

Analyze Purpose

- To determine the root causes
- Estimate population parameters with confidence intervals
- To construct hypothesis about the data and test them to determine significance

SIMPLER. FASTER. BETTER. LESS COSTLY.



Analyze Deliverables

- Update data collection plan
- Hypothesis testing
- Fishbone root cause analysis
- MSA analysis
- Analysis summary
- Update control charts

SIMPLER. FASTER. BETTER. LESS COSTLY.



Objectives

- Upon completion of this module, the Green Belt will understand:
 - Descriptive Statistics
 - Probability Distributions
 - Graphical Tools

SIMPLER. FASTER. BETTER. LESS COSTLY.



Types of Basic Descriptive Statistics

The Key Characteristics of Data	Statistics for Discrete Data	Statistics for Continuous Data
<ul style="list-style-type: none">CenterSpreadShapeStability over Time	<ul style="list-style-type: none">Defectives (Units)Statistics for Defects (Errors)	<ul style="list-style-type: none">T-TestANOVARegression

Data varies across its measurement scale

- Any set of data will have values that distribute across the measurement scale.
 - This is called a *data distribution*, or simply "distribution".
- Except in the rarest of circumstances, data will vary...even when nothing in the process seems to be changing.
 - Truth: Something is changing to cause the variation, we just may not be able to see what/how is changing.
- Knowing the data type and distribution is critical to choosing the right statistical tools.

SIMPLER. FASTER. BETTER. LESS COSTLY. **LEANOhio**

Key Terms

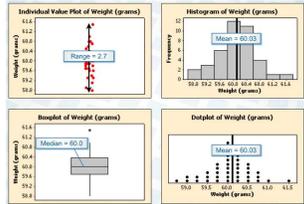
Statistics is the art and science of describing, interpreting, and analyzing data.

Descriptive statistics summarize important characteristics of the data such as central tendency or spread.

SIMPLER. FASTER. BETTER. LESS COSTLY. **LEANOhio**

Descriptive Statistics

We look at a variety of **Descriptive Statistics** to provide insights into different aspects of data –they provide **summary information** that we can't see in graph or a table of data values (for example, range, mean and median).



Four Characteristics of Data

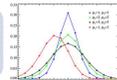
SIMPLER. FASTER. BETTER. LESS COSTLY.



The Key Characteristics of a Distribution

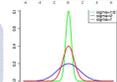
Where on the measure scale does the data appear to gather or "clump"?

• What is the **center** of the data?



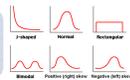
How does the data distribute around the center?

• What is the **spread** of the data?



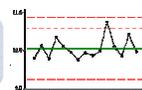
What values are more frequent and less frequent?

• What is the **shape** of the data?



How do the above characteristics behave over time?

• What is the **stability** of the data?



Four Key Characteristics of Continuous Data

Center	Spread (Variation)	Shape	Stability
<ul style="list-style-type: none">• Mean• Median• Mode	<ul style="list-style-type: none">• Range• Standard Deviation• Variance	<ul style="list-style-type: none">• Skew• Kurtosis	<ul style="list-style-type: none">• Control Charts

Measures of Center

Mean	<ul style="list-style-type: none">• The mathematical average of a set of data point values. (Sum of all data points/number of data points)
Median	<ul style="list-style-type: none">• The middle data point when the data is sorted by value, where 50% of the observed values are below and 50% are above. If there is an even number of data points, then average the two points in the middle
Mode	<ul style="list-style-type: none">• The most frequently occurring data point values

Calculating the Mean

Example: Average Custom Order Time (wks):
8 data points: 5,6,3,4,2,5,5,3

- 1) Add all 8 numbers
– $5+6+3+4+2+5+5+3=33$
- 2) Divide by number of occurrences
– There are 8 occurrences so: $33/8$
– $=4.125$
– Therefore the Mean or average= 4.125

SIMPLER. FASTER. BETTER. LESS COSTLY. **LEANOhio**

Calculating the Median

- ▶ Sort the data (from the lowest value to the highest value)
 - Take the 8 data points: 5,6,3,4,2,5,5,3
 - Sort in order from lowest to highest: 2,3,3,4,5,5,5,6
- ▶ The median is the middle number if the sorted data set has an odd number of observations.
 - This data set has an even number so there is no exact middle
- ▶ Otherwise, it is the average of the two middle numbers
 - 4 and 5 are the two middle points
 - Add $4+5=9$
 - Divide by 2: $9/2=4.5$
 - Median= 4.5

SIMPLER. FASTER. BETTER. LESS COSTLY.



Determining the Mode

- The mode is the most frequently occurring value in a data set.
 - Dataset: 5,6,3,4,2,5,5,3
 - Occurrence of each number:
 - 2=1 occurrence
 - 3= 2 occurrences
 - 4= 1 occurrence
 - 5= 3 occurrences
 - 6= 1 occurrence

SIMPLER. FASTER. BETTER. LESS COSTLY.



