resolution III**: Main effects are aliased with two-way interactions.
  - **Resolution IV**: Main effects are aliased with three-way interactions. Two-way interactions are aliased with other two-way interactions.
  - **Resolution V**: Main effects are aliased with four-way interactions. Two-way interactions are aliased with three-way interactions.

Determine Aliases

- To find the alias of a particular factor, we only need to multiply the factor of interest with the **identity (I)** of the design.
  - In 2^{3-1} fractional factorial design, ABC is the identity. To find the alias of factor A, we use A * ABC = BC.
  - In 2^{4-1} fractional factorial design, ABCD is the identity. To find the alias of factor A, we use A * ABCD = BCD.
  - In 2^{5-1} fractional factorial design, ABCDE is the identity. To find the alias of factor A, we use A * ABCDE = BCDE.

5.0 Control Phase

Black Belt Training: Control Phase

5.1 Lean Controls

5.1.1 Control Methods for 5S

5.1.2 Kanban

5.1.3 Poka-Yoke (Mistake Proofing)

5.2 Statistical Process Control (SPC)

5.2.1 Data Collection for SPC

5.2.2 I-MR Chart

5.2.3 Xbar-R Chart

5.2.4 U Chart

5.2.5 P Chart

5.2.6 NP Chart

5.2.7 X-S chart

5.2.8 CumSum Chart

5.2.9 EWMA Chart

5.2.10 Control Methods

5.2.11 Control Chart Anatomy

5.2.12 Subgroups, Variation, Sampling

5.3 Six Sigma Control Plans

5.3.1 Cost Benefit Analysis

5.3.2 Elements of the Control Plan

5.3.3 Elements of the Response Plan

5.1 Lean Controls

Black Belt Training: Control Phase

5.1 Lean Controls

- 5.1.1 Control Methods for 5S
- 5.1.2 Kanban
- 5.1.3 Poka-Yoke (Mistake Proofing)

5.2 Statistical Process Control (SPC)

- 5.2.1 Data Collection for SPC
- 5.2.2 I-MR Chart
- 5.2.3 Xbar-R Chart
- 5.2.4 U Chart
- 5.2.5 P Chart
- 5.2.6 NP Chart
- 5.2.7 X-S chart
- 5.2.8 CumSum Chart
- 5.2.9 EWMA Chart
- 5.2.10 Control Methods
- 5.2.11 Control Chart Anatomy
- 5.2.12 Subgroups, Variation, Sampling

5.3 Six Sigma Control Plans

- 5.3.1 Cost Benefit Analysis
- 5.3.2 Elements of the Control Plan
- 5.3.3 Elements of the Response Plan

5.1.1 Control Methods for 5S

What is 5S?

- 5S is a systematic method to organize, order, clean, and standardize a workplace…and keep it that way!
  - 5S is a methodology of organizing and improving the work environment.

- 5S is summarized in five Japanese words, all starting with the letter S:
  - **Seiri** (sorting)
  - **Seiton** (straightening)
  - **Seiso** (shining)
  - **Seiketsu** (standardizing)
  - **Shisuke** (sustaining).

- 5S was originally developed in Japan, and is widely used to optimize the workplace to increase productivity and efficiency.

5S Goals

- Reduced waste
- Reduced cost
- Establish a work environment that is:
  - self-explaining
  - self-ordering
  - self-regulating
  - self improving.
  - Where there is/are **no more:**
    - Wandering and/or searching
    - Waiting or delays
    - Secrets hiding spots for tools
    - Obstacles or detours
    - Extra pieces, parts, materials, etc.
    - Injuries
    - Waste.

5S Benefits

- Reduced changeovers
- Reduced defects
- Reduced waste
- Reduced delays
- Reduced injuries
- Reduced breakdowns
- Reduced complaints
- Reduced red ink
- Higher quality
- Lower costs
- Safer work environment
- Greater associate and equipment capacity

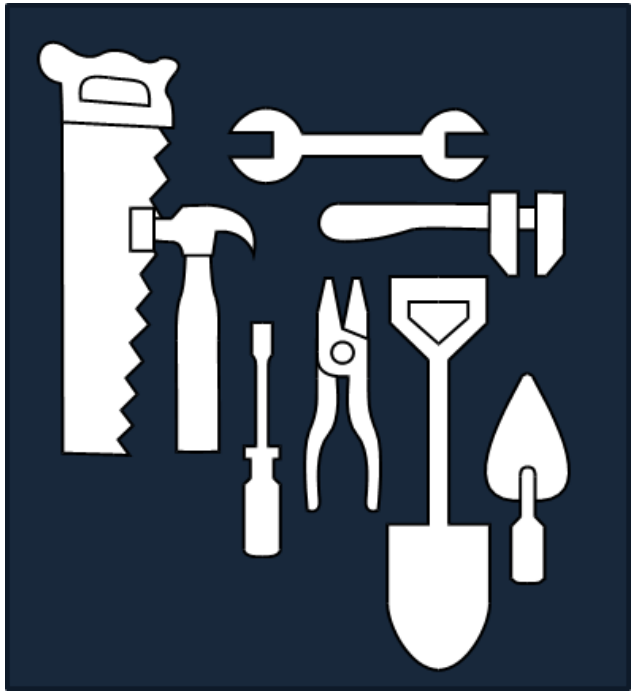
5S Systems Reported Results

- Cut in floor space: 60%
- Cut in flow distance: 80%
- Cut in accidents: 70%
- Cut in rack storage: 68%
- Cut in number of forklifts: 45%
- Cut in machine changeover time: 62%
- Cut in annual physical inventory time: 50%
- Cut in classroom training requirements: 55%
- Cut in nonconformance in assembly: 96%
- Increase in test yields: 50%
- Late deliveries: 0%
- Increase in throughput: 15%

Sorting (Seiri)

- Go through all the tools, parts, equipment, supply, and material in the workplace.
- Categorize them into two major groups: needed and unneeded.
- Eliminate the unneeded items from the workplace. Dispose of or recycle those items.
- Keep the needed items and sort them in the order of priority. **When in doubt…throw it out!**

Straightening (Seiton)



- **Straightening** in 5S is also called **setting in order.**
- Label each needed item.
- Store items at their best locations so that the workers can find them easily whenever they needed any item.
- Reduce the motion and time required to locate and obtain any item whenever it is needed.
- Promote an efficient work flow path.
- Use visual aids like the tool board image on this page.

Shining (Seiso)

- **Shining** in 5S is also called **sweeping**.
- Clean the workplace thoroughly.
- Maintain the tidiness of the workplace.
- Make sure every item is located at the specific location where it should be.
- Create the ownership in the team to keep the work area clean and organized.

Standardizing (Seiketsu)

- **Standardize** the workstation and the layout of tools, equipment, and parts.

- Create identical workstations with a consistent way of storing the items at their specific locations so that workers can be moved around to any workstation any time and perform the same task.

Sustaining (Shisuke)

- **Sustaining** in 5S is also called **self-discipline**.
- Create the culture in the team to follow the first four S's consistently.
- Avoid falling back to the old ways of cluttered and unorganized work environment.
- Keep the momentum of optimizing the workplace.
- Promote innovations of workplace improvement.
- Sustain the first fours S's using:
  - 5S Maps
  - 5S Schedules
  - 5S Job cycle charts
  - Integration of regular work duties
  - 5S Blitz schedules
  - Daily workplace scans.

Simplified Summary of 5S

- **Sort** – “when in doubt, move it out.”
- **Set in Order** – Organize all necessary tools, parts, and components of production. Use visual ordering techniques wherever possible.
- **Shine** – Clean machines and/or work areas. Set regular cleaning schedules and responsibilities.
- **Standardize** – Solidify previous three steps, make 5S a regular part of the work environment and everyday life.
- **Sustain** – Audit, manage, and comply with established five-s guidelines for your business or facility

5.1.2 Kanban

What is Kanban?

- The Japanese word “Kanban” means “signboard.”
- **Kanban system** is a “pull” production scheduling system to determine when to produce, what to produce, and how much to produce based on the demand.
- It was originally developed by Taiichi Ohno in order to reduce the waste in inventory and increase the speed of responding to the immediate demand.

Kanban System

- Kanban system is a demand-driven system.
- The customer demand is the signal to trigger or pull the production.
- Products are made only to meet the immediate demand. When there is no demand, there is no production.
- It is designed to minimize the in-process inventory and to have the right material with the right amount at the right location at the right time.

Kanban System

- Principles of the Kanban System:
  - Only produce products with exactly the same amount that customers consume.
  - Only produce products when customers consume.

- The production is driven by the *actual* demand from the customer side instead of the *forecasted* demand planned by the staff.

Kanban Card

- The **Kanban card** is the ticket or signal to authorize the production or movement of materials. It is the message of asking for more.
- It is sent from the end customer up to the chain of production.
- Upon receiving of a Kanban card, the production station would start to produce goods.
- The Kanban card can be a physical card or an electronic signal.

Kanban System Example

- The simplest example of a Kanban system is the supermarket operation.
- Customers visit the supermarkets and buy what they need.
- The checkout scanners send electronic Kanban cards to the local warehouse asking for more when the items are sold to customers.
- When the warehouse receives the Kanban cards, it starts to replenish the exact goods being sold.
- It the warehouse prepares more than what Kanban cards require, the goods would become obsolete. If it prepares less, the supermarket would not have the goods available when customers need them.

Kanban System Benefits

- Minimize in-process inventory
- Free up space occupied by unnecessary inventory
- Prevent overproduction
- Improve responsiveness to dynamic demand
- Avoid the risk of inaccurate demand forecast
- Streamline the production flow
- Visualize the work flow.

5.1.3 Poka-Yoke

What is Poka-Yoke?

- The Japanese term “poka-yoke” means “mistake-proofing.”
- It is a mechanism to eliminate defects as early as possible in the process.
- It was originally developed by Shigeo Shingo and was initially called “baka-yoke” (fool-proofing).

Two Types of Poka-Yoke

- Prevention
  - Preventing defects from occurring
  - Removing the possibility that an error could occur
  - Making the occurrence of an error impossible.

- Detection
  - Detecting defects once they occur
  - Highlighting defects to draw workers' attention immediately
  - Correcting defects so that they would not reach the next stage.

Three Methods of Poka-Yoke

- Contact Method
  - Use of shape, color, size, or any other physical attributes of the items.
- Constant Number Method
  - Use of a fixed number to make sure a certain number of motions are completed.
- Sequence Method
  - Use of a checklist to make sure all the prescribed process steps are followed in the right order.

Poka-Yoke Devices

- We are surrounded by poka-yoke devices daily.
  - Prevention Devices
    - Example: the dishwasher does not start to run when the door is open.
  - Detection Devices
    - Example: the car starts to beep when the passengers do not buckle their seatbelts.

- Poka-yoke devices can be in any format that can quickly and effectively prevent or detect mistakes.
  - Visual, electrical, mechanical, procedural, human etc.

Steps to Apply Poka-Yoke

- Step 1: Identify the process steps in need of mistake proofing.
- Step 2: Use the 5-why's method to analyze the possible mistakes or failures for the process step.
- Step 3: Determine the type of poka-yoke: prevention or detection.
- Step 4: Determine the method of poka-yoke: contact, constant number, or sequence.
- Step 5: Pilot the poka-yoke approach and make any adjustments if needed.
- Step 6: Implement poka-yoke in the operating process and maintain the performance.

5.2 Statistical Process Control

Black Belt Training: Control Phase

5.1 Lean Controls

5.1.1 Control Methods for 5S

5.1.2 Kanban

5.1.3 Poka-Yoke (Mistake Proofing)

5.2 Statistical Process Control (SPC)

5.2.1 Data Collection for SPC

5.2.2 I-MR Chart

5.2.3 Xbar-R Chart

5.2.4 U Chart

5.2.5 P Chart

5.2.6 NP Chart

5.2.7 X-S chart

5.2.8 CumSum Chart

5.2.9 EWMA Chart

5.2.10 Control Methods

5.2.11 Control Chart Anatomy

5.2.12 Subgroups, Variation, Sampling

5.3 Six Sigma Control Plans

5.3.1 Cost Benefit Analysis

5.3.2 Elements of the Control Plan

5.3.3 Elements of the Response Plan

5.2.1 Data Collection for SPC

What is SPC?

- **Statistical process control** (SPC) is a statistical method to monitor the performance of a process using control charts in order to keep the process in statistical control.

- Statistical process control can be used to distinguish between special cause variation and common cause variation in the process.

- It presents the voice of the process.

Common Cause Variation

- **Common cause variation** (also called chance variation) is the inherent natural variation in any processes.
- It is the random background noise, which cannot be controlled or eliminated from the process.
- Its presence in the process is expected and acceptable due to its relatively small influence on the process.

Special Cause Variation

- **Special cause variation** (also called assignable cause variation) is the unnatural variation in the process.
- It is the cause of process instability and leads to defects of the products or services.
- It is the signal of unanticipated change (either positive or negative) in the process.
- It is possible to eliminate the special cause variation from the process.

Process Stability

- A process is **stable** when:
  - There is not any special cause variation involved in the process
  - The process is in statistical control
  - The future performance of the process is predictable within certain limits
  - The changes happening in the process are all due to random inherent variation
  - There are not any trends, unnatural patterns, and outliers in the control chart of the process.

SPC Benefits

- Statistical process control can be used in different phases of Six Sigma projects to:
  - Understand the stability of a process
  - Detect the special cause variation in the process
  - Identify the statistical difference between two phases
  - Eliminate or apply the unnatural change in the process
  - Improve the quality and productivity.

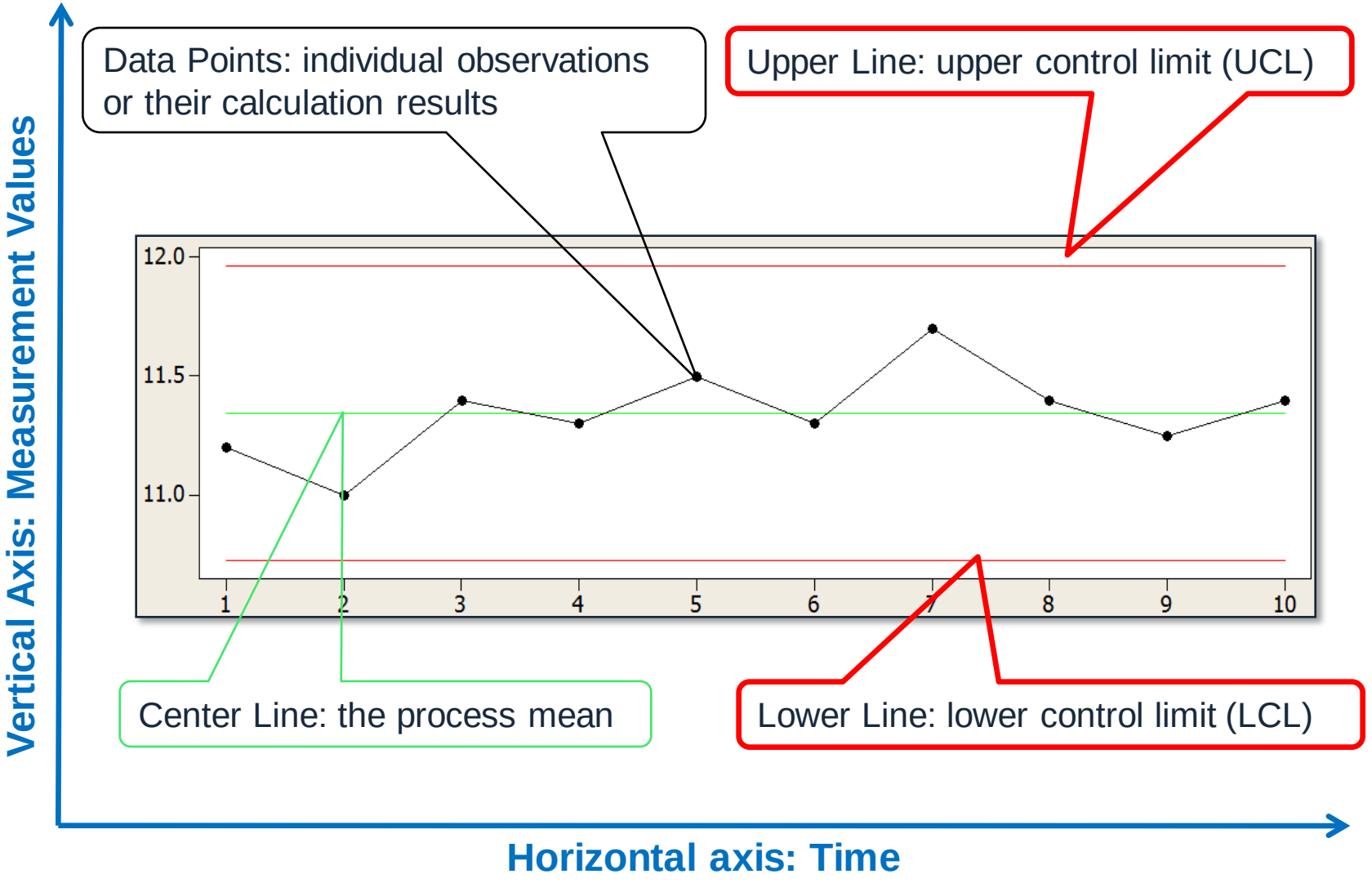
Control Charts

- **Control charts** are graphical tools to present and analyze the process performance in statistical process control.

- Control charts are used to detect special cause variation and determine whether the process is in statistical control (stable).

- Variation solutions:
  - Minimize the common cause variation
  - Eliminate the special cause variation when it leads to unanticipated negative changes in the outcome
  - Implement the special cause variation when it leads to unanticipated positive changes in the outcome.

Control Charts Elements

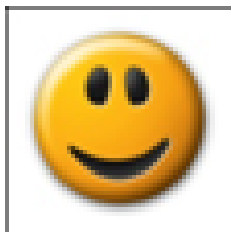
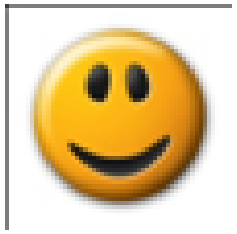


Control Charts Elements

- Control charts can work for both continuous data and discrete or count data.
- Control limits are approximately three sigma away from the process mean.
- A process is in statistical control when all the data points on the control charts fall within the control limits and have random patterns only.
- Otherwise, the process is out of control and we need to investigate the special cause variation in the process.

Possible Errors in SPC

- There are two types of possible errors in interpreting controls charts.

		Interpretation	
		Common Cause	Special Cause Variation
Truth	Common Cause Variation		Type I Error (False Positive)
	Special Cause Variation	Type II Error (False Negative)	

Possible Errors in SPC

- It is similar to the way of defining the type I and type II errors in hypothesis testing.

