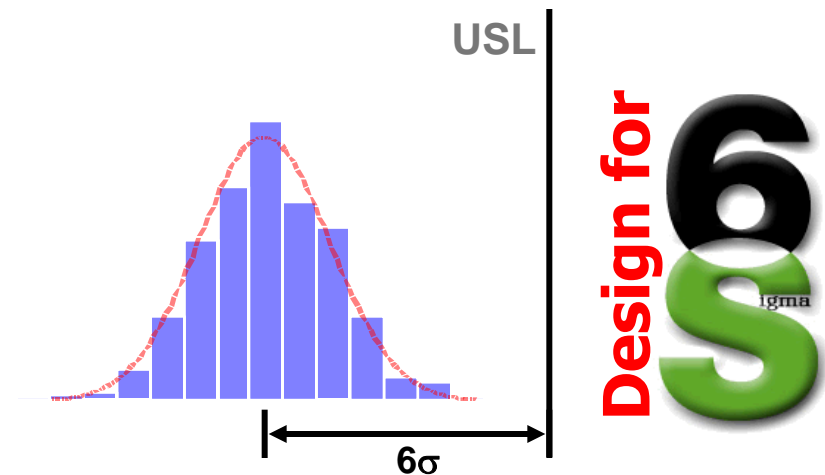


# Basic Statistics

## Basic Graphs



Design for  
**6σ**  
Sigma

# Objectives

---

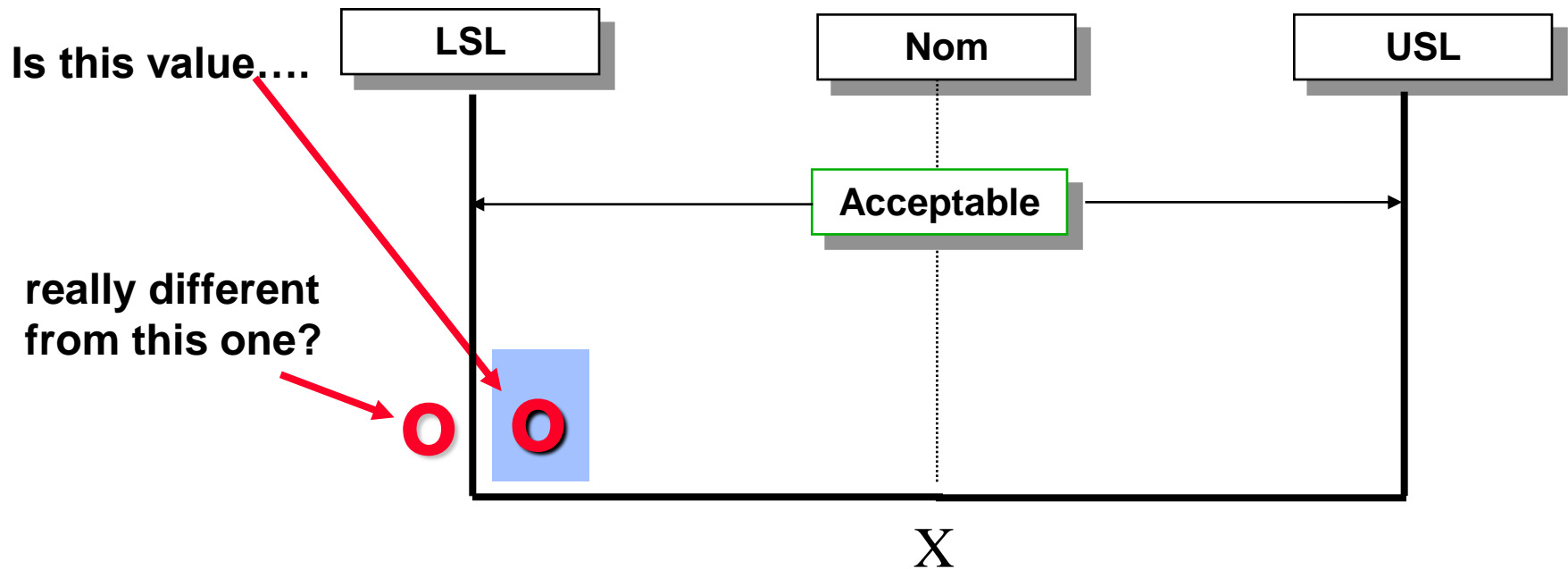
At the end of this module, the participant will be able to :

- Use graphs to visually represent variation
- Assess normality using the probability plot
- Describe a variable's distribution using graphs as well as shape, center, and spread metrics
- Use the normal distribution to calculate the probability of an outcome

# Traditional View of Performance

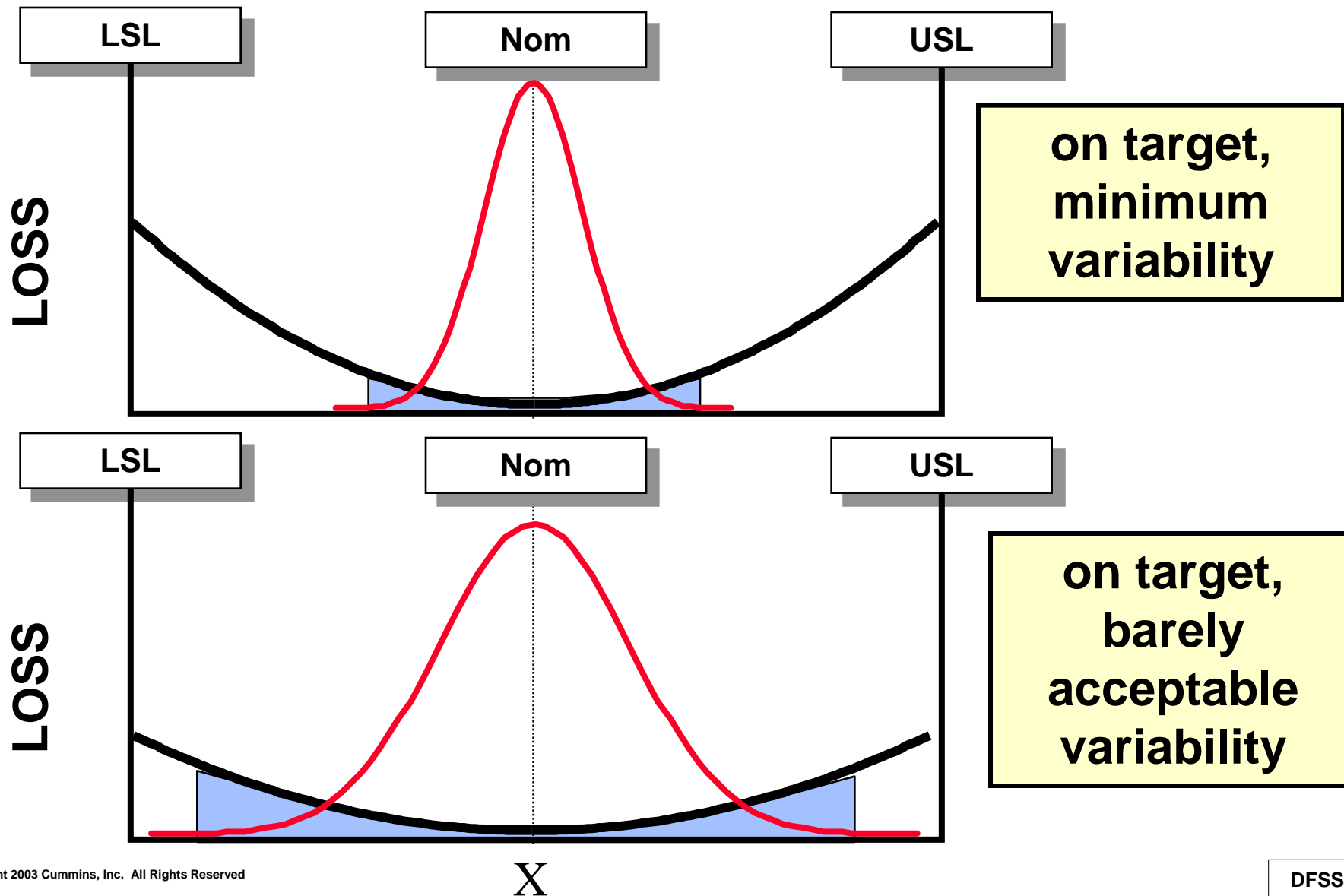
Our primary concern WILL NO LONGER BE  
“Are we in spec?”