- Control charts can be interpreted as a way of testing the hypothesis about the process stability.
  - Null Hypothesis (H_0): The process is stable (i.e., in statistical control).
  - Alternative Hypothesis (H_A): The process is unstable (i.e., out of statistical control).

Possible Errors in SPC

- Type I Error
  - False positive
  - False alarm
  - Considering true common cause variation as special cause variation
  - Type I errors waste resources spent on investigation.

- Type II Error
  - False negative
  - Miss
  - Considering true special cause variation as common cause variation
  - Type II errors neglect the need to investigate critical changes in the process.

Data Collection Considerations

- To collect data for plotting control charts, we need to consider:
  - What is the measurement of interest?
  - Are the data discrete or continuous?
  - How many samples do we need?
  - How often do we sample?
  - Where do we sample?
  - What is the sampling strategy?
  - Do we use the raw data collected or transfer them to percentages, proportions, rates, etc.?

Subgroups and Rational Subgrouping

- When sampling, we randomly select a group of items (i.e. a subgroup) from the population of interest.
- The *subgroup size* is the count of samples in a subgroup. It can be constant or variable.
- Depending on the subgroup sizes, we select different control charts accordingly.
- *Rational subgrouping* is the basic sampling scheme in SPC.
- The goal of rational subgrouping is to maximize the likelihood of detecting special cause variation. In other words, the control limits should only reflect the variation *between* subgroups.
- The number of subgroups, subgroup size, and frequency of sampling have great impact on the quality of control charts.

Impact of Variation

- The rational subgrouping strategy is designed to minimize the opportunity of having special cause variation *within* subgroups.

- If there is only random variation (background noise) within subgroups, all the special cause variation would be reflected between subgroups. It is easier to detect an out-of-control situation.

- Random variation is inherent and indelible in the process. We are more interested in identifying and taking actions on special cause variation.

Frequency of Sampling

- The frequency of sampling in SPC depends on whether we have sufficient data to signal the changes in a process with reasonable time and costs.
- The more frequently we sample, the higher costs it may trigger.
- We need the subject matter experts' knowledge on the nature and characteristics of the process to make good decisions on sampling frequency.

5.2.2 I-MR Chart

I-MR Chart

- The **I-MR chart** (also called individual-moving range chart or IR chart) is a popular control chart for continuous data with subgroup size equal to one.
- The I chart plots an individual observation as a data point.
- The MR chart plots the absolute value of the difference between two consecutive observations in individual charts as a data point.
- If there are *n* data points in the I chart, there are *n –1* data points in the MR chart.
- The I chart is valid only if the MR chart is in control.
- The underlying distribution of the I-MR chart is normal distribution.

I Chart Equations

- I Chart (Individuals Chart)
  - Data Point: x_i
  - Center Line: $\frac{\sum_{i=1}^{n} x_i}{n}$
  - Control Limits: $\frac{\sum_{i=1}^{n} x_i}{n} \pm 2.66 \times \overline{MR}$

where *n* is the number of observations

MR-Chart Equations

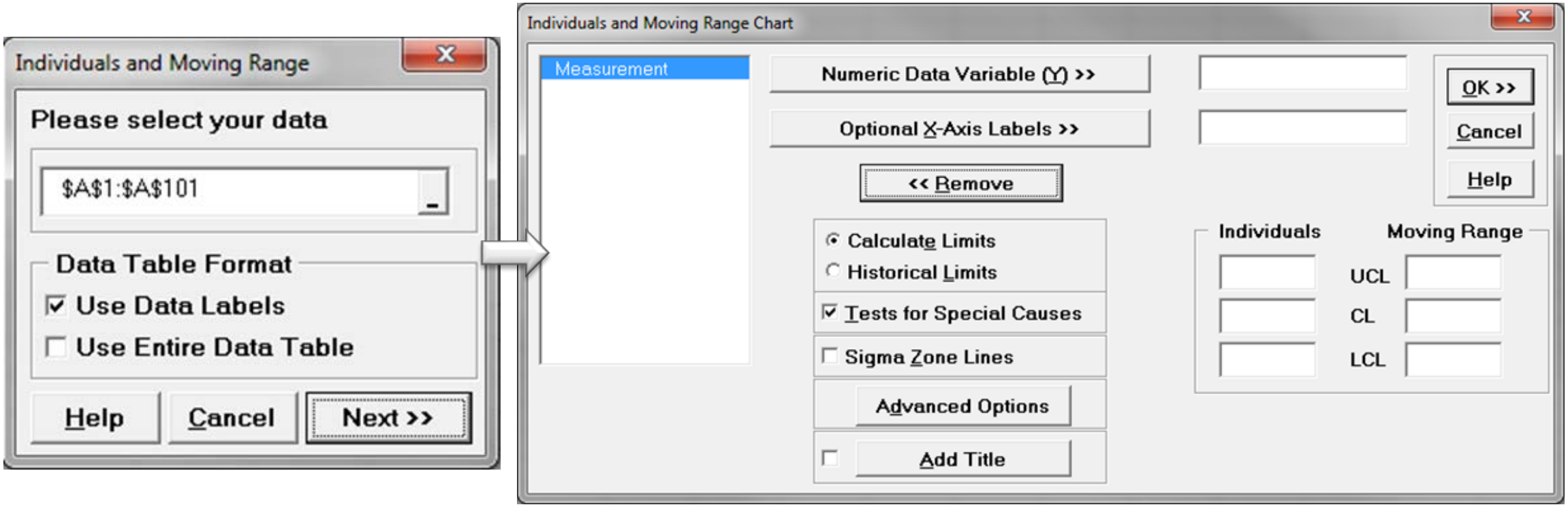
- MR Chart (Moving Range Chart)
  - Data Point: $\left|x_{i+1} - x_i\right|$
  - Center Line: $\dfrac{\left|x_{i+1} - x_i\right|}{n-1}$
  - Upper Control Limit: $3.267 \times \dfrac{\left|x_{i+1} - x_i\right|}{n-1}$
  - Lower Control Limit: 0

where *n* is the number of observations

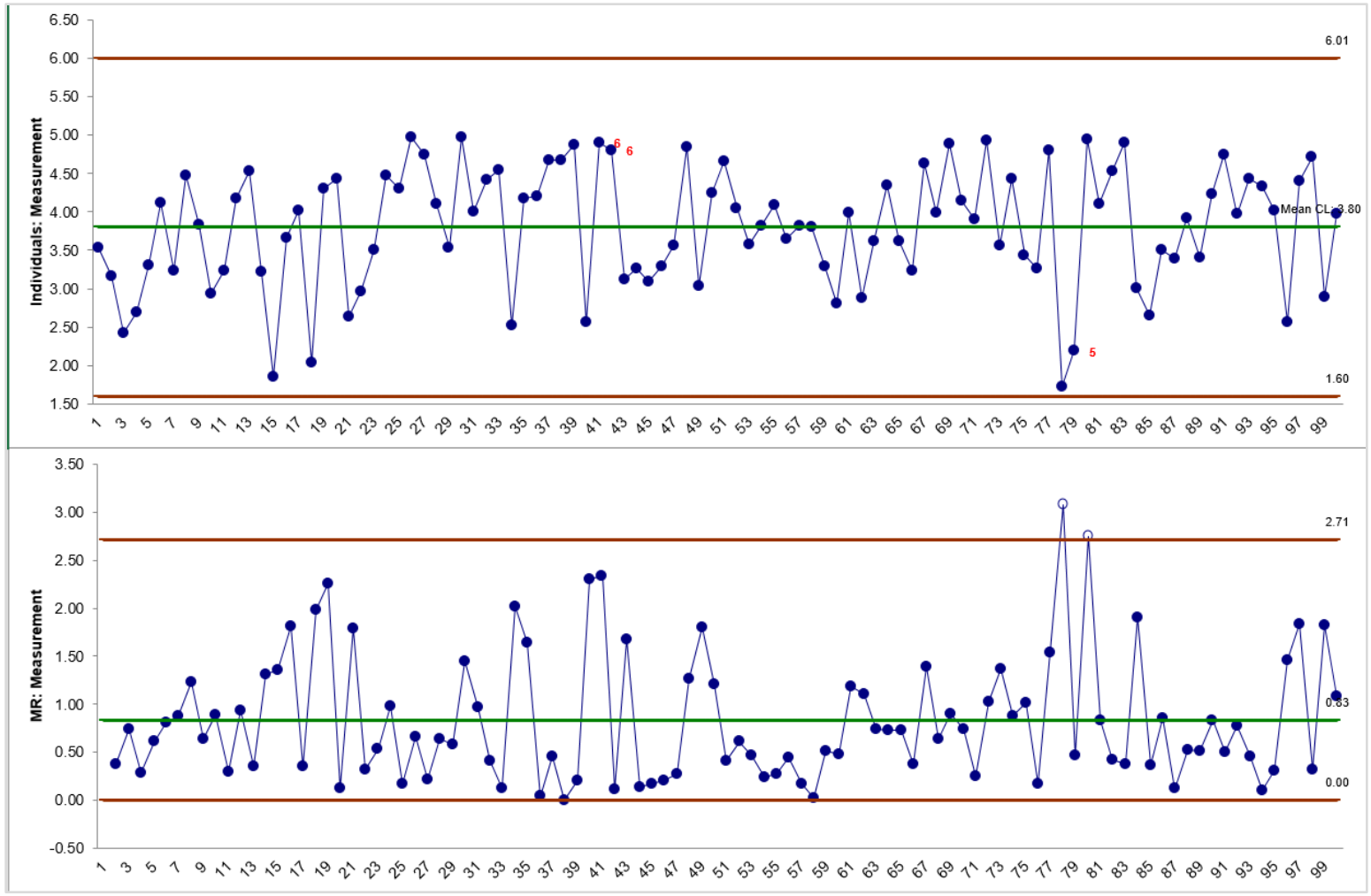
Use SigmaXL to Plot I-MR Charts

- Data File: “IR” tab in “Sample Data.xlsx”
- Steps in SigmaXL to plot IR charts
  - Select the entire range of data
  - Click SigmaXL -> Control Charts -> Individuals & Moving Range
  - A new window named “Individuals and Moving Range” appears with the selected range automatically populated into the box below “Please select your data”.
  - Click “Next>>”
  - A new window named “Individuals and Moving Range Chart” pops up.
  - Select “Measurement” as the “Numeric Data Variable (Y)”
  - Check the checkbox of “Test for special causes”
  - Click “OK>>”
  - The IR charts appear in the newly generated tab “Indiv & MR Charts (1)”.

Use SigmaXL to Plot I-MR Charts



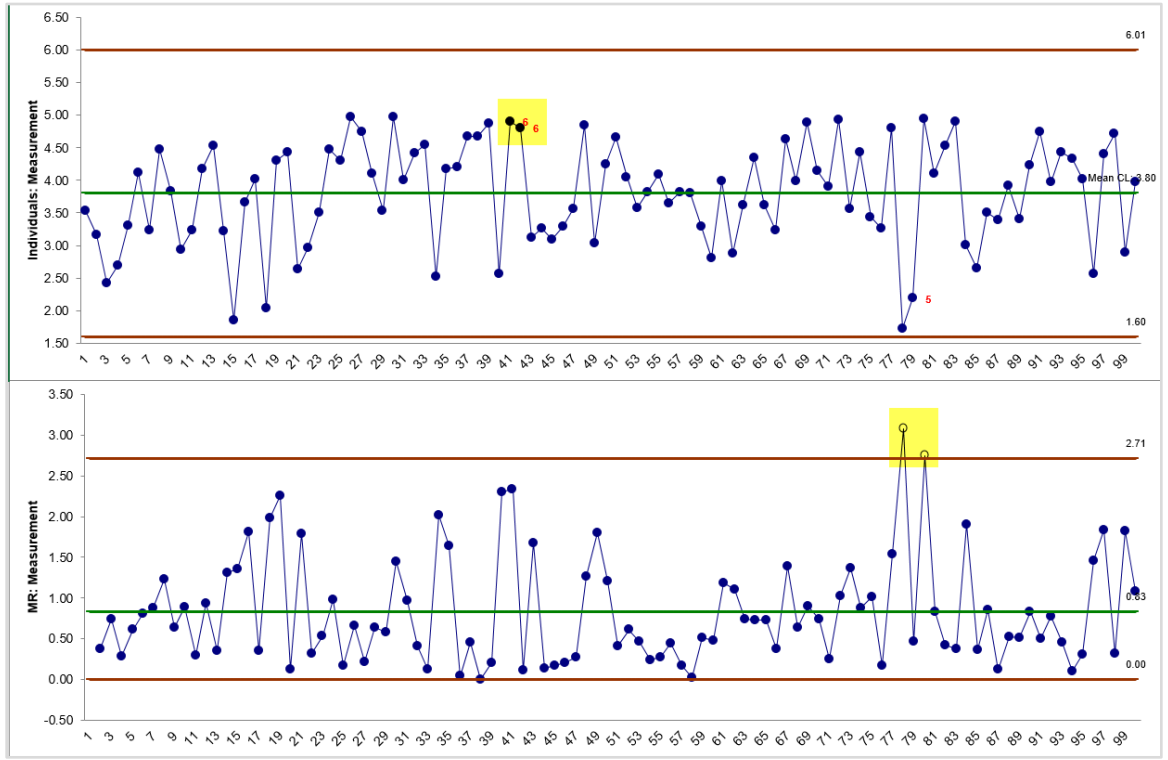
Use SigmaXL to Plot I-MR Charts



I-MR Charts Diagnosis

I Chart (Individuals Chart):

Since the MR Chart is out of control, the I chart is invalid.



MR Chart (Moving Range Chart):

Two data points fall beyond the upper control limit. It indicates the MR chart is out of control (i.e. the variations between every two contiguous individual samples are not stable over time). We need to further investigate the process, identify the root causes which trigger the outliers, and correct or apply the root causes to bring the process back to control.

5.2.3 Xbar-R Chart

Xbar-R Chart

- The **Xbar-R chart** is a control chart for continuous data with a constant subgroup size between two and ten.
- The Xbar chart plots the average of a subgroup as a data point.
- The R chart plots the difference between the highest and lowest values within a subgroup as a data point.
- The Xbar chart monitors the process mean and the R chart monitors the variation within subgroups.
- The Xbar is valid only if the R chart is in control.
- The underlying distribution of the Xbar-R chart is normal distribution.

Xbar Chart Equations

- Xbar chart
  - Data Point: $\overline{X}_i = \frac{\sum_{j=1}^{m} x_{ij}}{m}$
  - Center Line: $\overline{\overline{X}} = \frac{\sum_{i=1}^{k} \overline{X}_i}{k}$
  - Control Limits: $\overline{\overline{X}} \pm A_2 \overline{R}$

where *m* is the subgroup size and *k* is the number of subgroups. A_2 is a constant depending on the subgroup size.

R Chart Equations

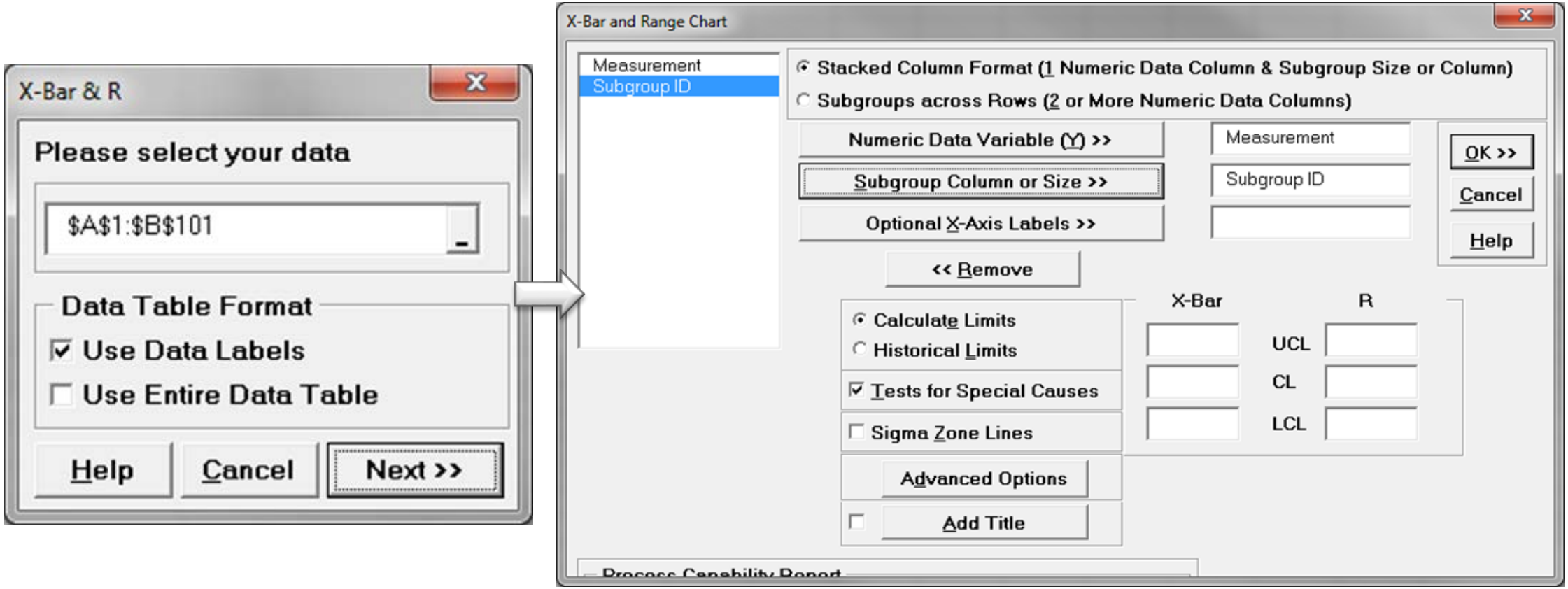
- R chart (Rage Chart)
  - Data Point: $R_i = \underset{j\in[1,m]}{Max}(x_{ij}) - \underset{j\in[1,m]}{Min}(x_{ij})$
  - Center Line: $\overline{R} = \frac{\sum_{i=1}^{k} R_i}{k}$
  - Upper Control Limit: $D_4 \times \overline{R}$
  - Lower Control Limit: $D_3 \times \overline{R}$

    where *m* is the subgroup size and *k* is the number of subgroups. D_3 and D_4 are constants depending on the subgroup size.