## Traditional View

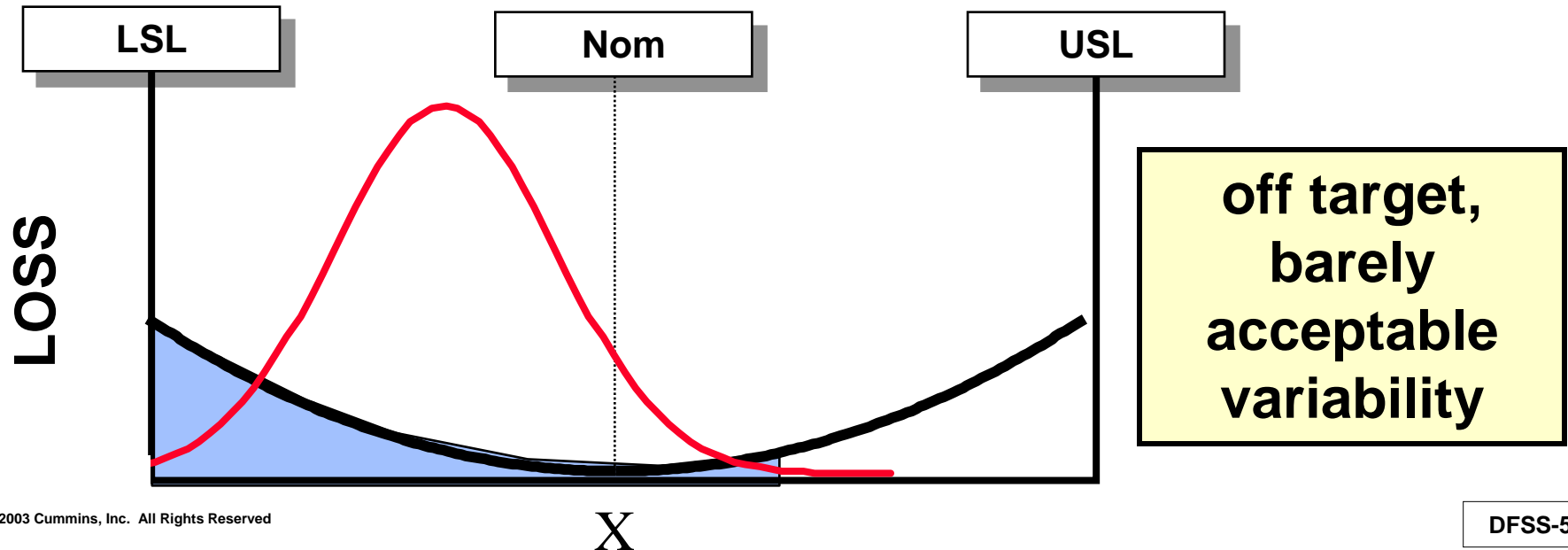
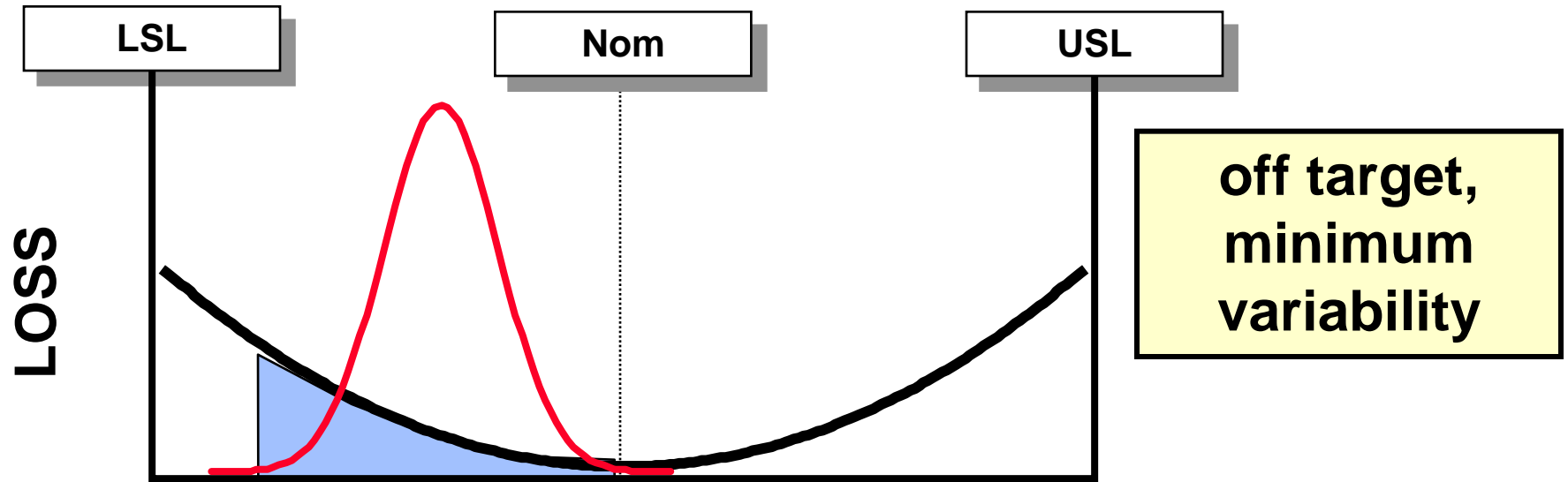


Yet we are ready to use one and throw the other out!

# Cost of Variation when On Target



# Cost of Variation when Off Target

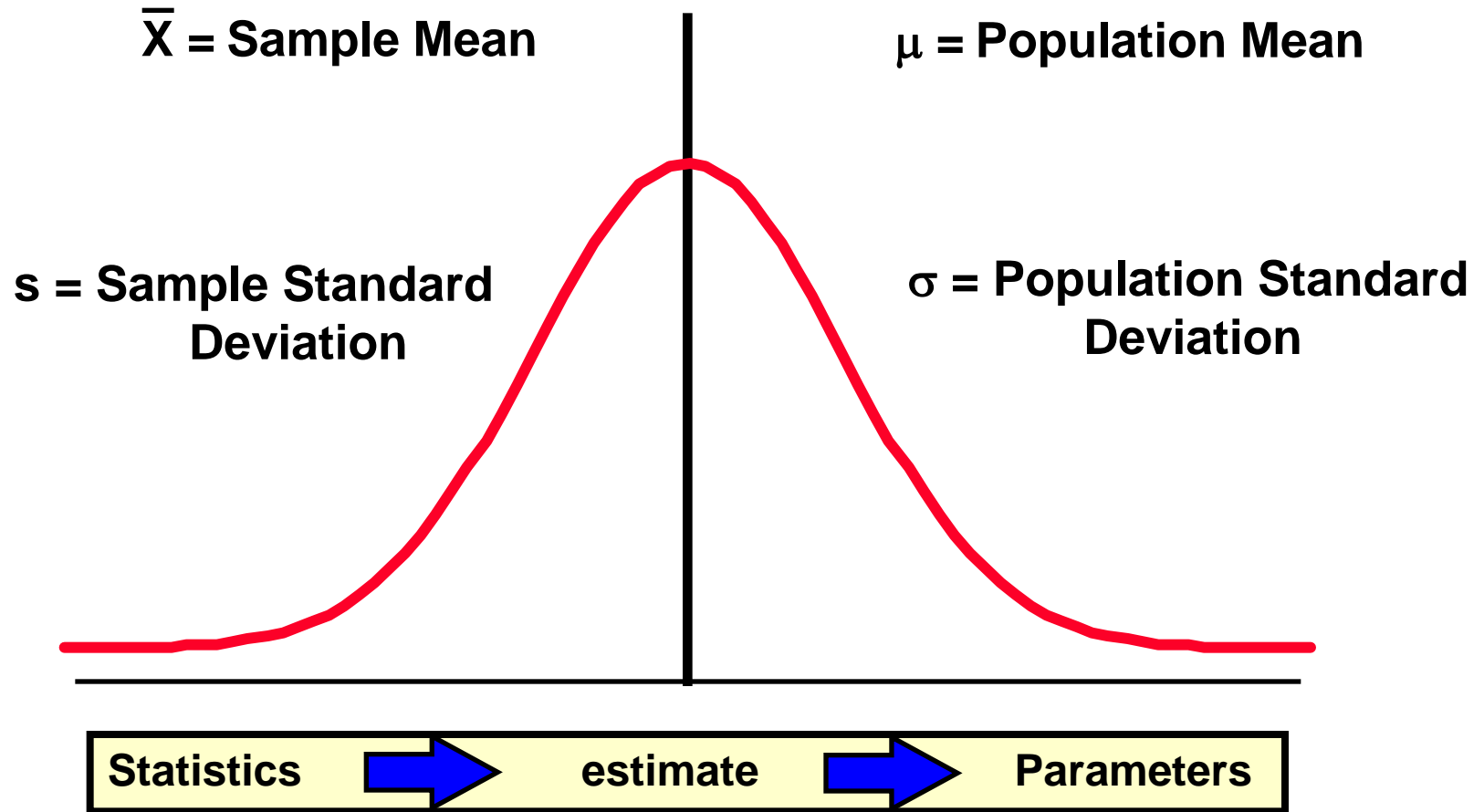


# How to Assess Performance

---

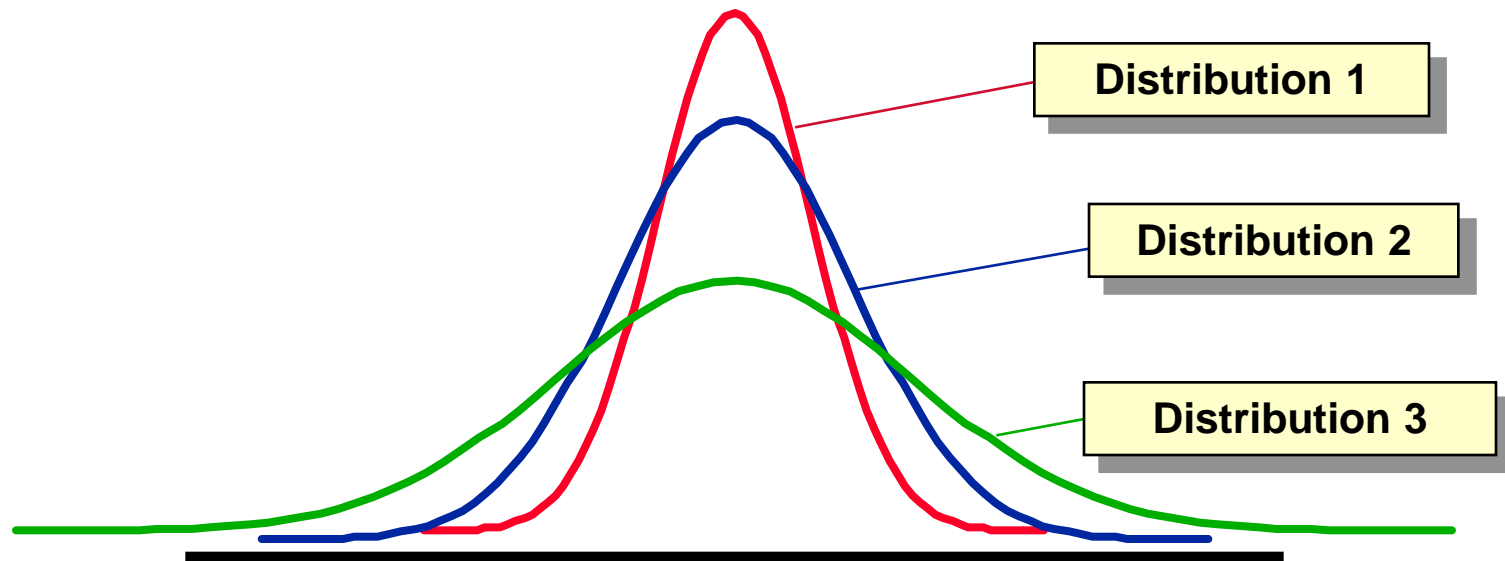
- From now on, we will be concerned with the center and spread of the critical parameters ... not only whether it's in spec or out of spec!
  
- We will assess product performance by
  - studying how the mean and variance (voice of the design) compare to the specification limits (voice of the customer).
    - use the mean of the distribution to determine if the response is on target & has sensitivity to noise
    - use the standard deviation of the distribution to determine variability & sensitivity to noise

# Sample Statistics vs Population Parameters



# The Normal Distribution

- Property 1: A normal distribution can be described completely by knowing only the :
  - mean and standard deviation

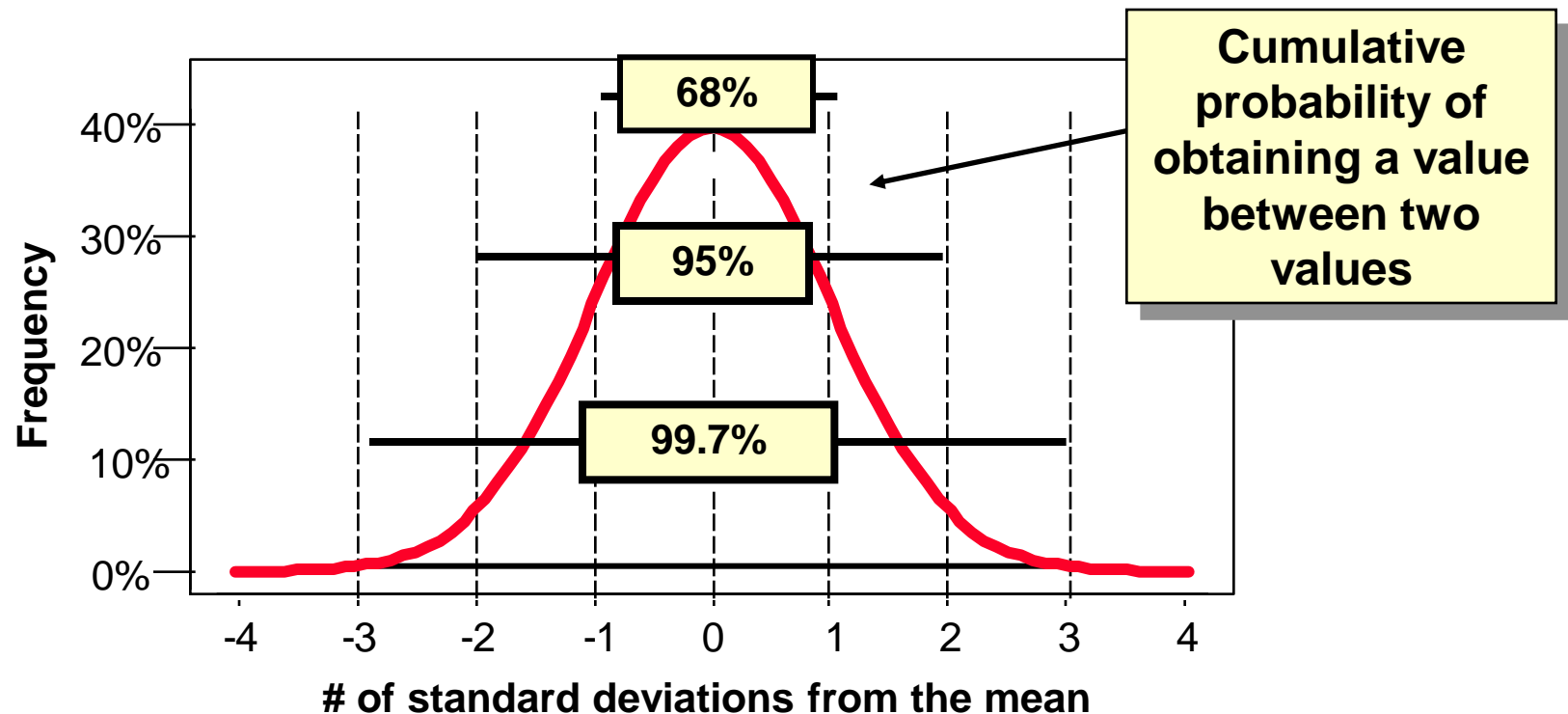


**What is the difference between these normal distributions?**

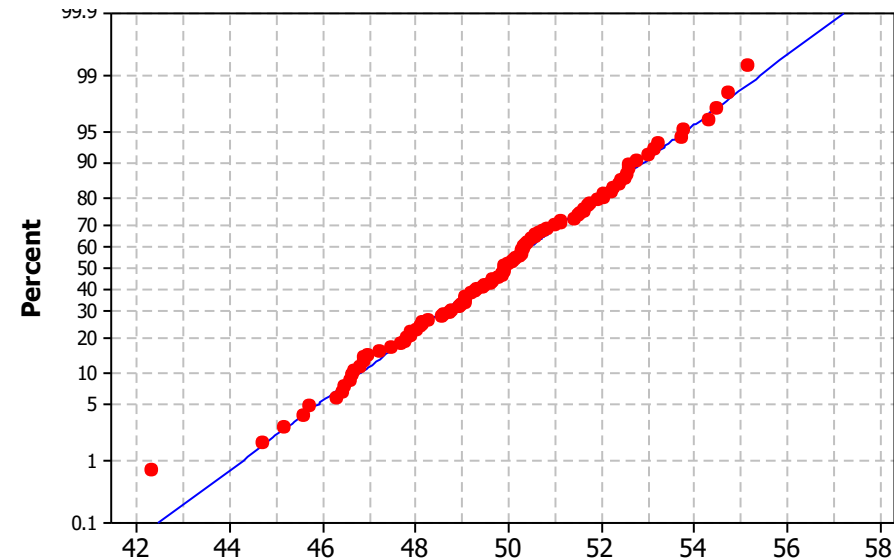
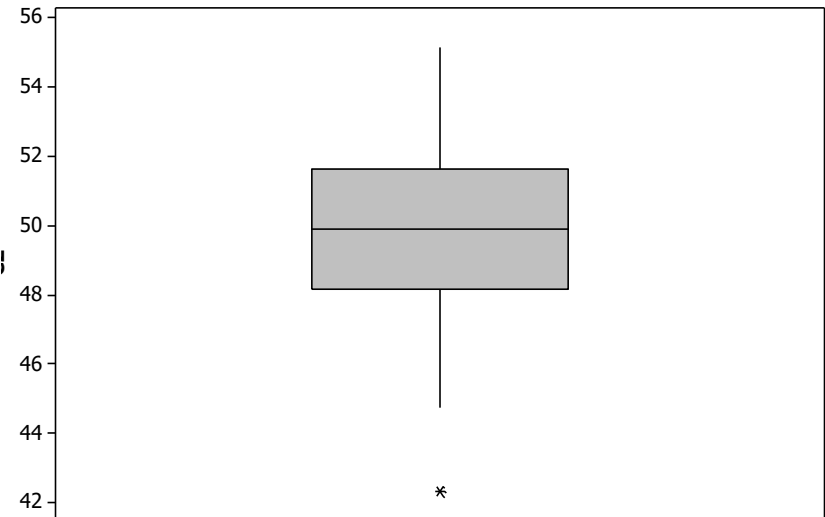
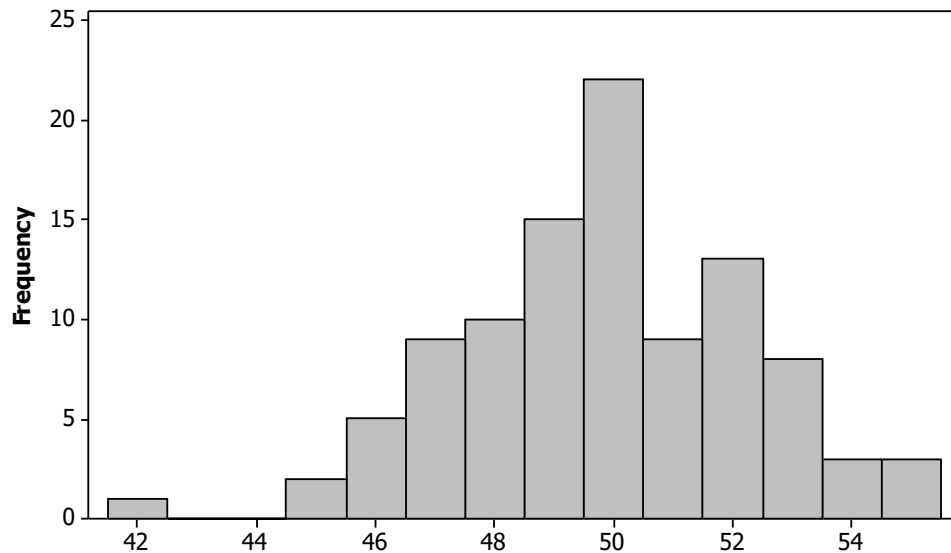
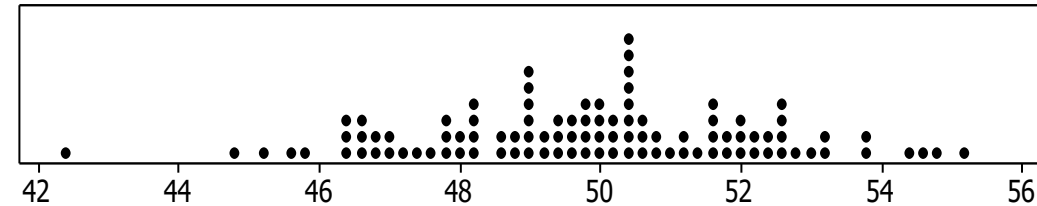