Use SigmaXL to Plot Xbar-R Charts

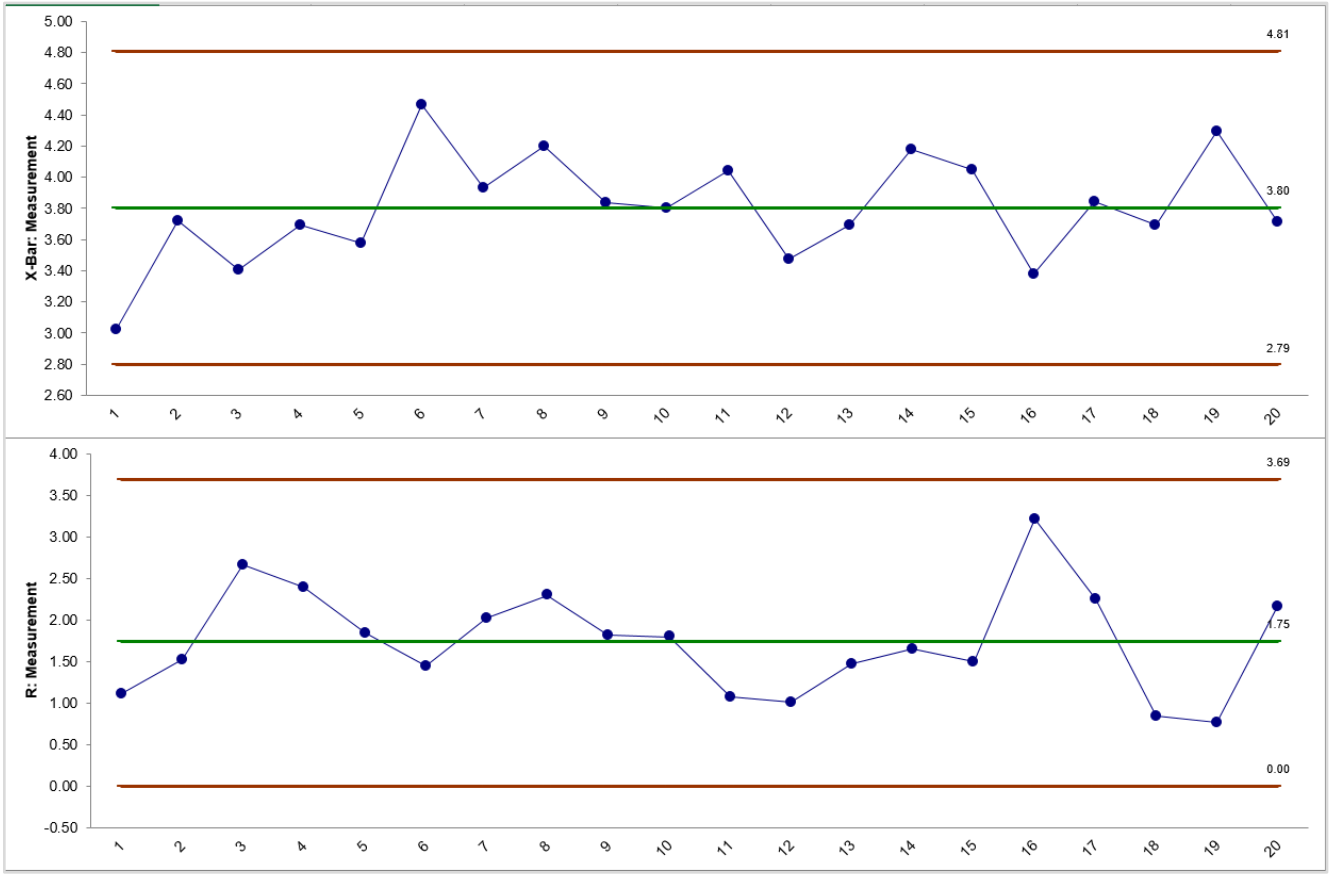
- Data File: “Xbar-R” tab in “Sample Data.xlsx”
- Steps in SigmaXL to plot Xbar-R charts
  - Select the entire range of the data
  - Click SigmaXL -> Control Charts -> X-Bar & R
  - A new window named “X-Bar & R” appears with the selected range automatically populated into the box below “Please select your data”.
  - Click “Next>>”
  - A new window named “X-Bar and Range Chart” pops up.
  - Select the “Measurement” as the “Numeric Data Variables (Y)”
  - Select the “Subgroup ID” as the “Subgroup Column or Size”
  - Check the checkbox of “Tests for Special Causes”
  - Click “OK>>”
  - The Xbar-R charts appear in the newly generated tab “Indiv & MR Charts (1)”.

Use SigmaXL to Plot Xbar-R Charts



Xbar-R Charts Diagnosis

Xbar-R Charts:
Since the R Chart is in control, the Xbar chart is valid. In both charts, there aren't any data points failing any tests for special causes (i.e. all the data points fall between the control limits and spread around the center line with a random pattern). We conclude that the process is in control.



5.2.4 U Chart

Defect vs. Defective

- A **defect** of a unit is the unit's characteristic that does not meet the customers' requirements.
- A **defective** is a unit that is not acceptable to the customers.
- One defective might have multiple defects.
- One unit might have multiple defects but be still usable to the customers.

U Chart

- The **U chart** is a control chart monitoring the average defects per unit.
- The U chart plots the count of defects per unit of a subgroup as a data point.
- It considers the situation when the subgroup size of inspected units for which the defects would be counted is not constant.
- The underlying distribution of the U chart is Poisson distribution.

U Chart Equations

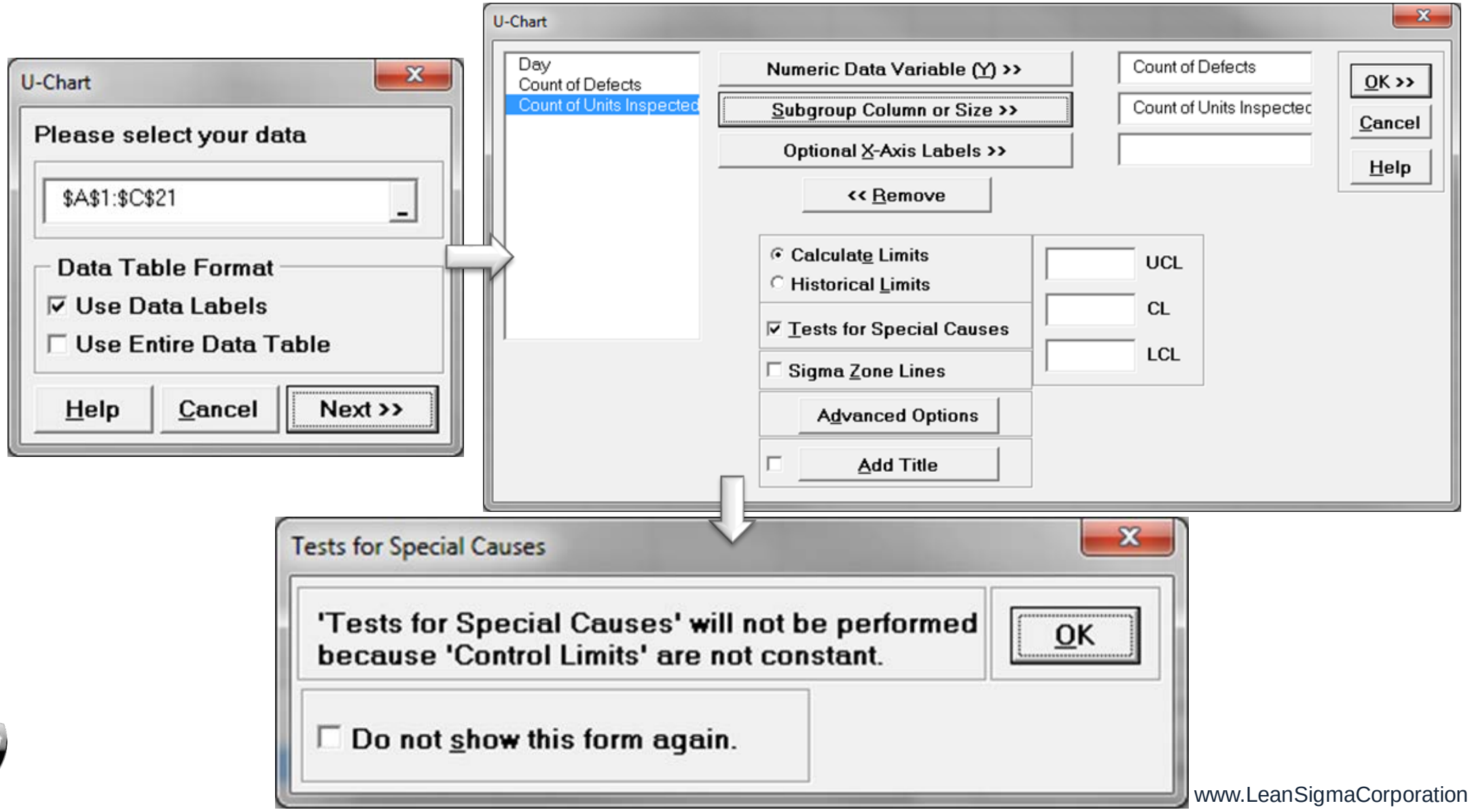
- U chart
  - Data Point: $u_i = \frac{x_i}{n_i}$
  - Center Line: $\bar{u} = \frac{\sum_{i=1}^{k} u_i}{k}$
  - Control Limits: $\bar{u} \pm 3 \times \sqrt{\frac{\bar{u}}{n_i}}$

where n_i is the subgroup size for the i^{th} subgroup;
k is the number of subgroups;
x_i is the number of defects in the i^{th} subgroup.

Use SigmaXL to Plot a U Chart

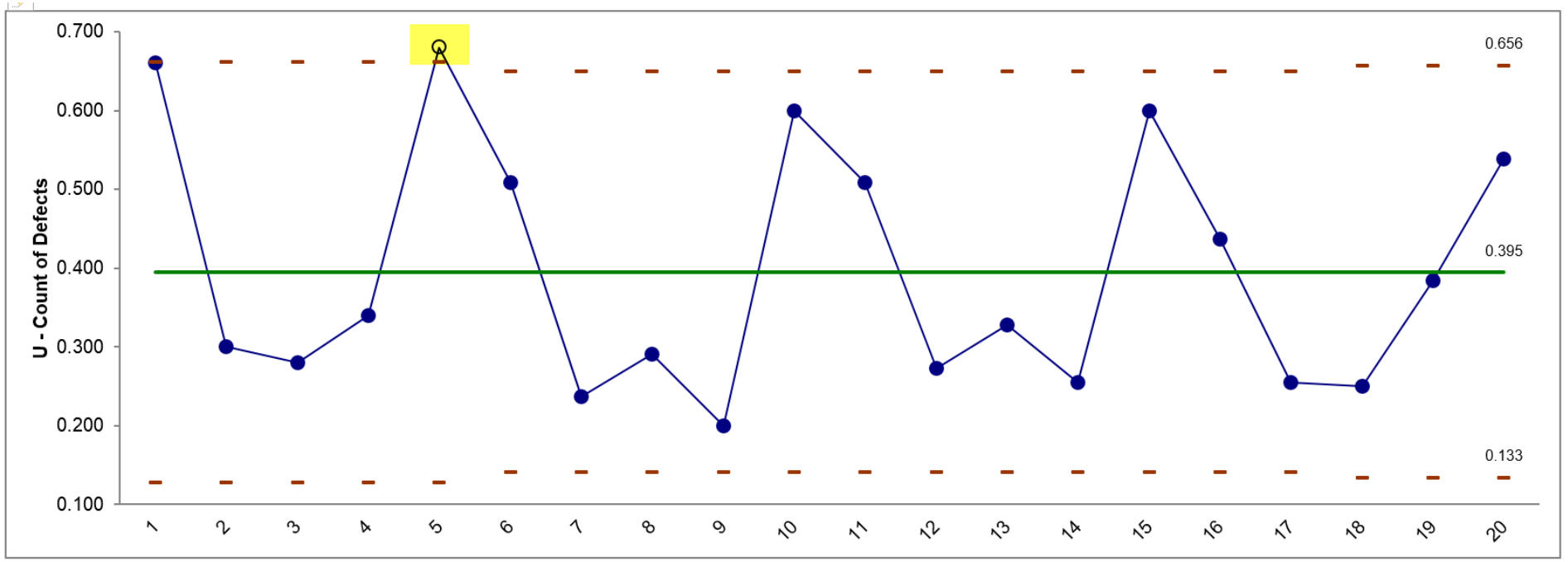
- Data File: “U” tab in “Sample Data.xlsx”
- Steps in SigmaXL to plot a U chart
  - Select the entire range of the data
  - Click SigmaXL -> Control Charts -> Attribute Charts -> U
  - A new window named “U-Chart” appears with the selected range of the data automatically populated into the box below “Please select your data”.
  - Click “Next>>”
  - A new window named “U-Chart” pops up.
  - Select “Count of Defects” as the “Numeric Data Variable (Y)”
  - Select “Count of Units Inspected” as the “Subgroup Column or Size”
  - Check the checkbox “Test for Special Causes”
  - Click “OK>>”
  - A new window named “Tests for Special Causes” pops up. Click “OK” to proceed.
  - The U chart appears in the newly generated tab “U-Chart (1)”.

Use SigmaXL to Plot a U Chart



U Chart Diagnosis

U Chart:
Since the sample sizes are not constant over time, the control limits are adjusted to different values accordingly. The highlighted data point falls beyond the upper control limit. We conclude that the process is out of control. Further investigation is needed to determine the special causes which triggered the unnatural pattern of the process.



5.2.5 P Chart

P Chart

- The **P chart** is a control chart monitoring the percentages of defectives.
- The P chart plots the percentage of defectives in one subgroup as a data point.
- It considers the situation when the subgroup size of inspected units is not constant.
- The underlying distribution of the P chart is binomial distribution.

P Chart Equations

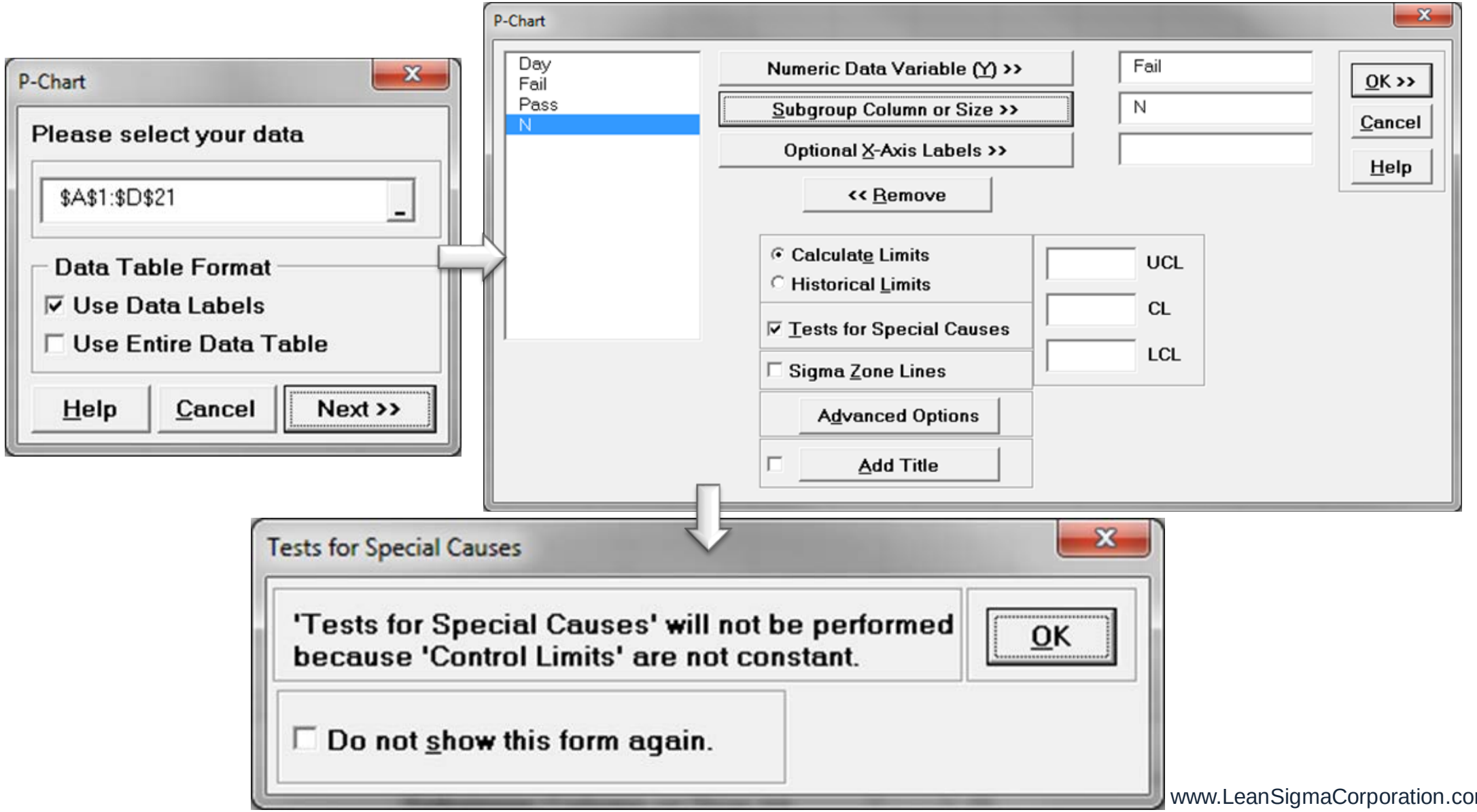
- P chart
  - Data Point: $p_i = \frac{x_i}{n_i}$
  - Center Line: $\bar{p} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} n_i}$
  - Control Limits: $\bar{p} \pm 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}$

where n_i is the subgroup size for the i^{th} subgroup;
k is the number of subgroups;
x_i is the number of defectives in the i^{th} subgroup.

Use SigmaXL to Plot a P Chart

- Data File: “P” tab in “Sample Data.xlsx”
- Steps in SigmaXL to plot a P chart
  - Select the entire range of the data
  - Click SigmaXL -> Control Chart -> Attribute Charts -> P
  - A new window named “P-Chart” appears with the selected range automatically populated into the box below “Please select your data”.
  - Click “Next>>”
  - A new window also named “P-Chart” pops up.
  - Select “Fail” as the “Numeric Data Variables (Y)”
  - Select “N” as the “Subgroup Column or Size”
  - Check the checkbox of “Test for Special Causes”
  - Click “OK>>”
  - A new window named “Tests for Special Causes” pops up. Click “OK” to proceed.
  - The P chart appears in the newly generated tab “P-Chart (1)”.

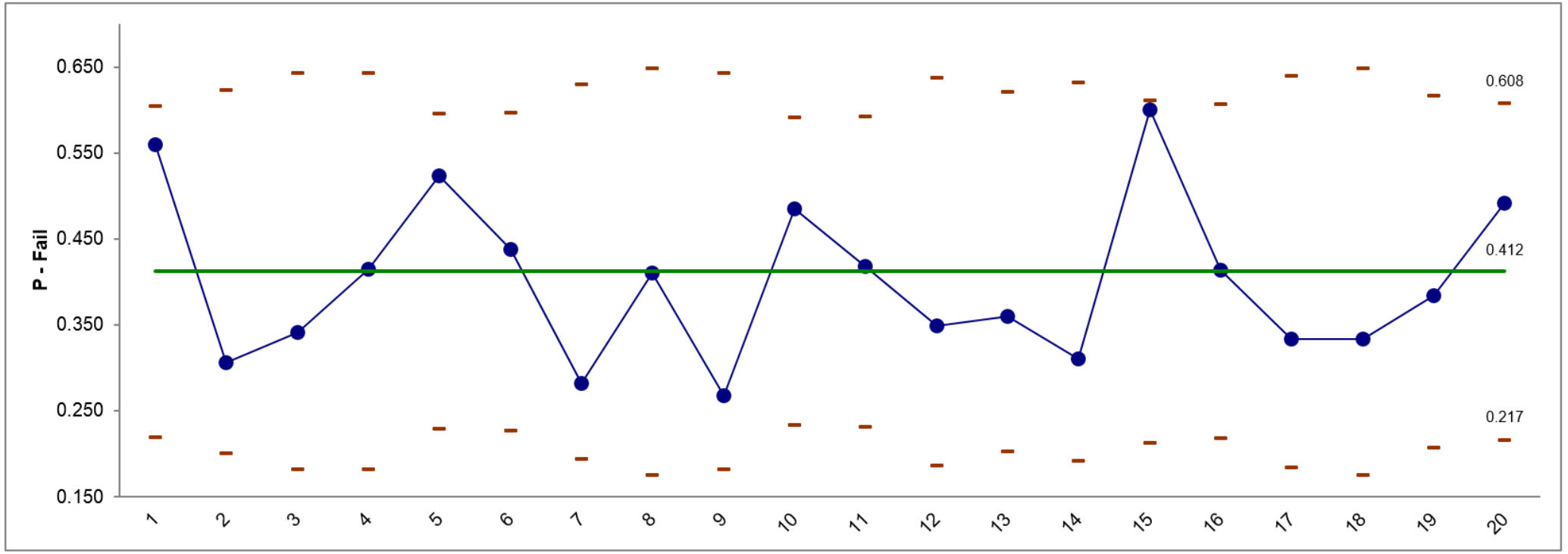
Use SigmaXL to Plot a P Chart



P Chart Diagnosis

P Chart:

Since the sample sizes are not constant over time, the control limits are adjusted to different values accordingly. All the data points fall within the control limits and spread randomly around the mean. We conclude that the process is in control.



5.2.6 NP Chart

NP Chart

- The **NP chart** is a control chart monitoring the count of defectives.
- The NP chart plots the number of defectives in one subgroup as a data point.
- The subgroup size of the NP chart is constant.
- The underlying distribution of the NP chart is binomial distribution.

NP Chart Equations

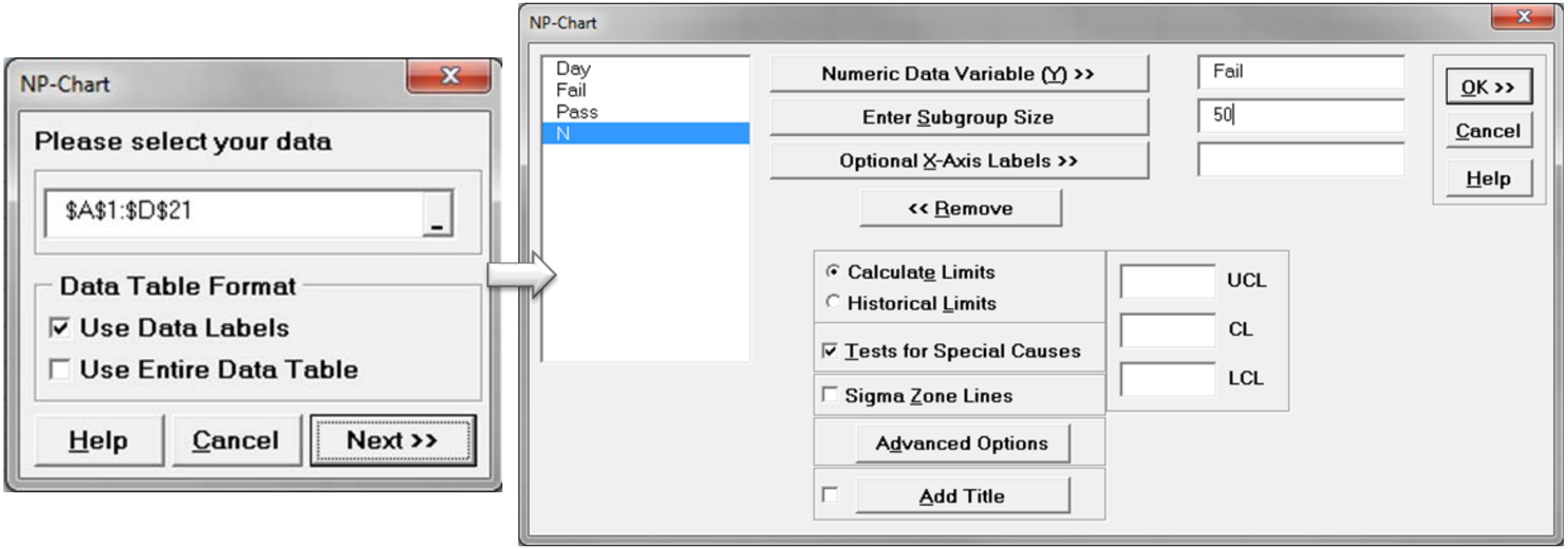
- NP chart
  - Data Point: x_i
  - Center Line: $n\bar{p} = \frac{\sum_{i=1}^{k} x_i}{k}$
  - Control Limits: $n\bar{p} \pm 3 \times \sqrt{n\bar{p}(1-\bar{p})}$

where n is the constant subgroup size;
k is the number of subgroups;
x_i is the number of defectives in the i^{th} subgroup.

Use SigmaXL to Plot an NP Chart

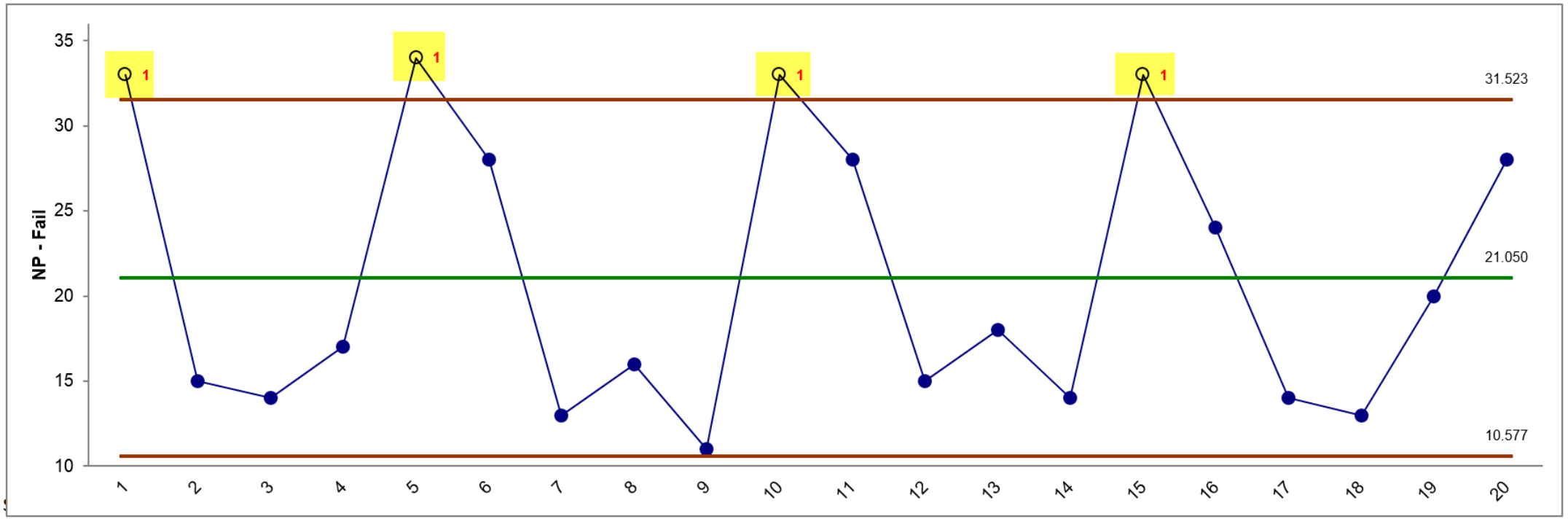
- Data File: “NP” tab in “Sample Data.xlsx”
- Steps in SigmaXL to plot a NP chart
  - Select the entire range of data
  - Click SigmaXL -> Control Charts -> Attribute Charts -> NP
  - A new window named “NP-Chart” appears with the selected range automatically populated into the box under “Please select your data”.
  - Click “Next>>”
  - A new window named “NP-Chart” pops up.
  - Select “Fail” as the “Numeric Data Variable (Y)”
  - Enter “50” as the “Subgroup Size”
  - Check the checkbox of “Tests of Special Causes”
  - Click “OK>>”
  - The NP chart appears in the newly generated tab “NP-Chart (1)”.

Use SigmaXL to Plot an NP Chart



NP Chart Diagnosis

NP Chart:
Four data points in the red circles fall beyond the upper control limit. We conclude that the NP chart is out of control. Further investigation is needed to determine the special causes which triggered the unnatural pattern of the process.



5.2.7 X-S Chart

X-S Chart

- The **X-S chart** (also called Xbar-S chart) is a control chart for continuous data with a constant subgroup size greater than ten.
- The Xbar chart plots the average of a subgroup as a data point.
- The S chart plots the standard deviation within a subgroup as a data point.
- The Xbar chart monitors the process mean and the S chart monitors the variability within subgroups.
- The Xbar is valid only if the S chart is in control.
- The underlying distribution of the Xbar-S chart is normal distribution.

Xbar Chart Equations

- Xbar Chart
  - Data Point: $\overline{X}_i = \frac{\sum_{j=1}^{m} x_{ij}}{m}$
  - Center Line: $\overline{\overline{X}} = \frac{\sum_{i=1}^{k} \overline{X}_i}{k}$
  - Control Limits: $\overline{\overline{X}} \pm A_3 \bar{s}$

where m is the subgroup size and k is the number of subgroups. A_3 is a constant depending on the subgroup size.

S Chart Equations

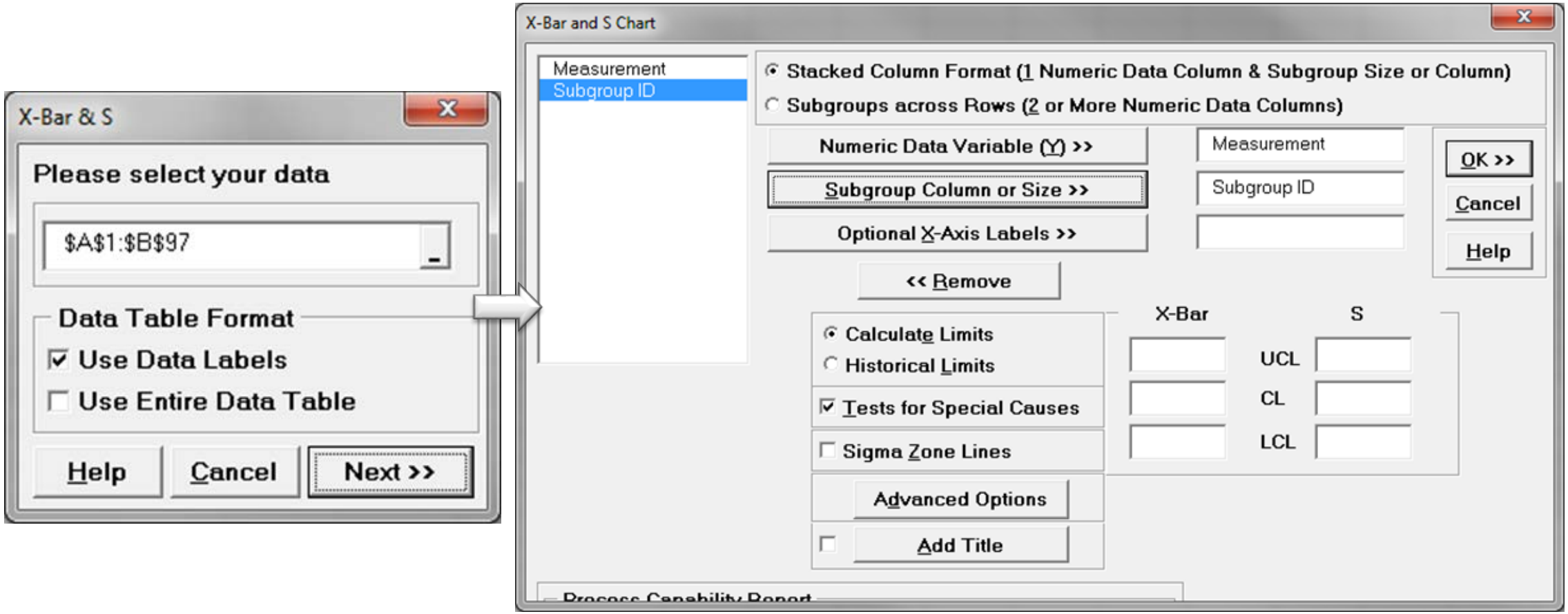
- S chart
  - Data Point: $s_i = \sqrt{\frac{\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2}{m-1}}$
  - Center Line: $\bar{s} = \frac{\sum_{i=1}^{k} s_i}{k}$
  - Upper Control Limit: $B_4 \times \bar{s}$
  - Lower Control Limit: $B_3 \times \bar{s}$

where *m* is the subgroup size and *k* is the number of subgroups.
B_3 and B_4 are constants depending on the subgroup size.

Use SigmaXL to Plot Xbar-S Charts

- Data File: “Xbar-S” tab in “Sample Data.xlsx”
- Steps in SigmaXL to plot Xbar-S charts
  - Select the entire range of the data
  - Click SigmaXL -> Control Charts -> X-Bar & S
  - A new window named “X-Bar & S” appears with the selected range automatically populated into the box below “Please select your data”
  - Click “Next>>”
  - A new window named “X-Bar & S chart” pops up.
  - Select “Measurement” as the “Numeric Data Variable (Y)”
  - Select “Subgroup ID” as “Subgroup Column or Size”
  - Check the checkbox “Tests for Special Causes”
  - Click “OK>>”
  - The Xbar-S charts appear in the newly generated tab “X-Bar & S Charts (1)”.

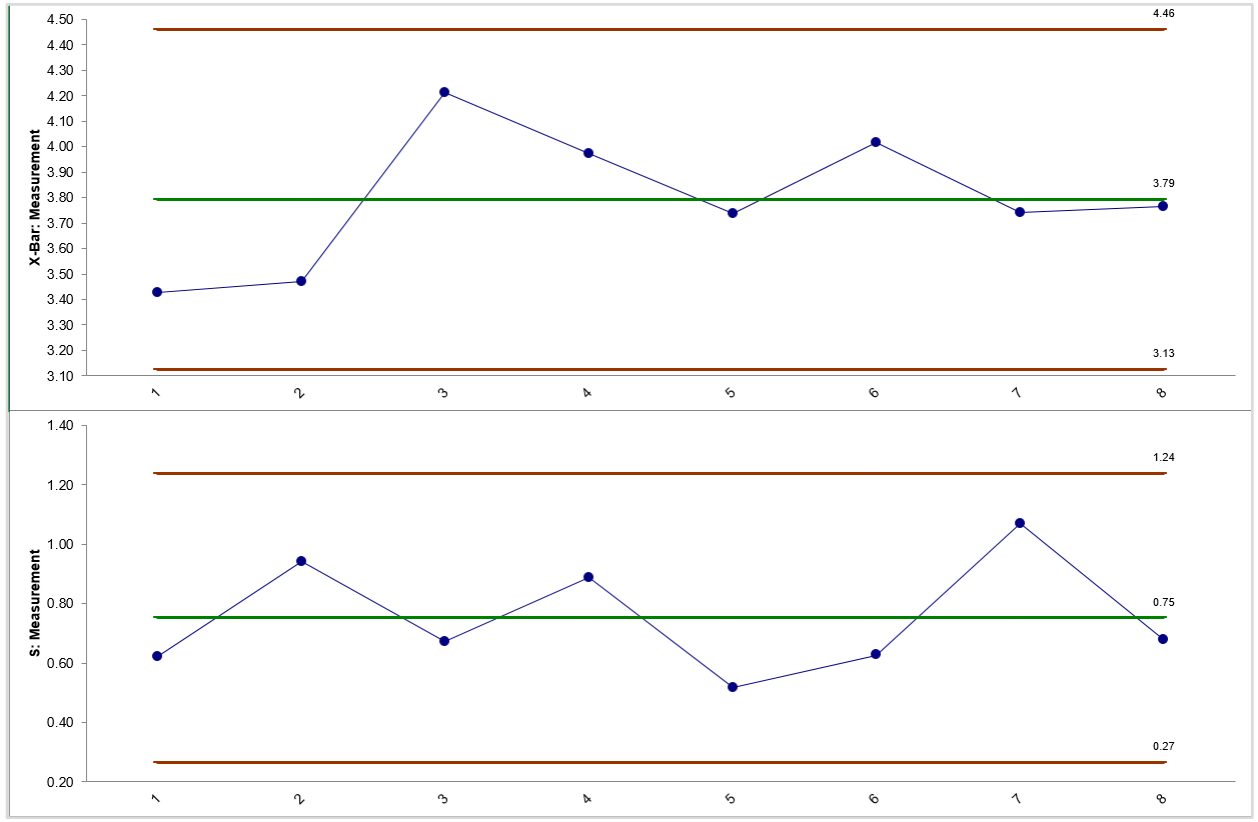
Use SigmaXL to Plot Xbar-S Charts



Xbar-S Charts Diagnosis

Xbar-S Charts:

Since the S Chart is in control, the Xbar chart is valid. In both charts, there aren't any data points failing any tests for special causes (i.e. all the data points fall between the control limits and spread around the center line with a random pattern). We conclude that the process is in control.