# The Normal Curve and Probability Areas Associated with the Standard Deviation

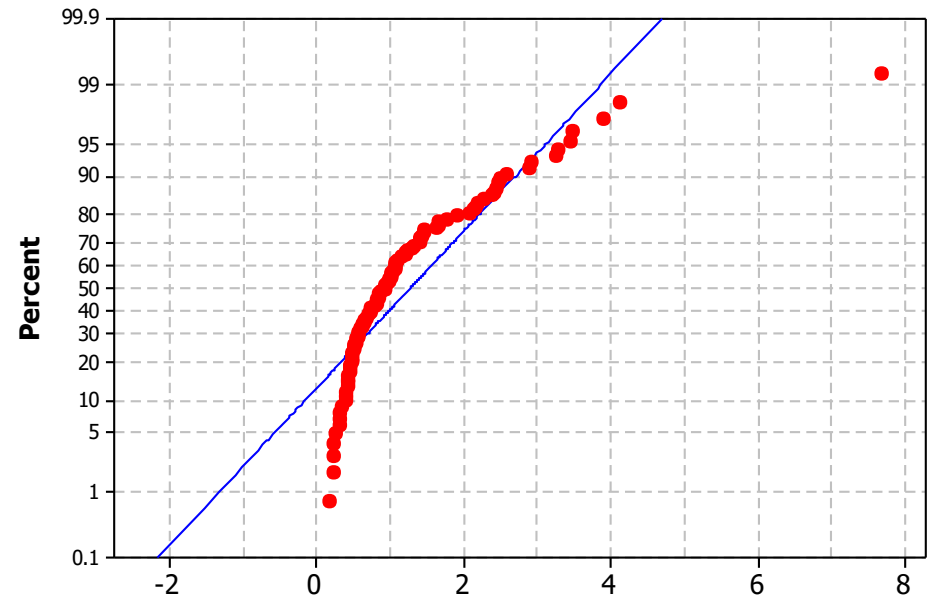
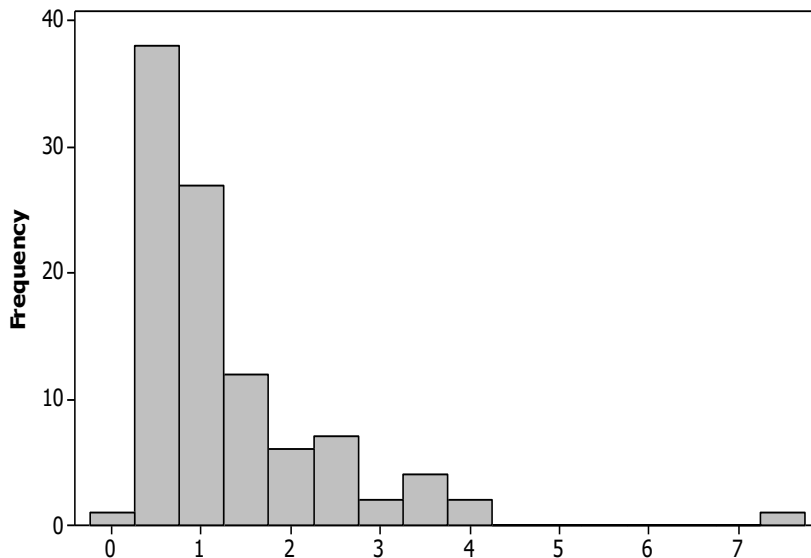
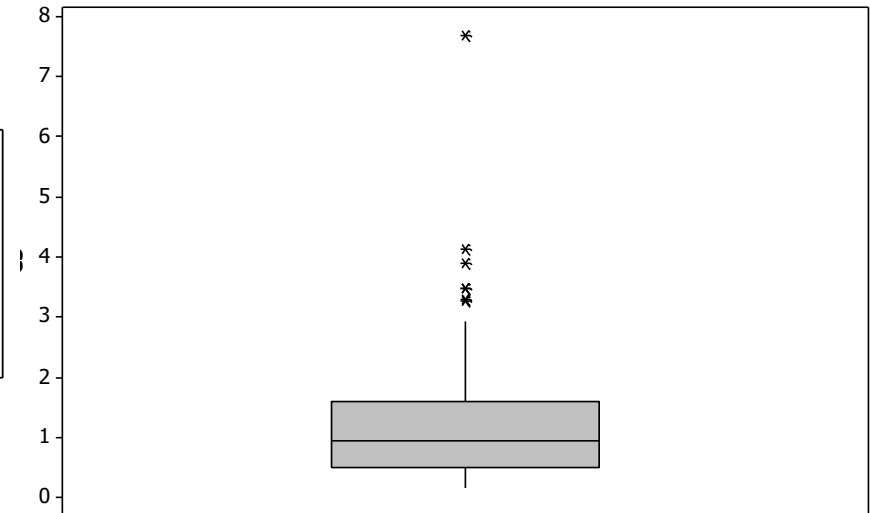
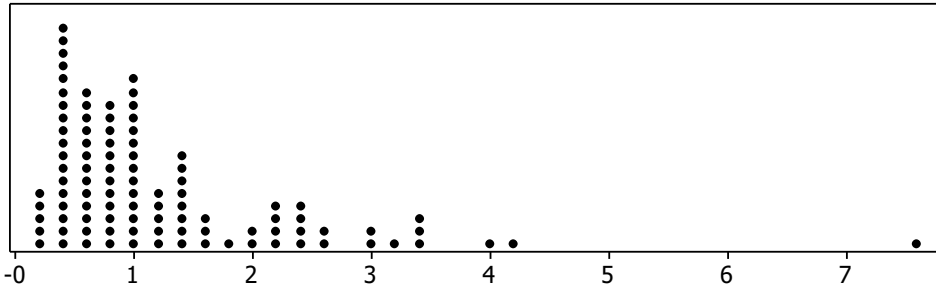
- Property 2: The area under the curve can be used to estimate the probability of certain “events” occurring



# Study Shape: Normal Data



# Study Shape: Non-Normal Data



# Measures of Center

## ■ Mean: arithmetic average of a set of values

- reflects the influence of all values
- strongly influenced by extreme values

Sample

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

Population

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

## ■ Median: center number after a set of numbers has been sorted

- is the 50% rank
- is “robust” to extreme scores

***When to use which measure?***

# Measures of Spread

- Range: distance between the extreme values of a data set (highest - lowest). The range is sensitive to extreme values.
- Variance: average squared deviation of each data point from the mean, Variance =  $s^2$  or  $\sigma^2$
- Standard Deviation: square root of the variance

Sample

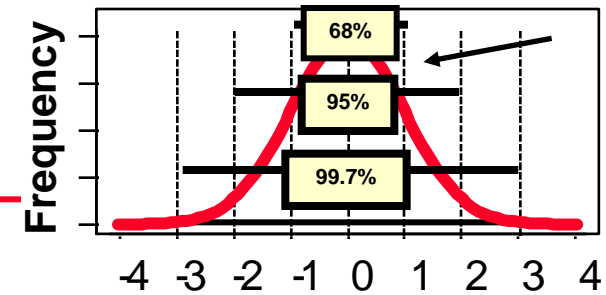
$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Population

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$$

***When to use which measure?***

# A Simple Example



- Calculate the standard deviation for the data : 2 1 3 5 4

	$X$	$X - \bar{X}$	$(X - \bar{X})^2$
1	2	-1	1
2	1	-2	4
3	3	0	0
4	5	2	4
5	4	1	1
$\Sigma$	15		10
Mean	3		
s-square			2.5
s			1.58

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

$$\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

# Exercise: Turbocharger Data

---

- CFR : Variable Geometry Turbine Efficiency (%)
- Sample : 100 turbos measured in the Charleston test cell
- Objective : Study shape, center, spread, outliers
- Question : What is the % of turbos with efficiency 62% or higher?
- Project File : Basic Statistics.mpj
- Worksheet : data for exercises, Column 1

# Exploratory Data Analysis (EDA)

---

- In general, data analysis should follow some basic steps:
  - Take a practical look at the raw data to identify any abnormalities (errors, unexpected values)
  - Explore the data graphically to get a visual sense of the data
  - Analyze the data statistically to get a numerical sense of the data (we'll learn how to do this in future modules)

**P - Practical**  
**G - Graphical**  
**A - Analytical**





# Basic Analysis Steps ... We'll do !

---

- **Practical** – take a visual look at data in worksheet

- **Graphical**

  - Graph > Histogram

  - Graph > Dotplot

  - Graph > Boxplot

  - Graph > Probability Plot

  - Stat > Basic Statistics > Normality Test

- **Analytical**

  - Stat > Basic Statistics > Display Descriptive Statistics

  - Stat > Basic Statistics > Graphical Summary

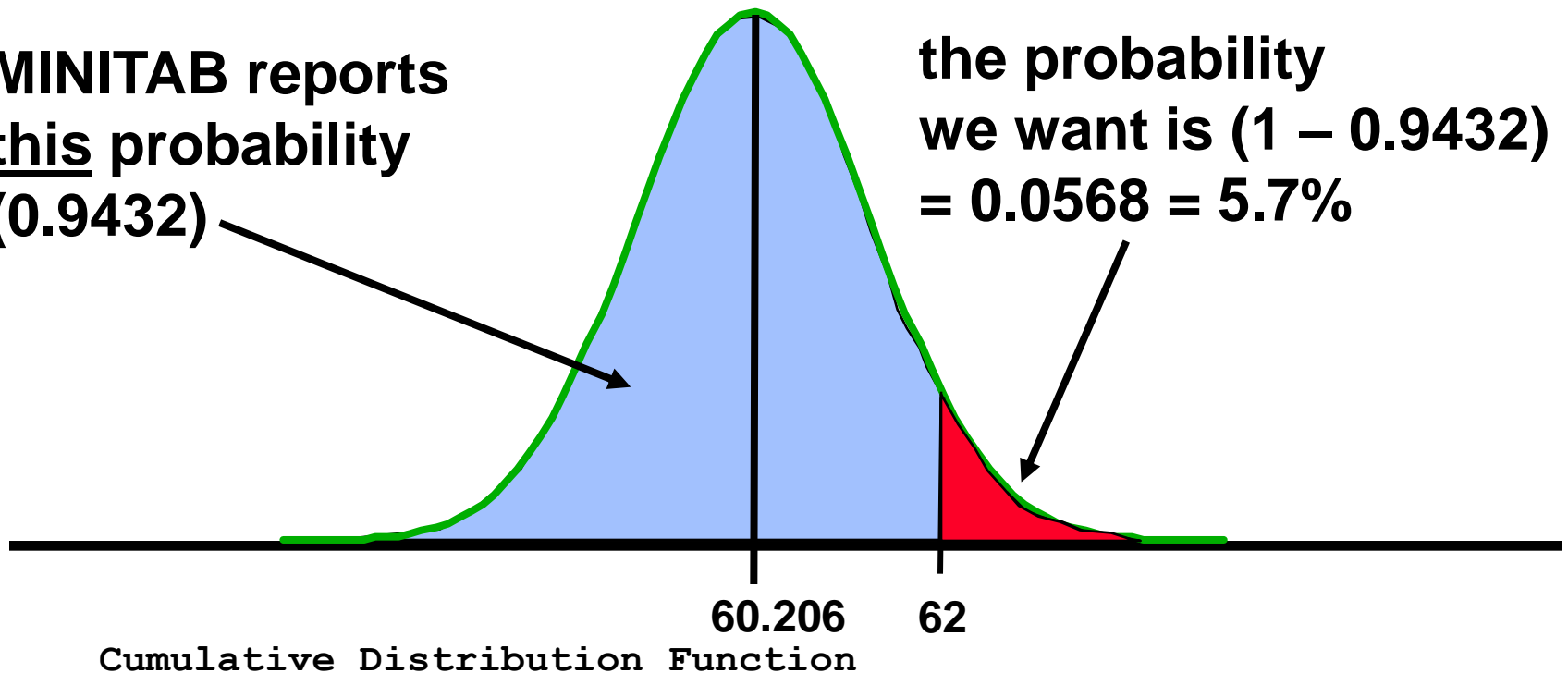
  - Calc > Probability Distributions > Normal (see next page)

# How to Use the Probabilities: Example

Calc > Probability Distributions > Normal

MINITAB reports  
this probability  
(0.9432)