5.2.8 CumSum Chart

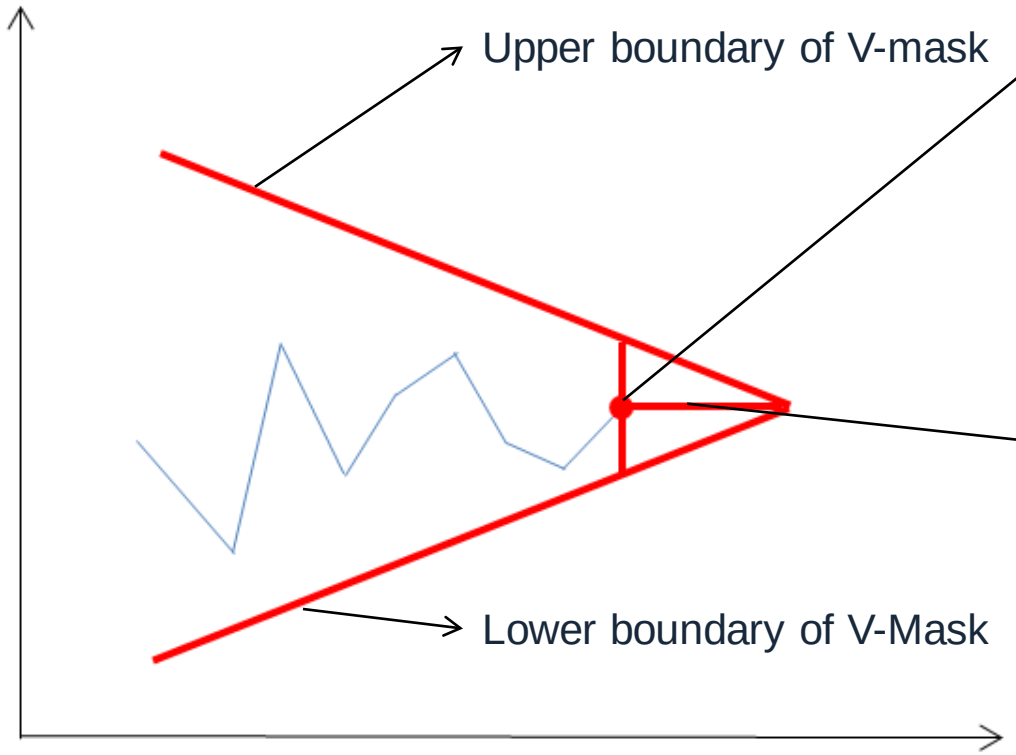
CumSum Chart

- The **CumSum chart** (also called cumulative sum control chart or CUSUM chart) is a control chart of monitoring the cumulative sum of the subgroup mean deviations from the process target.
- It detects the shift of the process mean from the process target over time.
- The underlying distribution of the CumSum chart is normal distribution.
- There are two types of CumSum charts:
  - One two-sided CumSum charts
  - Two one-sided CumSum charts.

Two-Sided CumSum

- In two-sided CumSum charts, we use a V-mask to identify out-of-control data points.
- The V-mask is a transparent overlay shape of a horizontal “V” applied on the top of the CumSum chart. Its origin point is placed on the last data point of the CumSum chart and its center line is horizontal.
- If all of the data points stay inside the V-mask, we consider the process is in statistical control.

V-Mask



Origin point on the V-mask, which is applied right on the last data point of the CumSum chart.

The center line of the V-mask, which should be parallel to the x axis (time).

If any data points in the CumSum chart to the left of the origin point are outside the V-mask, the process is considered out of statistical control.

Two-Sided CumSum Equations

- Two-Sided CumSum
  - Data Point: $\begin{cases} c_i = c_{i-1} + (\bar{x}_i - T) & i > 0 \\ c_i = 0 & i = 0 \end{cases}$
  - V-Mask Slope: $k\frac{\sigma}{\sqrt{m}}$
  - V-Mask Width at the Origin Point: $2h\frac{\sigma}{\sqrt{m}}$

where $\bar{x}_i$ is the mean of the i[th] subgroup;
T is the process target;
σ is the estimation of process standard deviation;
m is the subgroup size.

One-Sided CumSum

- We can also use two one-sided CumSum charts to detect the shift of the process mean from the process target.
- The upper one-sided CumSum detects the upward shifts of the process mean.
- The lower one-sided CumSum detects the downward shifts of the process mean.

One-Sided CumSum Equations

- One-Sided CumSum
  - Data Point: $\begin{cases} c_i = c_{i-1} + (\bar{x}_i - T) & i > 0 \\ c_i = 0 & i = 0 \end{cases}$
  - Center Line: T
  - Upper Control Limit: $c_i^+ = \max\left(0, \bar{x}_i - (T + k) + c_{i-1}^+\right)$
  - Lower Control Limit: $c_i^- = \max\left(0, (T - k) - \bar{x}_i + c_{i-1}^-\right)$

where $\bar{x}_i$ is the mean of the i[th] subgroup;
T is the process target;
k is the slope of the lower boundary of the V-mask.

5.2.9 EWMA Chart

EWMA Chart

- The **EWMA chart** (Exponentially-Weighted Moving Average Chart) is a control chart monitoring the exponentially-weighted average of previous and present subgroup means.
- The more recent data get more weight than more historical data.
- It detects the shift of the process mean from the process target over time.
- The underlying distribution of the EWMA chart is normal distribution.

EWMA Chart Equations

- EWMA Chart
  - Data Point: $z_i = \lambda \bar{x}_i + (1-\lambda) z_{i-1}$ where $0 < \lambda < 1$
  - Center Line: $\overline{\overline{X}}$
  - Control Limits: $\overline{\overline{X}} \pm k \cdot \frac{s}{\sqrt{n}} \sqrt{\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2i}\right]}$

where $\bar{x}_i$ is the mean of the i^{th} subgroup;
λ and k are user-defined parameters to calculate the EWMA data points and the control limits.

5.2.10 Control Methods

Control Methods

- There are many control methods available to keep the process stable and minimize the variation.

- Most common control methods:
  - SPC (statistical process control)
  - 5S method
  - Kanban
  - Poka-Yoke (mistake proofing).

SPC

- **SPC (Statistical Process Control)** is a quantitative control method to monitor the stability of the process performance by identifying the special cause variation in the process.
- It uses control charts to detect the unanticipated changes in the process.
- Which control chart to use depends on:
  - Whether the data are continuous or discrete
  - How large the subgroup size is
  - Whether the subgroup size is constant
  - Whether we are interested in measuring defects or defectives
  - Whether we are interested in detecting the shifts in the process mean.

5S

- **5S** is a systematic approach of cleaning and organizing the workplace.
  - Seiri (sorting)
  - Seiton (straightening)
  - Seiso (shining)
  - Seiketsu (standardizing)
  - Shisuke (sustaining)

Kanban

- A **Kanban** system is a demand-driven system.
- The customer demand is the signal to trigger or pull the production.
- Products are made only to meet the immediate demand. When there is no demand, there is no production.
- It was designed to reduce the waste in inventory and increase the speed of responding to immediate demand.

Poka-Yoke

- **Poka-yoke** is a mechanism to eliminate the defects as early as possible in the process.

- Contact Method
  - Use of the shape, color, size, or any other physical attributes of the items

- Constant Number Method
  - Use of a fixed number to make sure a certain number of motions are completed

- Sequence Method
  - Use of a checklist to make sure all the prescribed process steps are followed in the right order.

5.2.11 Control Chart Anatomy

Control Chart Calculations Summary

Chart	Center Line	Control Limits	σ_x
I Chart	$\frac{\sum_{i=1}^{n} x_i}{n}$	$\frac{\sum_{i=1}^{n} x_i}{n} \pm 3 \times \frac{\overline{MR}}{d_2}$	$\frac{\overline{MR}}{d_2}$
MR Chart	$\overline{MR} = \frac{\sum_{i=1}^{n-1} \lvert x_{i+1} - x_i \rvert}{n-1}$	$UCL = D_4 \times \overline{MR}$ $LCL = D_3 \times \overline{MR}$	
Xbar Chart (Xbar-R)	$\overline{\overline{X}} = \frac{\sum_{i=1}^{k} \overline{X}_i}{k}$	$\overline{\overline{X}} \pm A_2 \overline{R}$	$\frac{\overline{R}}{d_2}$
Xbar Chart (Xbar-S)	$\overline{\overline{X}} = \frac{\sum_{i=1}^{k} \overline{X}_i}{k}$	$\overline{\overline{X}} \pm A_3 \overline{s}$	$\frac{\overline{s}}{c_4}$
R Chart	$\overline{R} = \frac{\sum_{i=1}^{k} R_i}{k}$	$UCL = D_4 \times \overline{R}$ $LCL = D_3 \times \overline{R}$	
S Chart	$\overline{s} = \frac{\sum_{i=1}^{k} s_i}{k}$	$UCL = B_4 \times \overline{s}$ $LCL = B_3 \times \overline{s}$	
U Chart	$\overline{u} = \frac{\sum_{i=1}^{k} u_i}{k}$	$\overline{u} \pm 3 \times \sqrt{\frac{\overline{u}}{n_i}}$	$\sqrt{\frac{\overline{u}}{n_i}}$
P Chart	$\overline{p} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} n_i}$	$\overline{p} \pm 3 \sqrt{\frac{\overline{p}(1-\overline{p})}{n_i}}$	$\sqrt{\frac{\overline{p}(1-\overline{p})}{n_i}}$
NP Chart	$n\overline{p} = \frac{\sum_{i=1}^{k} x_i}{k}$	$n\overline{p} \pm 3 \times \sqrt{n\overline{p}(1-\overline{p})}$	$\sqrt{n\overline{p}(1-\overline{p})}$

Control Chart Constants

Subgroup Size	A2	A3	B3	B4	c4	d2	D3	D4
2	1.88	2.659	-	3.267	0.7979	1.128	-	3.267
3	1.023	1.954	-	2.568	0.8862	1.693	-	2.574
4	0.729	1.628	-	2.266	0.9213	2.059	-	2.282
5	0.577	1.427	-	2.089	0.94	2.326	-	2.114
6	0.483	1.287	0.03	1.97	0.9515	2.534	-	2.004
7	0.419	1.182	0.118	1.882	0.9594	2.704	0.076	1.924
8	0.373	1.099	0.185	1.815	0.965	2.847	0.136	1.864
9	0.337	1.032	0.239	1.761	0.9693	2.97	0.184	1.816
10	0.308	0.975	0.284	1.716	0.9727	3.078	0.223	1.777
15	0.223	0.789	0.428	1.572	0.9823	3.472	0.347	1.653
25	0.153	0.606	0.565	1.435	0.9896	3.931	0.459	1.541

Unnatural Patterns

- If there are *unnatural patterns* in the control chart of a process, we consider the process out of statistical control.

- Typical unnatural patterns in control charts:
  - Outliers
  - Trending
  - Cycling
  - Auto-correlative
  - Mixture.

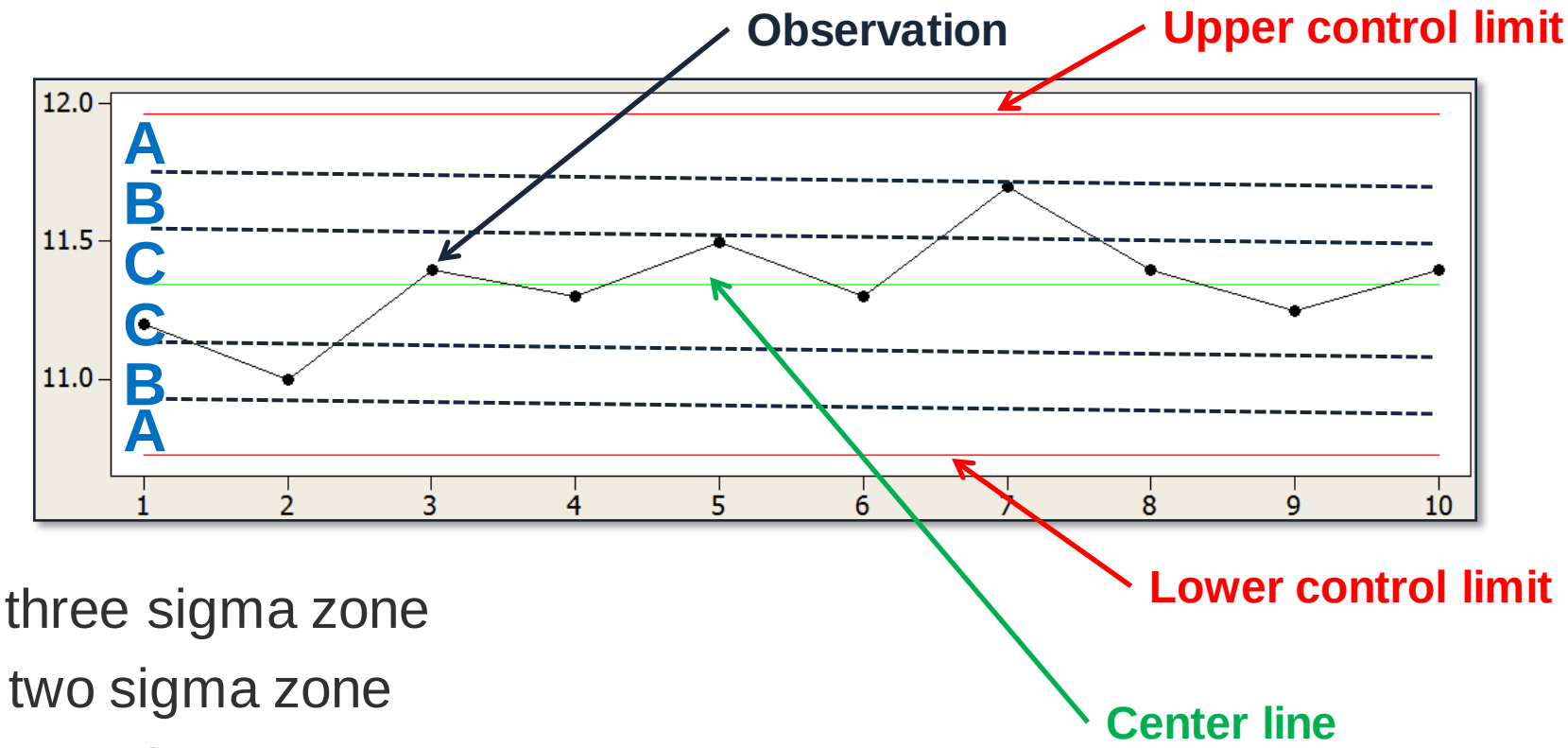
- A process is *in control* if all the data points on the control chart are randomly spread out within the control limits.

Western Electric Rules

- **Western Electric Rules** are the most popular decision rules to detect unnatural patterns in the control charts. They are a group of tests for special causes in a process.

- The area between the upper and lower control limits is separated into six subzones.
  - Zone A: between two and three standard deviations from the center line
  - Zone B: between one and two standard deviations from the center line
  - Zone C: within one standard deviation from the center line.

Western Electric Rules

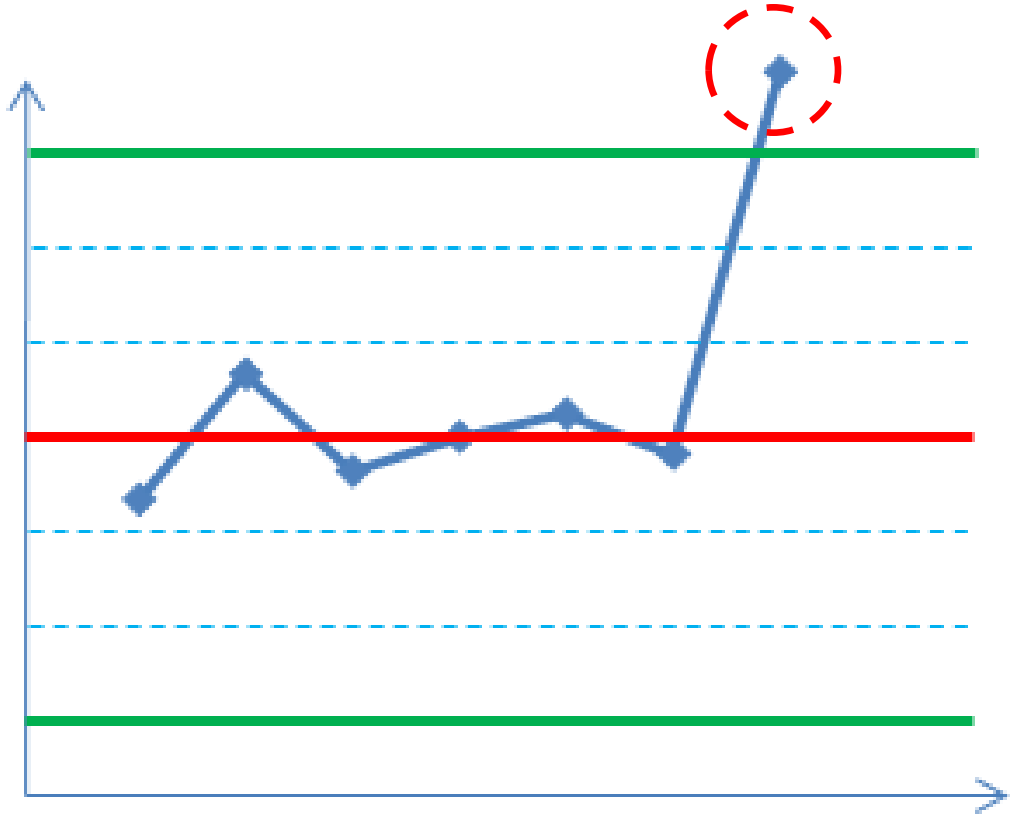


- Zone A: three sigma zone
- Zone B: two sigma zone
- Zone C: one sigma zone

- If a data point falls onto the dividing line of two consecutive zones, the point belongs to the outer zone.