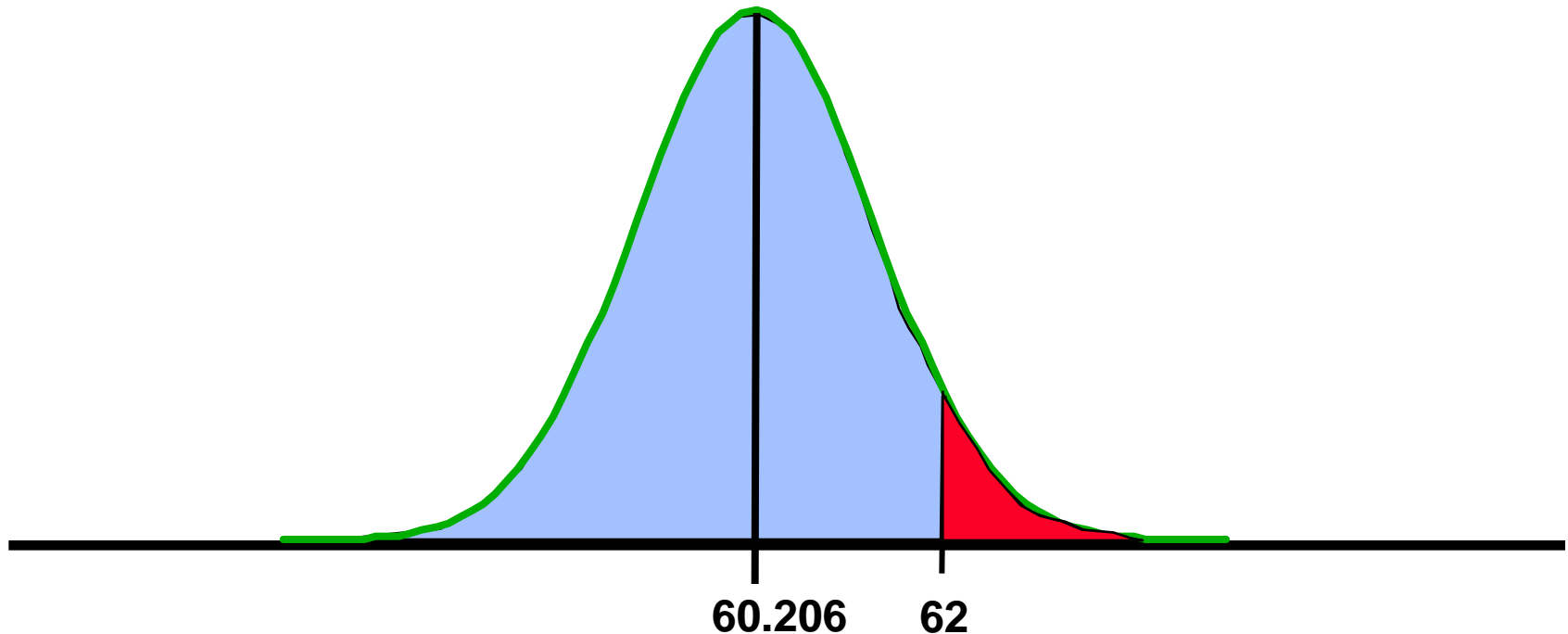
the probability  
we want is  $(1 - 0.9432)$   
 $= 0.0568 = 5.7\%$



Normal with mean = 60.206 and standard deviation = 1.134

x	P( X ≤ x )
62	0.9432

# How to Use the Probabilities: Reality Check



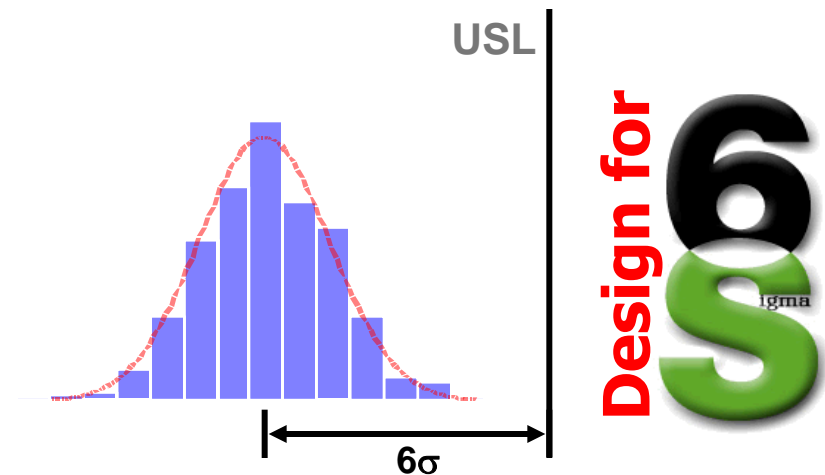
Does 5.7% make sense? How would you check it?

$$\bar{X} + 1s = 60.206 + 1(1.134) = 61.34 \Rightarrow 16\% \text{ above}$$

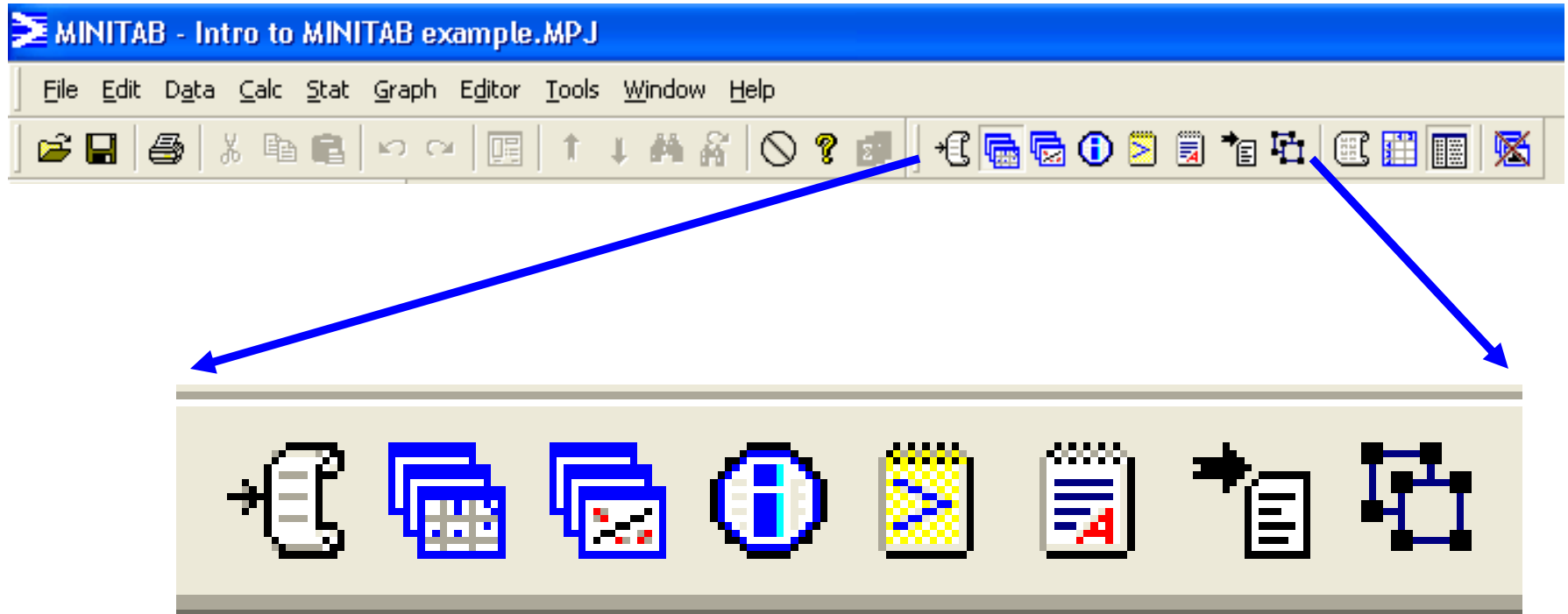
$$\bar{X} + 2s = 60.206 + 2(1.134) = 62.474 \Rightarrow 2.5\% \text{ above}$$

$$\bar{X} + 3s = 60.206 + 3(1.134) = 63.608 \Rightarrow 0.15\% \text{ above}$$

# MINITAB “Getting Around My File” Tips



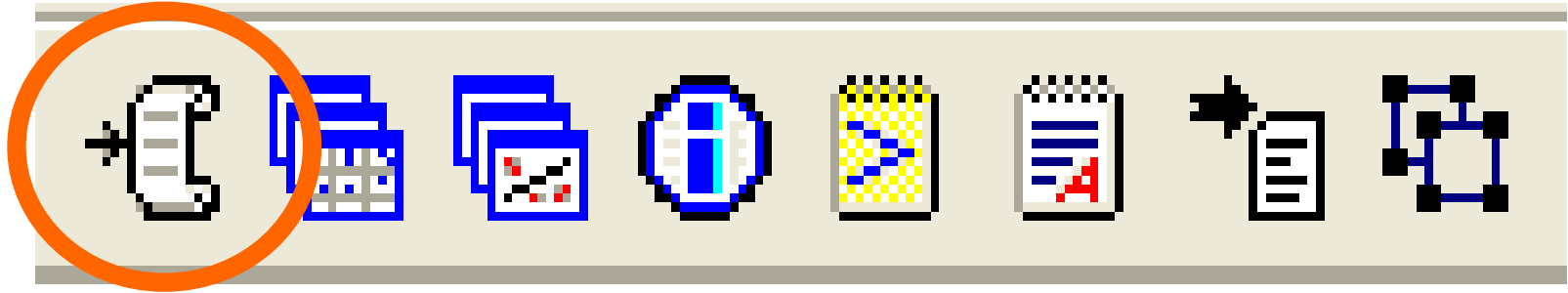
# Project Manager Tool Bar



- Open MINITAB file: Basic Statistics.mpj
  - Worksheet: xxxxxxxx
- A Project Manager toolbar allows quick & easy access to everything in your MINITAB project

# Session Folder

---



- Go to session folder by clicking this button once
- Double-clicking on a specific item takes you to that analysis in the session window
- right-mouse menu (edit title, delete, print, open StatGuide, etc.)

# Session Folder Example

**Project Manager**

- Session
  - CAPA for Height CPK=1.0
  - 12/6/2002 10:02:32 AM
  - Two-Sample T-Test and CI: Pin
  - Two-Sample T-Test and CI: Pin
  - Two-Sample T-Test and CI: Pin
  - Test for Equal Variances
  - Test for Equal Variances
  - Test for Equal Variances: Pin 1
  - Test for Equal Variances
  - Test for Equal Variances: Pin 1
  - Test for Equal Variances
  - Test for Equal Variances: Pin 2
  - Two-Sample T-Test and CI: Pin
  - Descriptive Statistics: Pin 1, Pin
  - Descriptive Statistics Graph: Pin
  - Descriptive Statistics Graph: Pin
  - Descriptive Statistics Graph: Pin
  - Descriptive Statistics: ran
  - Two-way ANOVA: Height versus Test, Pin
  - ANOVA: Height versus Test, Pin
  - One-way ANOVA: Height versus Test, Pin
  - One-way ANOVA: Height versus Test, Pin
  - One-way ANOVA: Height versus Test, Pin
  - One-way ANOVA: Height versus Test, Pin
  - ANOVA: Height versus Test, Pin
  - General Linear Model: Height versus Test, Pin
  - General Linear Model: Height versus Test, Pin
  - ANOVA: Height versus Test, Pin
  - 12/6/2002 2:19:13 PM
  - 2/9/03 7:34:21 PM
  - Factorial Design
  - Normal Prob Plot: Pin 1
  - 3/21/2003 8:23:35 AM
  - 3/21/2003 9:03:53 AM
  - 3/21/2003 10:13:56 AM
  - Descriptive Statistics: Data 1
  - 3/21/2003 10:42:57 AM
  - 6/9/2004 12:26:24 PM

**Session**

**ANOVA: Height versus Test, Pin**

Factor	Type	Levels	Values
Test	fixed	3	1-1 1-2 1-3
Pin	fixed	10	1 2 3 4 5 6 7 8 9 10

**Analysis of Variance for Height**

Source	DF	SS	MS	F	P
Test	2	0.00881	0.00440	0.43	0.655
Pin	9	0.12327	0.01370	1.35	0.282
Error	18	0.18326	0.01018		
Total	29	0.31534			

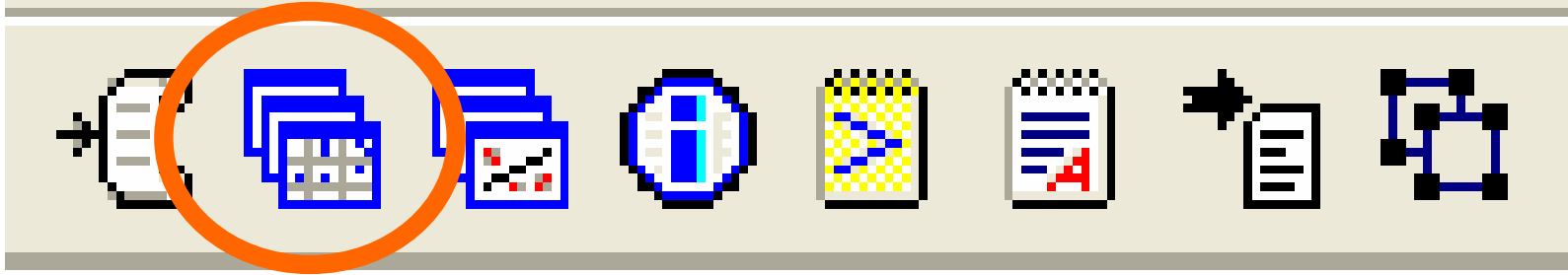
**One-way ANOVA: Height versus Pin**

**Analysis of Variance for Height**

Source	DF	SS	MS	F	P
Pin	9	0.12327	0.01370	1.43	0.242
Error	20	0.19207	0.00960		
Total	29	0.31534			

Individual 95% CIs For Mean

# Worksheets Folder



- Go to worksheet folder by clicking this button once
- Double-clicking on a specific worksheet brings it to the front, for analysis
- right-mouse menu (tile, print, save, close, rename, description, etc.)



# Worksheets Folder Example

MINITAB - Intro to MINITAB example.MPJ

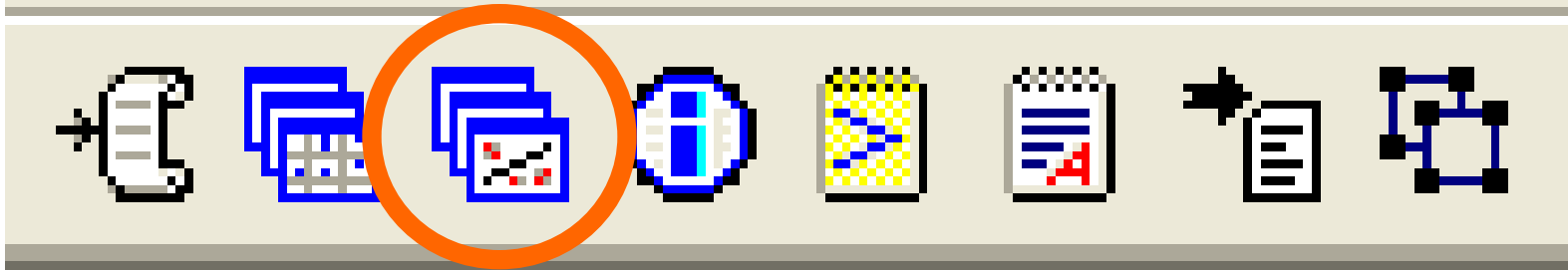
File Edit Data Calc Stat Graph Editor Tools Window Help

Project Manager: Worksheet 1, Pin DOE, Pin Data

Pin DOE \*\*\*

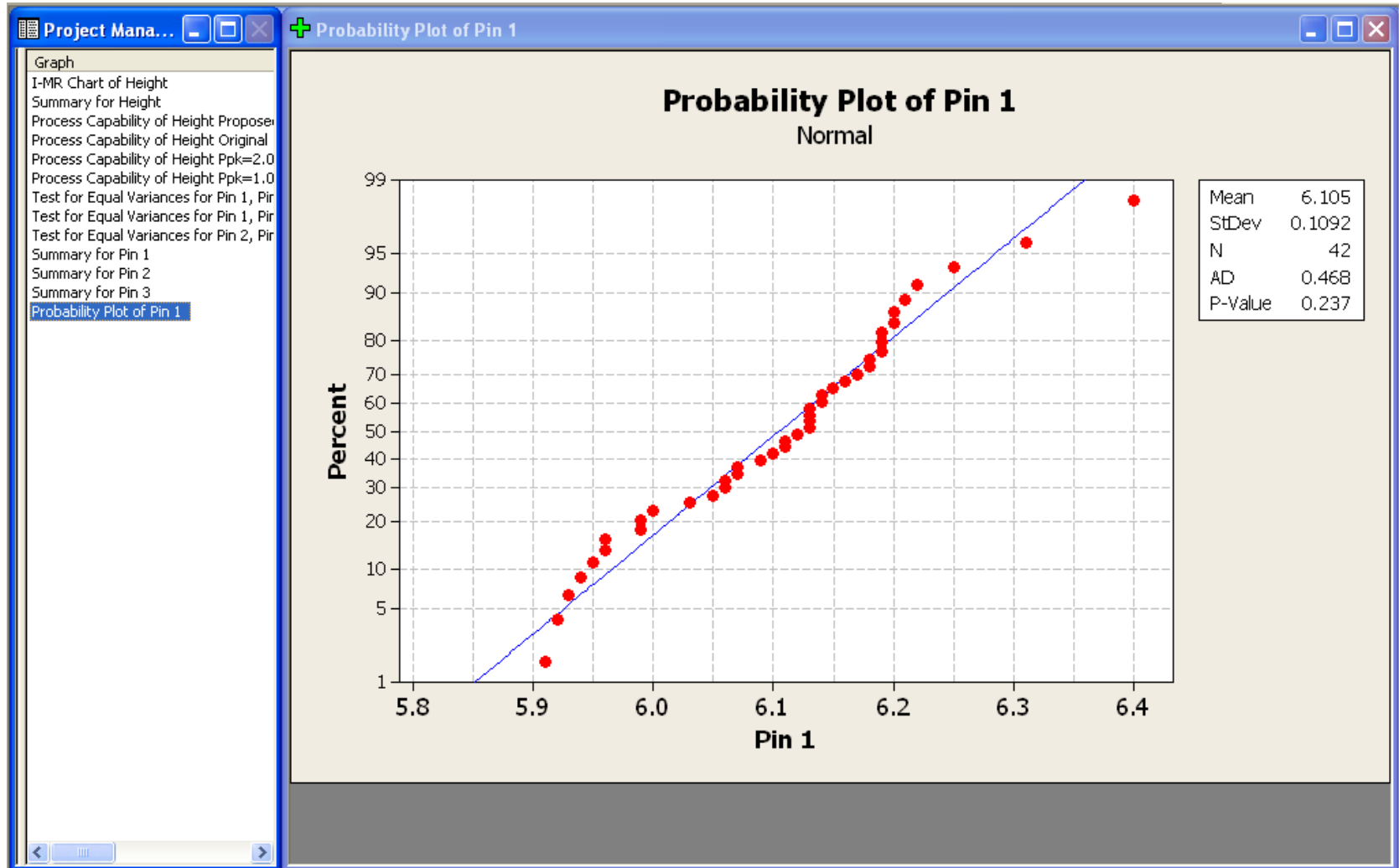
	C1	C2	C3	C4	C5	C6	C7
	StdOrder	RunOrder	CenterPt	Blocks	Pin Width	Pin Length	
1	2	1	1	1	3	5	
2	3	2	1	1	2	7	
3	1	3	1	1	2	5	
4	4	4	1	1	3	7	
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							

# Graphs Folder

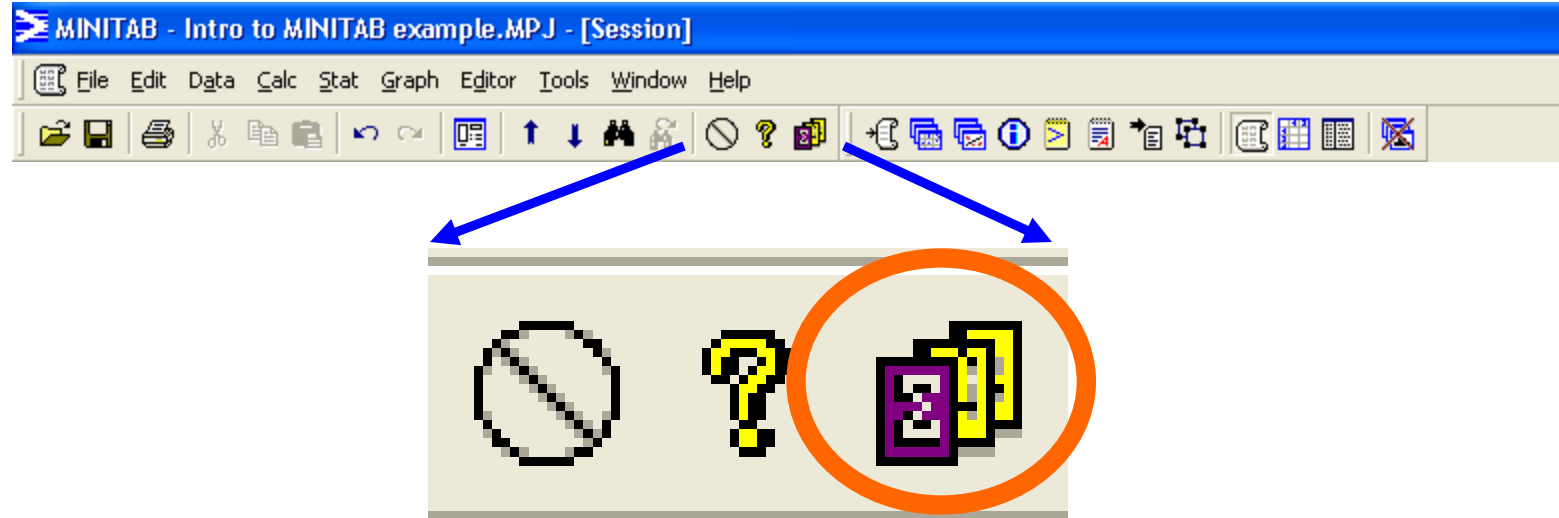


- Go to graph folder by clicking this button once
- Double-clicking on a specific graph brings it to the front
- right-mouse menu (tile, print, save, close, rename, copy, etc.)