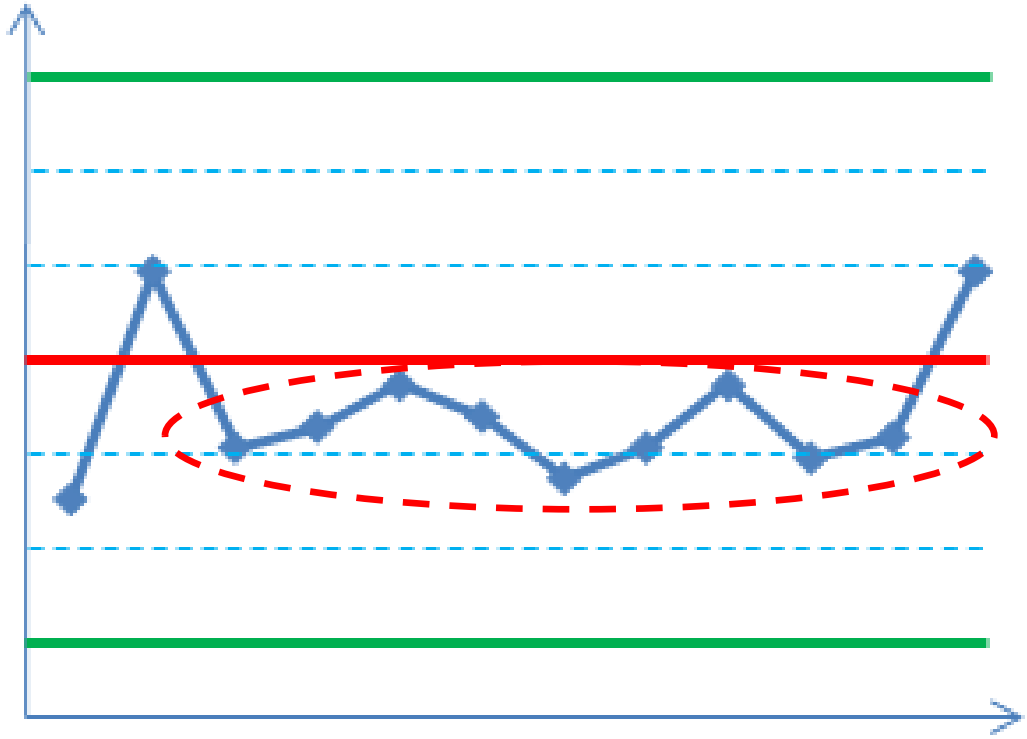
Western Electric Rules

- Test 1: 1 point more than 3 standard deviations from the center line (i.e., 1 point beyond zone A)



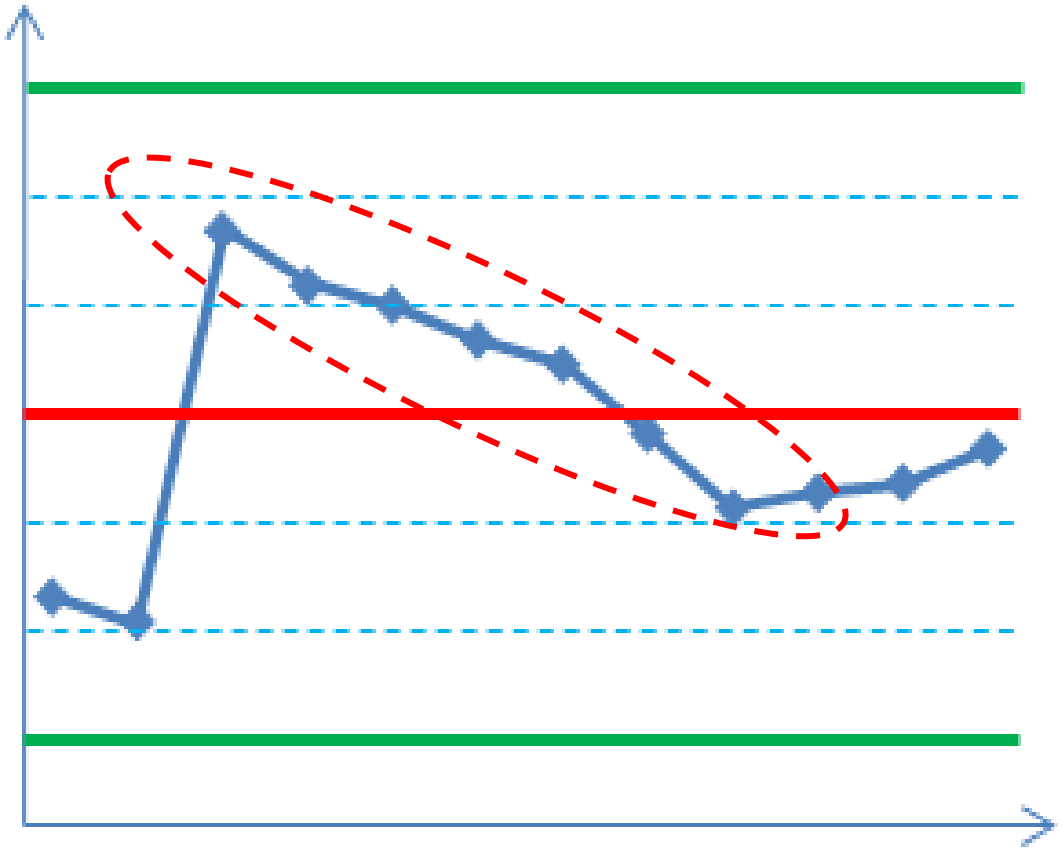
Western Electric Rules

- Test 2: 9 points in a row on the same side of the center line



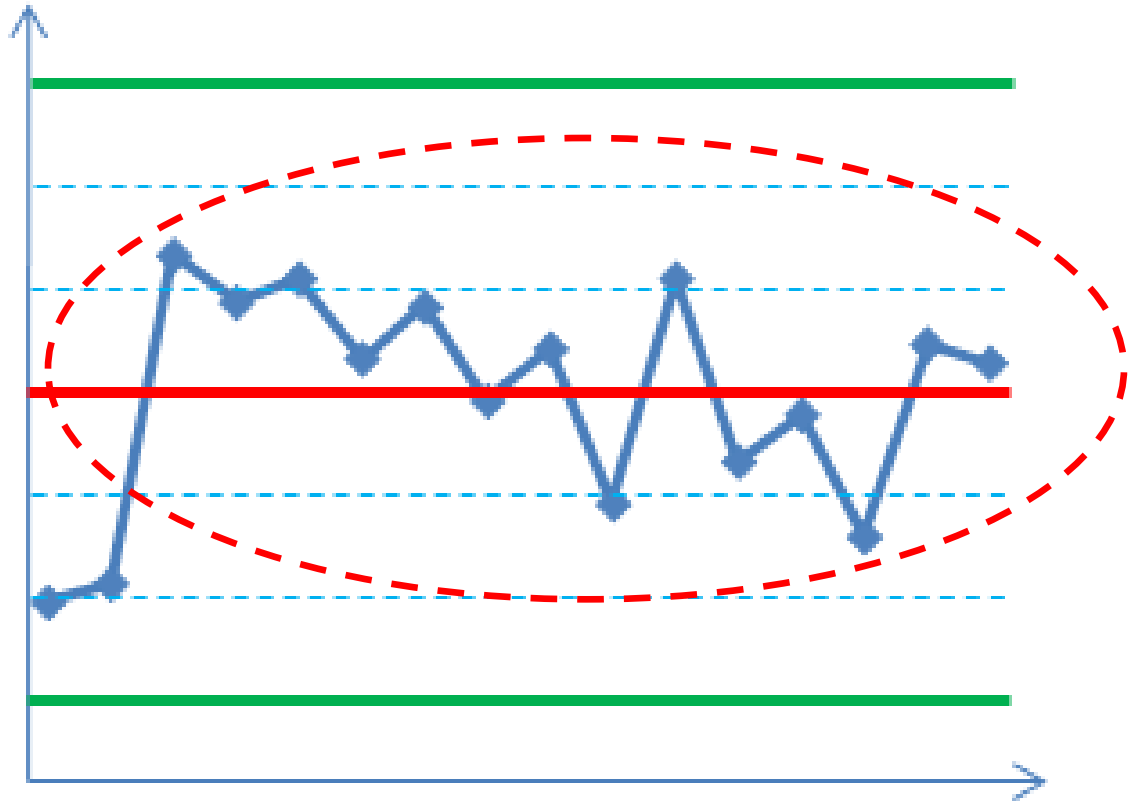
Western Electric Rules

- Test 3: 6 points in a row steadily increasing or steadily decreasing



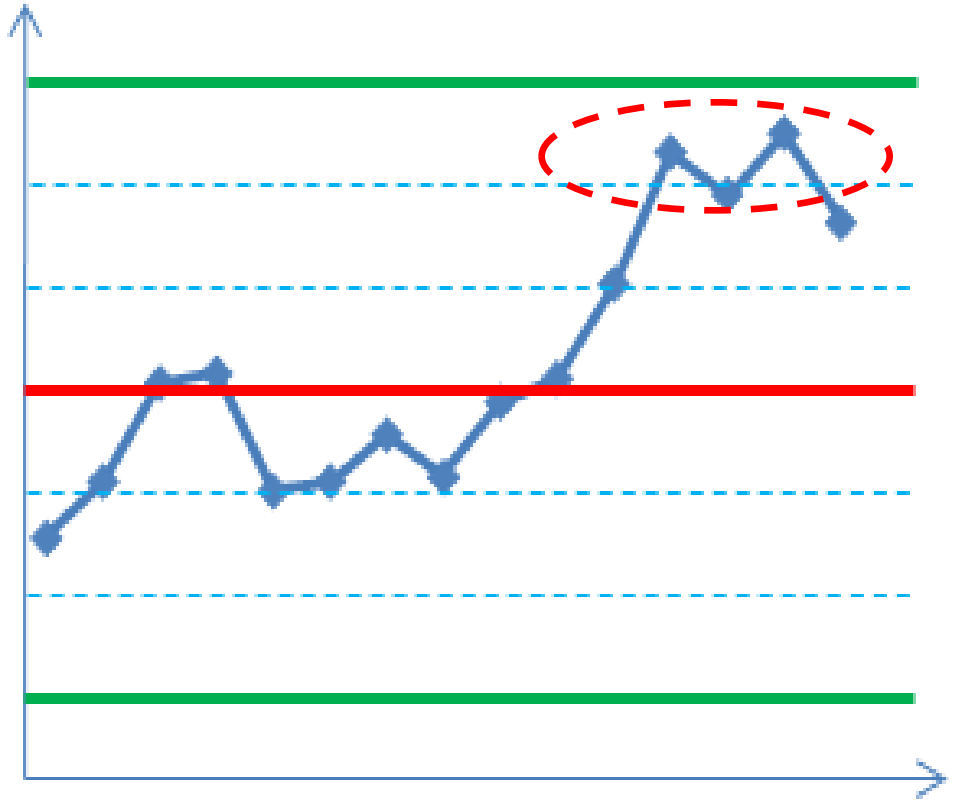
Western Electric Rules

- Test 4: 14 points in a row alternating up and down



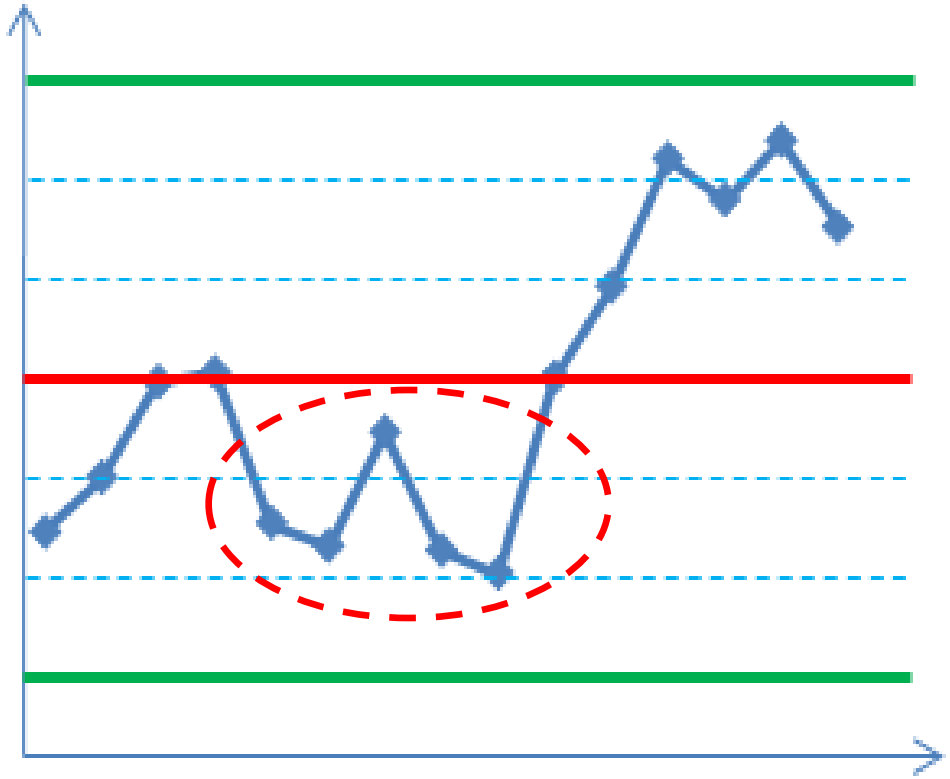
Western Electric Rules

- Test 5: 2 out of 3 points in a row at least 2 standard deviations from the center line (in zone A or beyond) on the same side of the center line



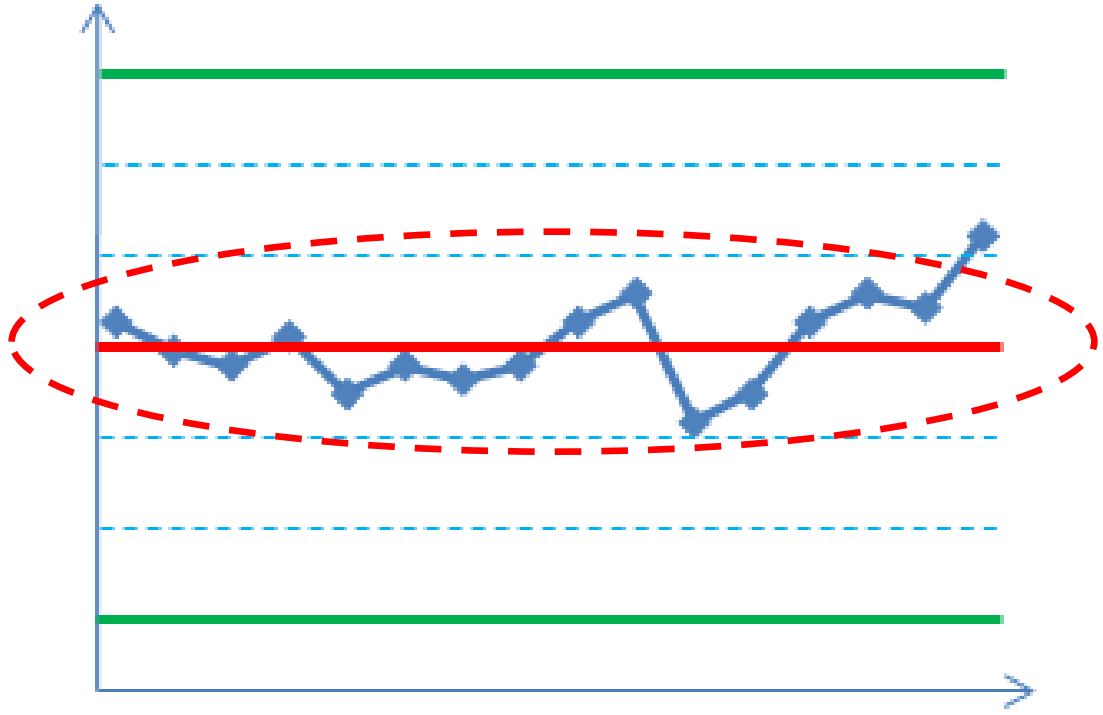
Western Electric Rules

- Test 6: 4 out of 5 points in a row at least 1 standard deviation from the center line (in zone B or beyond) on the same side of the center line



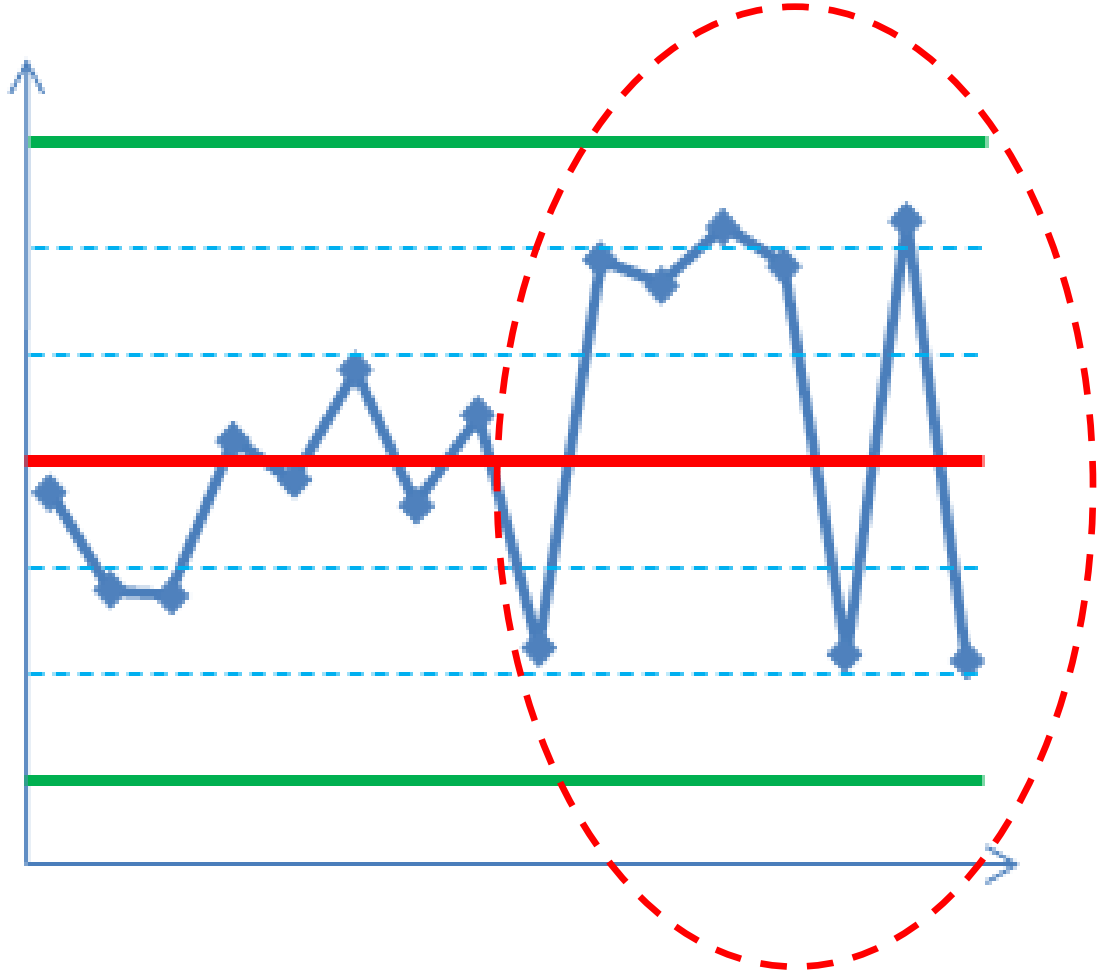
Western Electric Rules

- Test 7: 15 points in a row within 1 standard deviation from the center line (in zone C) on either side of the center line



Western Electric Rules

- Test 8: 8 points in a row beyond 1 standard deviation from the center line (beyond zone C) on either side of the center line



Next Steps

- If no data points fail any tests for special causes, the process is in *statistical control*.
- If any data point fails any tests for special causes, the process is *unstable* and we will need to investigate the observation thoroughly to discover and take actions on the special causes leading to the changes.
- Process stability is the prerequisite of process capability analysis.