# Graphs Folder Example



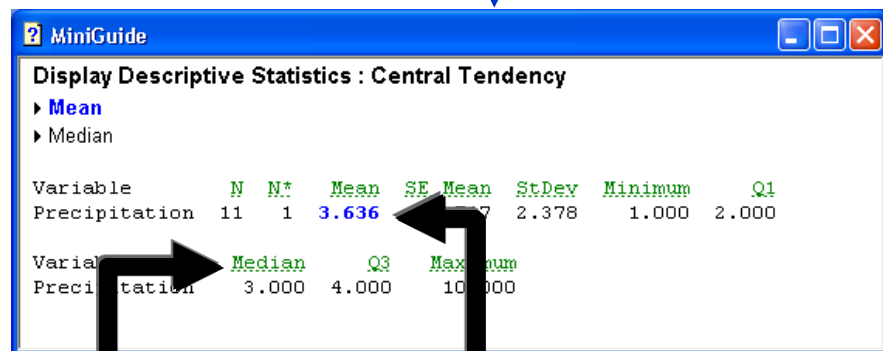
# StatGuide



- Provides statistical guidance on your analysis after you run a STAT procedure
- The tone is informal and practical, not like a textbook

# StatGuide Layout

Example



MiniGuide

Display Descriptive Statistics : Central Tendency

► Mean  
► Median

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1
Precipitation	11	1	3.636	0.7	2.378	1.000	2.000

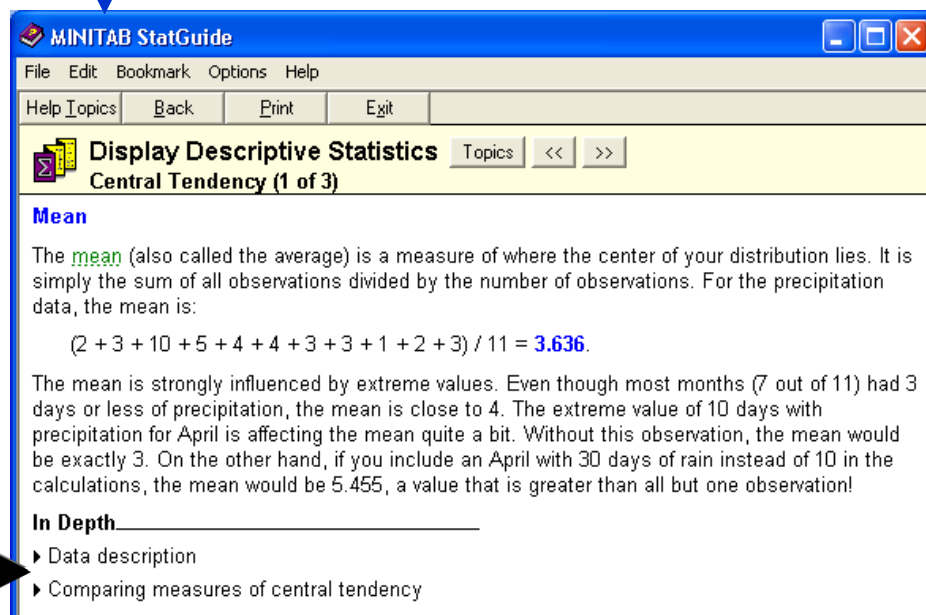
Variable	Median	Q3	Maximum
Precipitation	3.000	4.000	10.000

Arrows indicate that the 'Mean' value (3.636) in the first table is linked to the 'Mean' definition in the right panel, and the 'Median' value (3.000) in the second table is linked to the 'Median' definition in the right panel.

Relevant output highlighted

Hot-linked definitions

Detailed description



MINITAB StatGuide

File Edit Bookmark Options Help

Help Topics Back Print Exit

Display Descriptive Statistics Central Tendency (1 of 3)

Mean

The **mean** (also called the average) is a measure of where the center of your distribution lies. It is simply the sum of all observations divided by the number of observations. For the precipitation data, the mean is:

$$(2 + 3 + 10 + 5 + 4 + 4 + 3 + 3 + 1 + 2 + 3) / 11 = 3.636.$$

The mean is strongly influenced by extreme values. Even though most months (7 out of 11) had 3 days or less of precipitation, the mean is close to 4. The extreme value of 10 days with precipitation for April is affecting the mean quite a bit. Without this observation, the mean would be exactly 3. On the other hand, if you include an April with 30 days of rain instead of 10 in the calculations, the mean would be 5.455, a value that is greater than all but one observation!

In Depth

- Data description
- Comparing measures of central tendency

Arrows indicate that the 'Mean' definition in this panel is linked to the 'Mean' value in the MiniGuide window, and the 'In Depth' section is linked to the 'Links to additional topics' text.

Links to additional topics

# Breakout Exercise: Dominoes

---

- CFR : Thickness of dominoes
- Sample : a box of dominoes
- Objective : Study shape, center, spread, outliers of the box of dominoes you measure

# How to estimate spread with a complex Y (variance components)

- What if the response is a function of several random variables?
- Suppose the critical parameter is  $Y = \text{BSNOx} + \text{BSHC}$
- How can the parameters of Y be estimated?
- $\text{Mean}(Y) = \text{Mean}(\text{BSNOx}) + \text{Mean}(\text{BSHC})$
- Do we add standard deviations? NO!
- **Standard deviations do not add**
- Need to add Variances ( $\sigma^2$ )!

Think  
Pythagorean!

$$\hat{\sigma}_Y^2 = \hat{\sigma}_{\text{BSNOx}}^2 + \hat{\sigma}_{\text{BSHC}}^2$$

$$\hat{\sigma}_Y = \sqrt{\hat{\sigma}_{\text{BSNOx}}^2 + \hat{\sigma}_{\text{BSHC}}^2}$$

# Graphs : Priorities

---

- This is the case where we have data in categories and frequency within each category
  - test cell data on type of failure modes and how many
  - RPN's of potential causes of failure in an FMEA
- Use graphs to :
  - determine how to prioritize the work
  - look for trends or systematic issues with the data
- The graphs we use to study this type of data are :
  - pareto chart



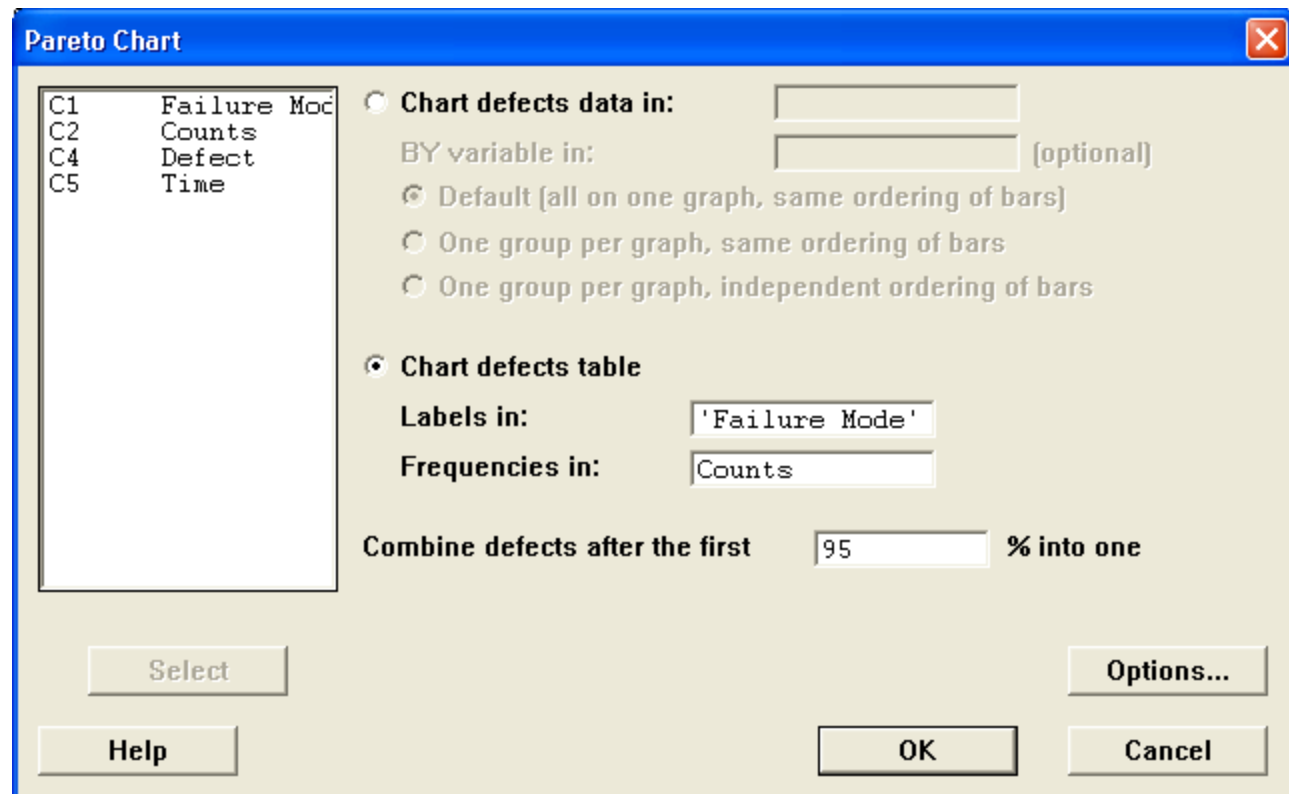
# Example : Graphs - Priorities

---

- Critical Parameter : Counts by Failure Mode
- Objective : To identify failure modes which account for 80% of the failures
- Question : What failure modes need to be addressed?
- Open worksheet “Pareto Data”

# Example : Graphs - Priorities

Stat > Quality Tools > Pareto Chart  
use defect table (C1 & C2)



The image shows the Minitab Pareto Chart dialog box. On the left, a list of variables includes C1 (Failure Mode), C2 (Counts), C4 (Defect), and C5 (Time). The 'Chart defects data in:' section has three radio buttons: 'Default (all on one graph, same ordering of bars)' is selected, followed by 'One group per graph, same ordering of bars' and 'One group per graph, independent ordering of bars'. The 'Chart defects table' section is also selected, with 'Labels in:' set to 'Failure Mode' and 'Frequencies in:' set to 'Counts'. The 'Combine defects after the first' field is set to 95, followed by '% into one'. At the bottom, there are buttons for 'Select', 'Options...', 'Help', 'OK', and 'Cancel'.

Variable	Description
C1	Failure Mode
C2	Counts
C4	Defect
C5	Time

**Pareto Chart**

☐ Chart defects data in:   
BY variable in:  (optional)

☒ Default (all on one graph, same ordering of bars)  
☐ One group per graph, same ordering of bars  
☐ One group per graph, independent ordering of bars

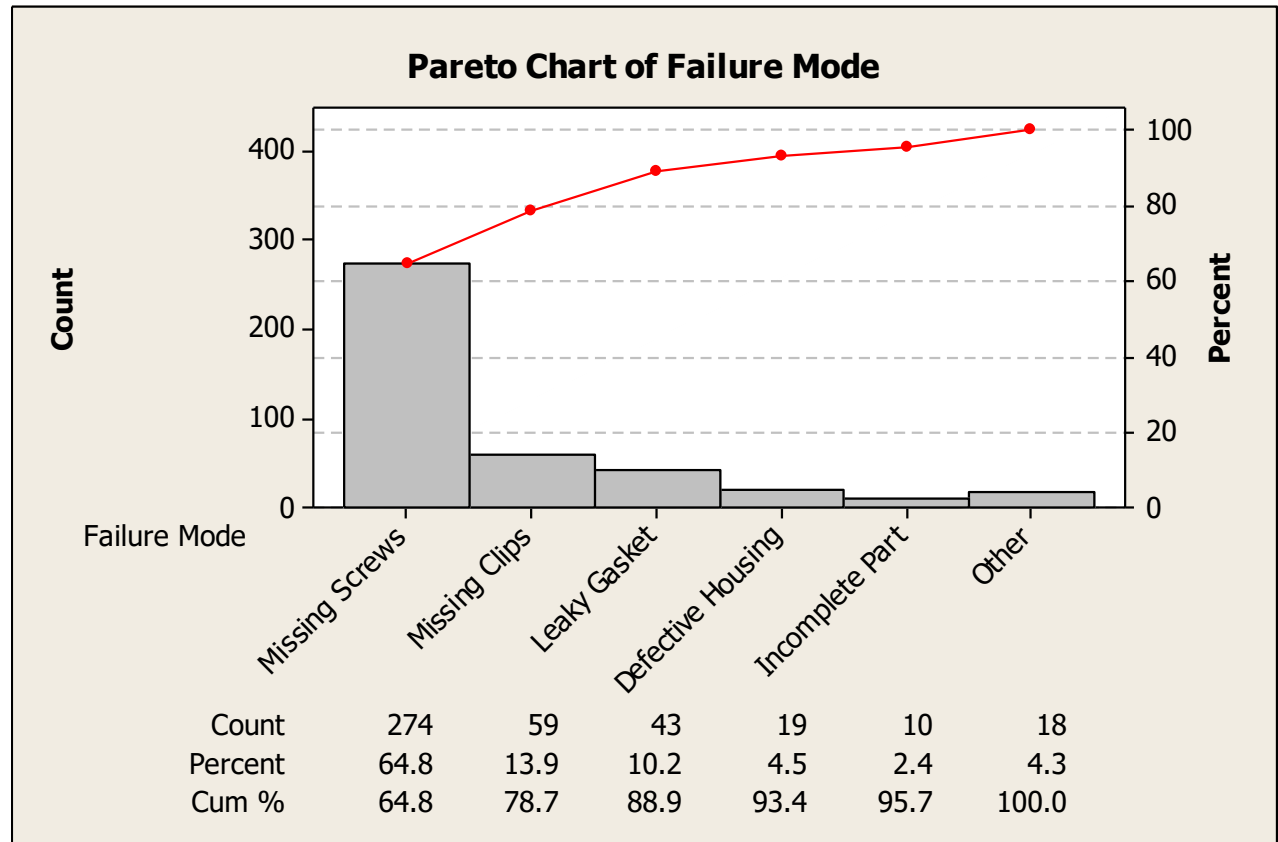
☒ Chart defects table  
Labels in:   
Frequencies in:

Combine defects after the first  % into one

Select Options... Help OK Cancel

# Example : Graphs - Priorities

Failure Mode	Counts
Missing Screws	274
Missing Clips	59
Defective Housing	19
Leaky Gasket	43
Scrap	4
Unconnected Wire	8
Missing Studs	6
Incomplete Part	10



# Example : Graphs - Priorities

---

- Critical Parameter : Defect Count
- Objective : Determine if there are any systematic issues in the data that need to be addressed
- Question : Is there an effect due to Time?
- Open worksheet “Pareto Data”

# Example : Graphs - Priorities

Stat > Quality Tools > Pareto Chart  
use defect data (C4 & C5)

**Pareto Chart**

C1	Failure Mode
C2	Counts
C4	Defect
C5	Time

☒ **Chart defects data in:** Defect

**BY variable in:** Time (optional)

☒ **Default** (all on one graph, same ordering of bars)

☐ One group per graph, same ordering of bars

☐ One group per graph, independent ordering of bars

☐ **Chart defects table**

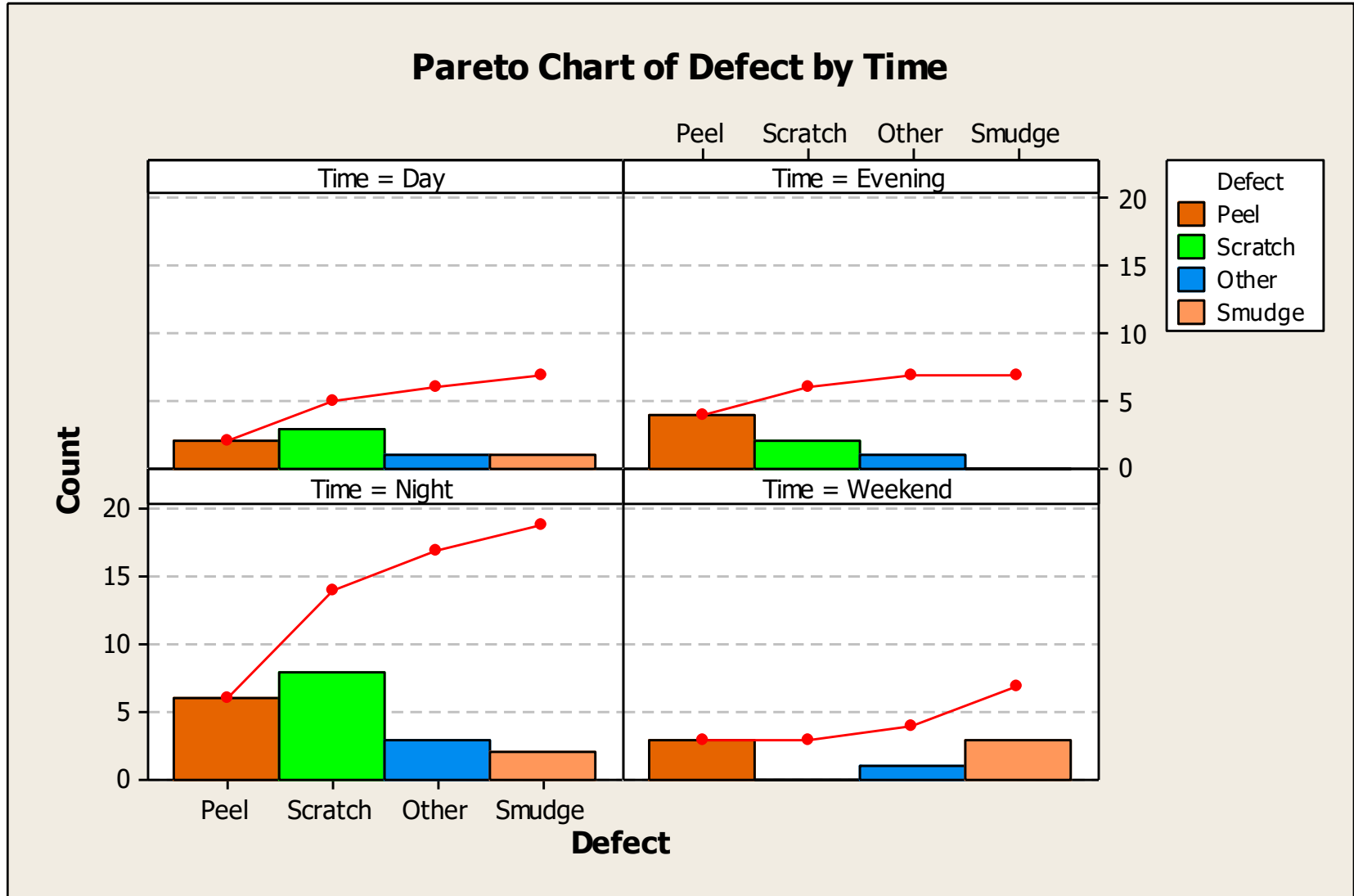
**Labels in:** 'Failure Mode'

**Frequencies in:** Counts

**Combine defects after the first** 95 % into one

Select Options... Help OK Cancel

# Example : Graphs - Priorities



# Example Summary : Priorities

---

## ■ For the Failure Mode data

- need to eliminate missing screws & missing clips
- perhaps these can go to zero
- typically never fix a failure mode 100% unless you remove the component
- would probably want to address the 3rd failure mode as well to ensure 80% reduction

## ■ For the Defect data by Time

- weekend had no scratches - WHY?
- would want to identify what in the process is different
- a watch-out is that the counts aren't normalized by opportunity (don't know total production by time)

# Objectives Revisited

---

Should be able to:

- Use graphs to visually represent variation
- Assess normality using the probability plot
- Describe a variable's distribution using graphs as well as shape, center, and spread metrics
- Use the normal distribution to calculate the probability of an outcome