5.2.12 Subgroups & Sampling

Subgroups

- **Rational subgrouping** is the basic sampling scheme in SPC (Statistical Process Control).
- When sampling, we randomly select a group (i.e. a subgroup) of items from the population of interest.
- The subgroup size is the count of samples in a subgroup. It can be constant or variable.
- Depending on the subgroup sizes, we select different control charts accordingly.

Impact of Variation

- The rational subgrouping strategy is designed to minimize the opportunity of having special cause variation within subgroups.

- If there is only random variation (background noise) within subgroups, all the special cause variation would be reflected between subgroups. It is easier to detect the out-of-control situation.

- Random variation is inherent and indelible in the process. We are more interested in identifying and taking actions on special cause variation.

Frequency of Sampling

- The **frequency of sampling** in SPC depends on whether we have sufficient data to signal the changes in a process with reasonable time and costs.

- The more frequently we sample, the higher the costs sampling may trigger.

- We need the subject matter experts' knowledge on the nature and characteristics of the process to make good decisions on sampling frequency.

5.3 Six Sigma Control Plans

Black Belt Training: Control Phase

5.1 Lean Controls

5.1.1 Control Methods for 5S

5.1.2 Kanban

5.1.3 Poka-Yoke (Mistake Proofing)

5.2 Statistical Process Control (SPC)

5.2.1 Data Collection for SPC

5.2.2 I-MR Chart

5.2.3 Xbar-R Chart

5.2.4 U Chart

5.2.5 P Chart

5.2.6 NP Chart

5.2.7 X-S chart

5.2.8 CumSum Chart

5.2.9 EWMA Chart

5.2.10 Control Methods

5.2.11 Control Chart Anatomy

5.2.12 Subgroups, Variation, Sampling

5.3 Six Sigma Control Plans

5.3.1 Cost Benefit Analysis

5.3.2 Elements of the Control Plan

5.3.3 Elements of the Response Plan

5.3.1 Cost Benefit Analysis

What is Cost-Benefit Analysis?

- The **cost-benefit analysis** is a systematic method to assess and compare the financial costs and benefits of multiple scenarios in order to make sound economic decisions.

- A cost-benefit analysis is recommended to be done at the beginning of the project based on estimations of the experts from the finance team in order to determine whether the project is financially feasible.

- It is recommended to update the cost-benefit analysis at each DMAIC phase of the project.

Why Cost-Benefit Analysis?

- In the Define phase of the project, the cost-benefit analysis helps us understand the financial feasibility of the project.

- In the middle phases of the project, updating and reviewing the cost-benefit analysis helps us compare potential solutions and make robust data-driven decisions.

- In the Control phase of the project, the cost-benefit analysis helps us track the project's profitability.

Return on Investment

- **Return on investment** (also called ROI, rate of return, or ROR) is the ratio of the net financial benefits (either gain or loss) of a project or investment to the financial costs.

$$ROI = \frac{TotalNetBenefits}{TotalCosts} \times 100\%$$

where

$$TotalNetBenefits = TotalBenefits - TotalCosts$$

Return on Investment (ROI)

- The return on investment is used to evaluate the financial feasibility and profitability of a project or investment.
  - If ROI < 0, the investment is not financially viable.
  - If ROI = 0, the investment has neither gain nor loss.
  - If ROI > 0, the investment has financial gains.
- The higher the ROI, the more profitable the project.

Net Present Value (NPV)

- The **net present value** (also called NPV, net present worth, or NPW) is the total present value of the cash flows calculated using a discount rate.

$$NPV = \frac{NetCashFlow_t}{(1+r)^t}$$

Where

$NetCashFlow_t$ is the net cash flow happening at time t;
r is the discount rate;
t is the time of the cash flow.

Cost Estimation

- Examples of costs triggered by the project:
  - Administration
  - Asset
  - Equipment
  - Material
  - Delivery
  - Real estate
  - Labor
  - Training
  - Consulting.

Benefits Estimation

- Examples of benefits generated by the project:
  - Direct revenue increase
  - Waste reduction
  - Operation cost reduction
  - Quality and productivity improvement
  - Market share increase
  - Cost avoidance
  - Customer satisfaction improvement
  - Associate satisfaction improvement.

Challenges in Cost and Benefit Estimation

- Different analysts might come up with different cost and benefit estimations due to their subjectivity in determining:
  - The discount rate
  - The time length of the project and its impact
  - Potential costs of the project
  - The tangible/intangible benefits of the project
  - The specific contribution of the project to the relevant financial gains/loss.

5.3.2 Elements of Control Plans

Control Plans

- The **control plans** ensure that the changes introduced by a Six Sigma project are sustained over time.
- Benefits of the Control phase:
  - Methodical roll-out of changes including standardization of processes and work procedures
  - Ensure compliance with changes through methods like auditing and corrective actions
  - Transfer solutions and learning across the enterprise
  - Plan and communicate standardized work procedures
  - Coordinate ongoing team and individual involvement
  - Standardize data collection and procedures
  - Measure process performance, stability, and capability
  - Plan actions that mitigate possible out-of-control conditions
  - Sustain changes over time.

What is a Control Plan?

- A **control plan** is a management planning tool to identify, describe, and monitor the process performance metrics in order to meet the customer specifications steadily.
- It proposes the plan of monitoring the stability and capability of inputs and outputs of critical process steps in the Control phase of a project.
- It covers the data collection plan of gathering the process performance measurements.
- Control plans are the most overlooked element of most projects. It is critical that a good solution be solidified with a great control plan!

Control Plan Elements

- Control Plan
  - The clear and concise summary document that details key process steps, CTQs metrics, measurements, and corrective actions.
- Standard Operating Procedures (SOPs)
  - Supporting documentation showing the “who does what, when, and how” in completing the tasks.
- Communication Plan
  - Document outlining messages to be delivered and the target audience.
- Training Plan
  - Document outlining the necessary training for employees to successfully perform new processes and procedures.
- Audit Checklists
  - Document that provides auditors with the audit questions they need to ask.
- Corrective Actions
  - Activities that need to be conducted when an audit fails.

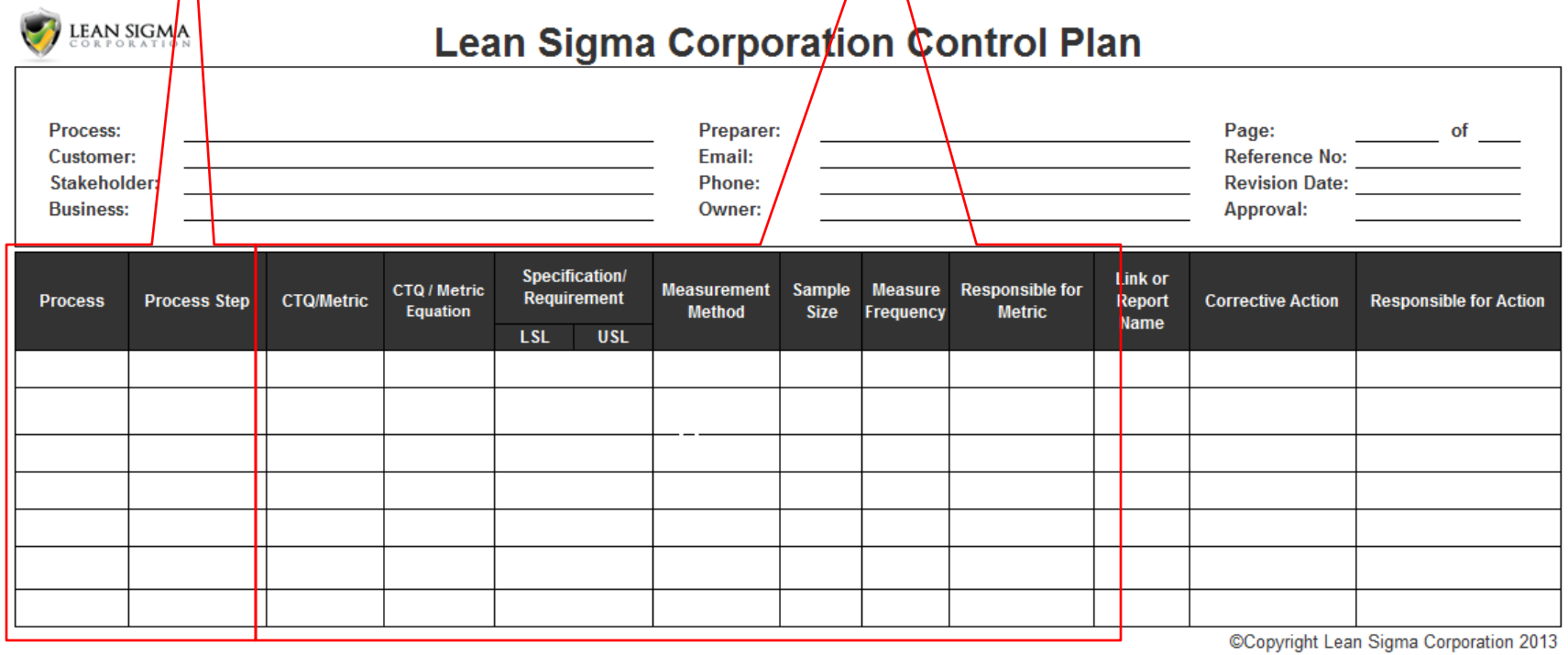
Control Plan

- The control plan identifies critical process steps that have significant impact on the products or services and the appropriate controls mechanisms.
- The control plan includes measurement systems that monitor and help manage key process step performance.
- Specified limits and targets of the performance metrics are clearly defined and communicated.
- Sampling plans to collect the measurements are declared:
  - How many samples are needed?
  - How often do we need to sample?
  - Where should we sample?

Control Plan

Key processes and process steps are identified

Critical information regarding key measurements is documented and clarified



LEAN SIGMA CORPORATION

Lean Sigma Corporation Control Plan

Process: _____ Preparer: _____ Page: _____ of ____
Customer: _____ Email: _____ Reference No: _____
Stakeholder: _____ Phone: _____ Revision Date: _____
Business: _____ Owner: _____ Approval: _____

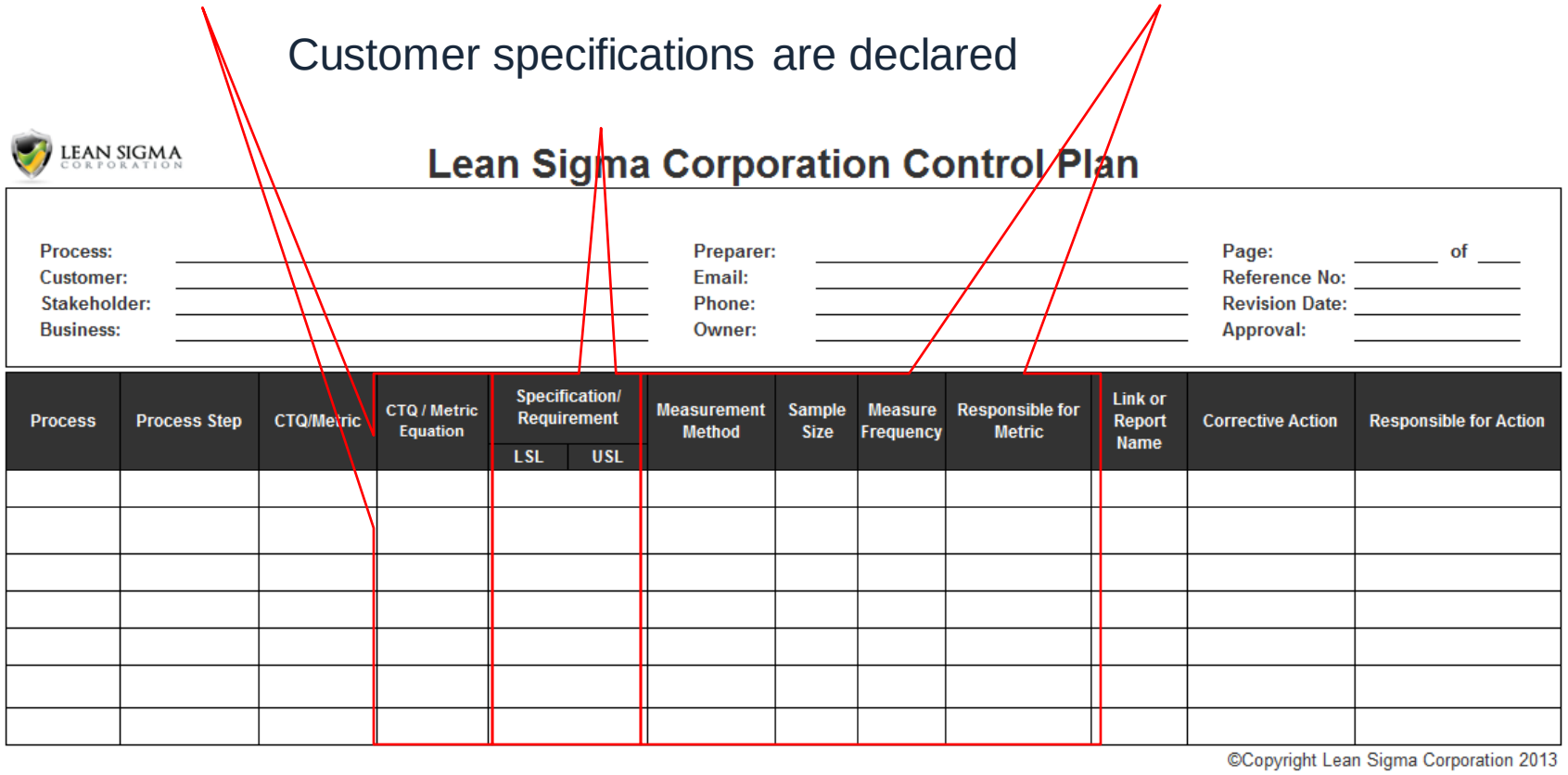
Process	Process Step	CTQ/Metric	CTQ / Metric Equation	Specification/ Requirement		Measurement Method	Sample Size	Measure Frequency	Responsible for Metric	Link or Report Name	Corrective Action	Responsible for Action
				LSL	USL							

Control Plan

Measurements are clearly defined with equations

Other key measurement information is documented: sample size, measurement frequency, people responsible for the measurement, etc.

Customer specifications are declared



LEAN SIGMA CORPORATION

Lean Sigma Corporation Control Plan

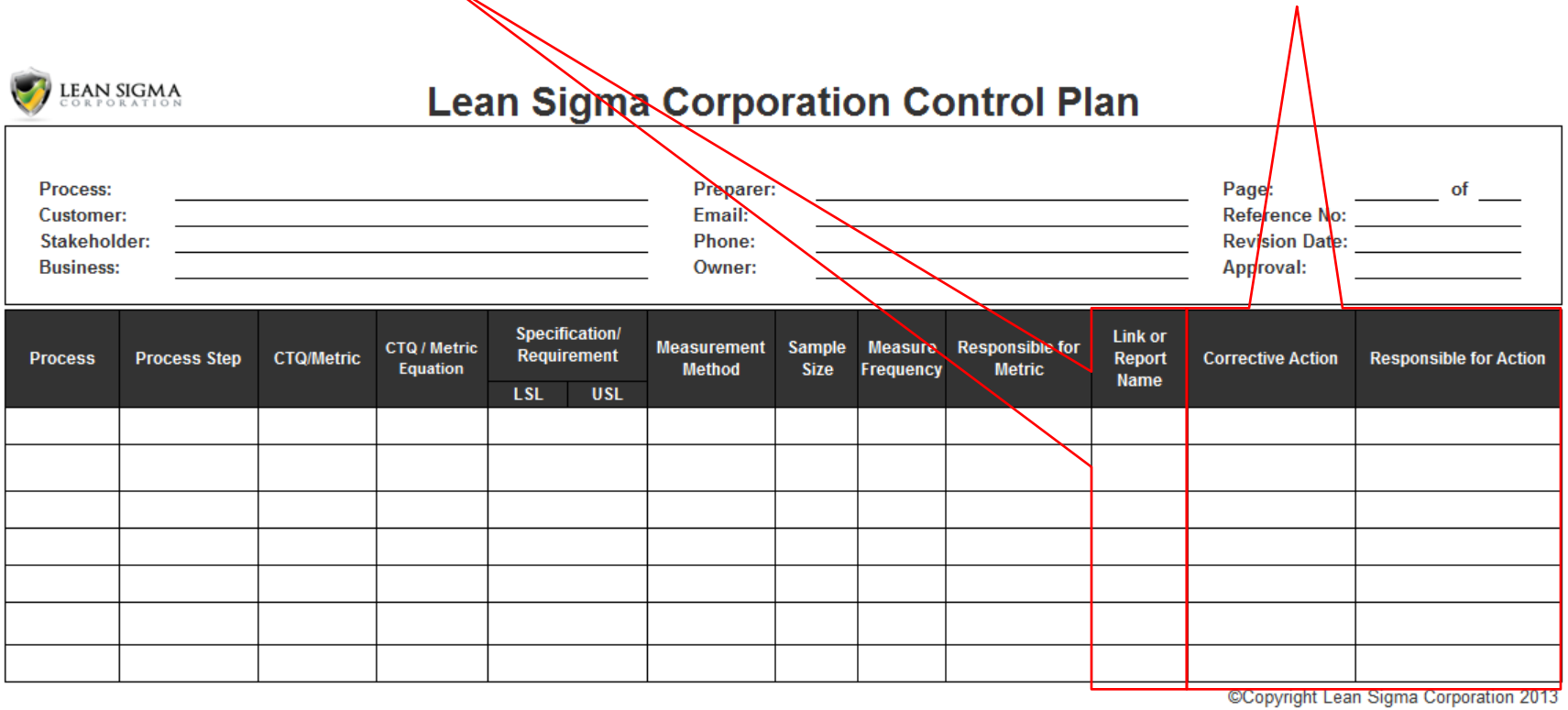
Process:		Preparer:		Page:	of
Customer:		Email:		Reference No:	
Stakeholder:		Phone:		Revision Date:	
Business:		Owner:		Approval:	

Process	Process Step	CTQ/Metric	CTQ / Metric Equation	Specification/ Requirement		Measurement Method	Sample Size	Measure Frequency	Responsible for Metric	Link or Report Name	Corrective Action	Responsible for Action
				LSL	USL							

Control Plan

Where will this measurement or report be found? Good control plans provide linking information or other report reference information.

Control plans identify the mitigating action or corrective actions required in the event the measurement falls out of spec or control. Responsible parties are also declared.



LEAN SIGMA CORPORATION

Lean Sigma Corporation Control Plan

Process: _____ Preparer: _____ Page: _____ of ____
Customer: _____ Email: _____ Reference No: _____
Stakeholder: _____ Phone: _____ Revision Date: _____
Business: _____ Owner: _____ Approval: _____

Process	Process Step	CTQ/Metric	CTQ / Metric Equation	Specification/ Requirement		Measurement Method	Sample Size	Measure Frequency	Responsible for Metric	Link or Report Name	Corrective Action	Responsible for Action
				LSL	USL							

Control Plan Example

Process Name:	Custom Kit Assembly			**Prepared by:**	John Doe	**Page:**	1 of 1
Customer:	Assemble for Me Inc.	Int/Ext	Ext	**Email:**	Johndoe@Custommanufacturersinc.com	**Reference No:**	001-01
Stakeholder:	Production Supervisor			**Phone:**	555-555-5515	**Revision Date:**	3/9/09
Business:	Custom Manufacturers Inc			**Control Plan Owner:**	Production Supervisor	**Approval:**	Yes

Process	Process Step	CTQ/Metric	CTQ / Metric Equation	Specification/ Requirement		Measurement Method	Sample Size	Measure Frequency	Responsible for Metric	Link or Report Name	Corrective Action	Responsible for Action
				LSL	USL							
Parts Picking	Picking Inventory	Picking Accuracy	# correct parts/ # parts picked	93.73%	99.86%	Inspection at Assembly Setup Station	All Assembly Jobs	Daily	Assembly Supervisor	Pk Accuracy	Audit Picking Procedures	Inventory Supervisor
											Conduct Gage R&R on Pick Counting Methods	Division Black Belt
Assembly	Custom Assembly	Assembly Accuracy	# Good kits / # Kits	98.65%	99.73%	Quality Inspection	38 Random Kits	Daily	Quality Control Mgr	Kit Accuracy	Audit Assembly Procedures	Assembly Supervisor
											Audit Setup Procedures	Setup Associate Supervisor
											Audit Picking Procedures	Inventory Supervisor
Shipping	Shipping	Shipping Accuracy	# Good products /# products sampled	99.73 %	100%	Distribution QC	52 Products	Daily	Distribution Mgr	Ship Accuracy	Audit Shipping Procdures	Shipping Supervisor

Standard Operating Procedures (SOPs)

- **Standard Operating Procedures** (SOPs) are documents that focus on process steps, activities, and specific tasks required to complete an operation.

- SOPs should not be much more than two to four pages.

- SOPs should be written to the user's level of required detail and information.
  - The level of detail is dependent on the position's required skills and training

- Good SOPs are auditable, easy to follow, and not difficult to find.
  - Auditable characteristics are: observable actions and countable frequencies. Results should be evident to a third party (*compliance to the SOP must be measurable*).

SOP Elements

- SOPs are intended to impart high value information in concise and well-documented manner.
- SOP Title and Version Number:
  - Provide a title and unique identification number with version information.
- Date:
  - List the original creation date; add all revision dates.
- Purpose:
  - State the reason for the SOP and what it intends to accomplish.
- Scope:
  - Identify all functions, jobs, positions, and/or processes governed or affected by the SOP.

SOP Elements

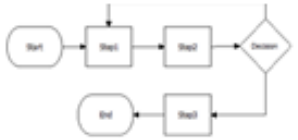
- Responsibilities:
  - Identify job functions and positions (not people) responsible for carrying out activities listed in the SOP.
- Materials:
  - List all material inputs: parts, files, data, information, instruments, etc.
- Process Map:
  - Show high level or level two to three process maps or other graphical representations of operating steps.
- Process Metrics:
  - Declare all process metrics and targets or specifications.
- Procedures:
  - List actual steps required to perform the function.
- References:
  - List any documents that support the SOP.

SOP Template

Standard Operating Procedure Template

SOP Name/Title:		
Document Storage Location/Source:		Document No:
SOP Originator:	Approving Position:	Effective Date:
Name:	Name:	Last Edited Date:
Signature:	Signature:	Other:

1. Purpose
2. Scope
3. Responsibilities
4. Materials
5. Related Documents
6. Definitions
7. Process Map



8. Procedures

Step	Action	Responsible
1		
2		
3		

9. Process Metrics
10. Resources

Communication Plans

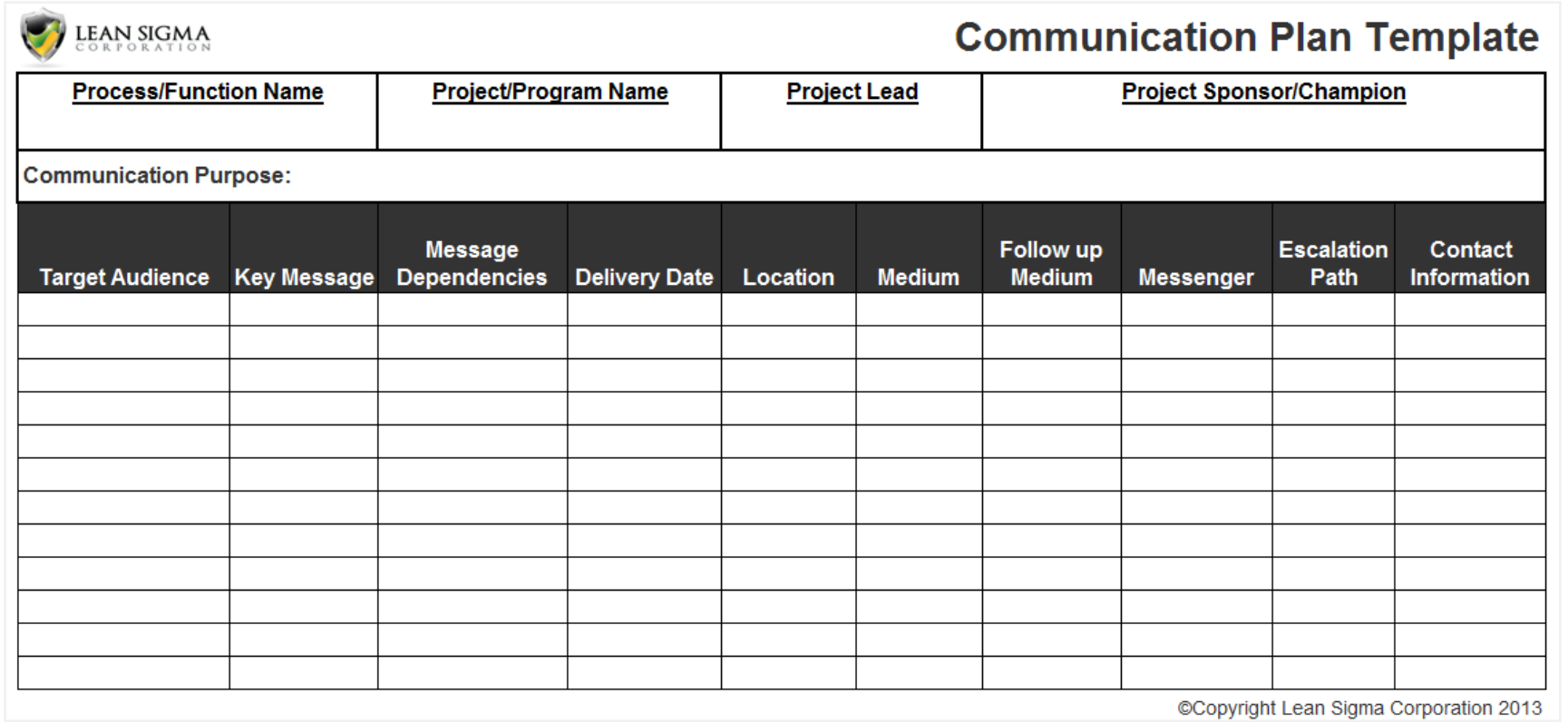
- **Communication plans** are documents that focus on planning and preparing for the dissemination of information.

- Communication plans organize messages and ensure that the proper audiences receive the correct message at the right time.

- A good communication plan identifies:
  - Audience
  - Key points/message
  - Medium (how the message is to be delivered)
  - Delivery schedule
  - Messenger
  - Dependencies and escalation points
  - Follow-up messages and delivery mediums.

- Communication plans help develop and execute strategies for delivering changes to an organization.

Communication Plan Template



LEAN SIGMA CORPORATION

Communication Plan Template

Process/Function Name	Project/Program Name	Project Lead	Project Sponsor/Champion

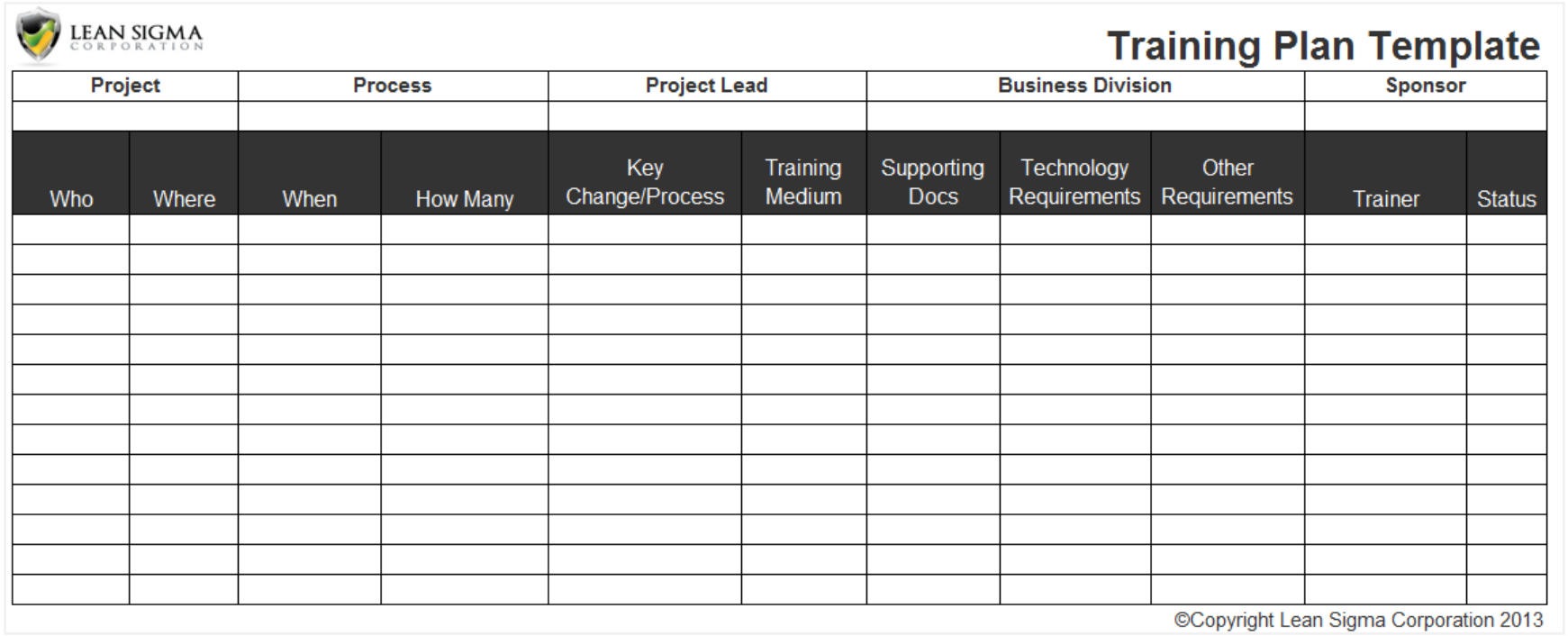
Communication Purpose:

Target Audience	Key Message	Message Dependencies	Delivery Date	Location	Medium	Follow up Medium	Messenger	Escalation Path	Contact Information

Training Plans

- **Training plans** are used to manage the delivery of training for new processes and procedures.
- Most GB or BB projects will require changes to processes and/or procedures that must be executed or followed by various employees.
- Training plans should incorporate all SOPs related to performing new or modified tasks.
- Training plans use and support existing SOPs and do not supersede them.
- Training plans should include logistics:
  - One-on-one or classroom
  - Instruction time
  - Location of training materials
  - Master training reference materials
  - Instructors and intended audience
  - Trainee names.

Training Plan Template



LEAN SIGMA CORPORATION

Training Plan Template

Project	Process	Project Lead	Business Division	Sponsor

Who	Where	When	How Many	Key Change/Process	Training Medium	Supporting Docs	Technology Requirements	Other Requirements	Trainer	Status

Audits

- What is an **audit**?
  - ISO 9000 defines an Audit as “a systematic and independent examination to determine whether quality activities and related results comply with planned arrangements and whether these arrangements are implemented effectively and are suitable to achieve objectives.”

- Audits are used to ensure actions, processes, procedures, and other tasks are performed as expected.

Audit Guidelines

- Audits should be directed by managers, supervisors, and other accountable positions.
- An audit's purpose must be well-defined and executed by independent unbiased personnel.
- Auditors must:
  - Be qualified to perform their tasks
  - Attend and successfully complete an internal auditing training session
  - Be able to identify whether or not activities are being followed according to the defined SOP
  - Base conclusions on facts and objective evidence
  - Use a well documented audit checklist.

Audits should confirm compliance or declare non-compliance.

Audit Checklists

- Auditors should review the SOPs before preparing checklists or ensure that existing checklists properly reference SOPs.

- Audit checklists:
  - Serve as guides for identifying items to be examined
  - Are used in conjunction with understanding of the procedure
  - Ensure a well-defined audit scope
  - Identify needed facts during audits
  - Provide places to record gathered facts.

- Checklists should include:
  - A review of training records
  - A review of maintenance records
  - Questions or observations that focus on expected behaviors
  - Questions should be open-ended where possible
  - Definitive observations yes/no, true/false, present/absent, etc.

Audit Checklist Template

LEAN SIGMA CORPORATION

Audit Checklist

Target Area:	Statement of Audit Objective:	Auditor:	Audit Date:
Audit Technique	Auditable Item, Observation, Procedure etc.	Individual Auditor Rating (Circle Rating)	
Observation	Have all associates been trained?	YES	NO
Observation	Is training documentation available?	YES	NO
Observation	Is training documentation current?	YES	NO
Observation	Are associates wearing proper safety gear?	YES	NO
Observation	Are SOP's available?	YES	NO
Observation	Are SOP's current?	YES	NO
Observation	Is quality being measured	YES	NO
Observation	Is sampling being conducted in random fashion	YES	NO
Observation	Is sampling meeting it's sample size target?	YES	NO
Observation	Are control charts in control	YES	NO
Observation	Are control charts current?	YES	NO
Observation	Is the process capability index >1.0?	YES	NO
Number of Out of Compliance Observations			
Total Observations			
Audit Yield			#DIV/0!
Corrective Actions Required			
Auditor Comments			

5.3.3 Response Plan Elements

What is a Response Plan?

- A **response plan** should be a component of as many control plan elements as possible.
- Response plans are a management planning tool to describe corrective actions necessary in the event of out-of-control situations.
- There is never any guarantee that processes will always perform as designed. Therefore, it is wise to prepare for occasions when special causes are present.
- Response plans help us mitigate risks and, as already mentioned, should be part of several control plan elements.

Response Plan Elements

- Action triggers
  - When do we need to take actions to correct a problem or issue?
- Action recommendation
  - What activities are required in order to solve the problem in the process? The action recommended can be short-term (quick fix) or long-term (true process improvement).
- Action respondent
  - Who is responsible for taking actions?
- Action date
  - When did the actions happen?
- Action results
  - What actions have been taken?
  - When were actions taken?
  - What are the outcomes of the actions taken?

Response Plan Elements

LEAN SIGMA CORPORATION

Lean Sigma Corporation Control Plan

Process: _____ Preparer: _____ Page: _____ of ____
Customer: _____ Email: _____ Reference No: _____
Stakeholder: _____ Phone: _____ Revision Date: _____
Business: _____ Owner: _____ Approval: _____

Process	Process Step	CTQ/Metric	CTQ / Metric Equation	Specification/ Requirement		Measurement Method	Sample Size	Measure Frequency	Responsible for Metric	Link or Report Name	Corrective Action	Responsible for Action
				LSL	USL							

Note the response plan element in this control plan template

Response Plan Elements

LEAN SIGMA CORPORATION

Communication Plan Template

Process/Function Name	Project/Program Name	Project Lead	Project Sponsor/Champion

Communication Purpose:

Target Audience	Key Message	Message Dependencies	Delivery Date	Location	Medium	Follow up Medium	Messenger	Escalation Path	Contact Information

Note the response plan element in this communication plan template

Response Plan Elements

LEAN SIGMA CORPORATION

Audit Checklist

Target Area:	Statement of Audit Objective:	Auditor:	Audit Date:
Audit Technique	**Auditable Item, Observation, Procedure etc.**	**Individual Auditor Rating (Circle Rating)**	
Observation	Have all associates been trained?	YES	NO
Observation	Is training documentation available?	YES	NO
Observation	Is training documentation current?	YES	NO
Observation	Are associates wearing proper safety gear?	YES	NO
Observation	Are SOP's available?	YES	NO
Observation	Are SOP's current?	YES	NO
Observation	Is quality being measured	YES	NO
Observation	Is sampling being conducted in random fashion	YES	NO
Observation	Is sampling meeting it's sample size target?	YES	NO
Observation	Are control charts in control	YES	NO
Observation	Are control charts current?	YES	NO
Observation	Is the process capability index >1.0?	YES	NO
Number of Out of Compliance Observations			
Total Observations			
Audit Yield			#DIV/0!
Corrective Actions Required			
Auditor Comments			

Note the response plan element in this audit checklist

LEAN SIGMA
CORPORATION

Lean Six Sigma Black Belt Training

Featuring Examples from SigmaXL