Do I need a fancy Stats Program?

- You do not need a complex stats program to compute the type of data you will collect for your project
- Excel is capable of computing Descriptive Statistics, Charts and Graphs, and most statistical analyses
- Minitab is much easier and faster in most cases but it is not required
- Use YouTube videos if you get stuck in Minitab or Excel!

SIMPLER. FASTER. BETTER. LESS COSTLY.



Excel Central Tendency Exercise

- Turn on Data Analysis Package in Excel
 - Go to “File”, “Options”
 - Click “add-ins”
 - Click “go”
 - Select “Analysis Toolpak” and then “ok”
 - “Data Analysis” should appear under the “Data” tab on the far right.
- Calculate the Mean, Median, and Mode for the following data set using Excel.
 - 1,5,9,12,4,8,10,7,5,6

Mean	6.7
Median	6.5
Mode	5

Four Key Characteristics of Continuous Data

Center	Spread (Variation)	Shape	Stability
<ul style="list-style-type: none"> • Mean • Median • Mode 	<ul style="list-style-type: none"> • Range • Standard Deviation • Variance 	<ul style="list-style-type: none"> • Skew • Kurtosis 	<ul style="list-style-type: none"> • Control Charts

Variation

- Accuracy vs. Precision
- Measures of Variation
 - Range
 - Standard Deviation
 - Variance

SIMPLER. FASTER. BETTER. LESS COSTLY.



Accuracy vs. Precision

Accuracy describes Centering



How close to target?

Precision describes Spread



How close together?

Range

- **Range:** is the difference between the maximum value and the minimum value in a data set.
 - Range = Maximum Value - Minimum Value
- The purpose is to measure the dispersion (range) between the highest and lowest values of a data set.

Range is the difference between the maximum value and the minimum value in a data set.

SIMPLER. FASTER. BETTER. LESS COSTLY. **LEANOhio**

Calculating the Range

- Example: Determine the range for the following data set – the number of abandoned calls per day in Customer Service:
 - 20, 32, 19, 45, 30, 23, 5, 29, 34, 26, 18
 - Arrange from highest to lowest:
 - 5, 18, 19, 20, 26, 26, 29, 30, 32, 34, 45
 - Take highest number minus lowest number:
 - $45 - 5 = 40$
 - Range= 40

SIMPLER. FASTER. BETTER. LESS COSTLY. **LEANOhio**

Deviation

► Deviation: This is the distance between a data point value and the mean.

$$Deviation = (X - \bar{X})$$

► These deviations for each data point will be used to calculate and describe the variation in a set of data.

SIMPLER. FASTER. BETTER. LESS COSTLY. **LEANOhio**

Deviation

- Deviation = distance from mean

Variance

- Variance: accounts for all data by measuring the distance or difference between each value and the mean. These differences are called deviations.
- To compute variance, square each deviation. Add them to get the sum of the squared deviations. And Finally, divide by the number of values minus one. Roughly speaking, the variance is the average of the squared

The **variance** is a measure of variation in a data set. It is the sum squared deviations divided by the number of values minus one.

Standard Deviation

- We can use standard deviation to quantify variability in the same units as we measure our data. The Standard Deviation is the square root of the variance. We use S^2 to denote the variance, and S to denote the standard deviation.

The **standard deviation** is the square root of the variance. It measures variation in data units.

Standard Deviation

- Standard Deviation: Measures the average dispersion about the mean.
 - On average- how far are values from the mean?
 - Average Deviation from the mean
- Population Standard Deviation (σ):

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

Excel Example

- Using the following dataset, calculate the standard deviation and variation.
 - Number of grant applications filed as incomplete:
 - 20, 87, 29, 46, 42, 76, 62, 55, 38, 80
 - Standard Deviation= 22.50
 - Variance= 506.28

SIMPLER. FASTER. BETTER. LESS COSTLY.



Minitab Example

- You can run Standard Deviation and Variance statistics on multiple data sets at the same time to get a visual comparison.
 - This can be done in Excel or Minitab, just make sure each dataset is in its own colum.

SIMPLER. FASTER. BETTER. LESS COSTLY.



Four Key Characteristics of Continuous Data

Center	Spread (Variation)	Shape	Stability
<ul style="list-style-type: none">• Mean• Median• Mode	<ul style="list-style-type: none">• Range• Standard Deviation• Variance	<ul style="list-style-type: none">• Skew• Kurtosis	<ul style="list-style-type: none">• Control Charts

Shape

- Quantitative Methods
 - Skewness and Kurtosis
 - Goodness of Fit
 - **Normality Tests**
- You can compute this data in Excel or Minitab by running descriptive statistics

SIMPLER. FASTER. BETTER. LESS COSTLY.



Probability Distributions

- **Probability Distribution:** This is the tendency of a large numbers of observations from a process to group themselves around some central value with a certain amount of variation or “scatter” on either side.

SIMPLER. FASTER. BETTER. LESS COSTLY.



Probability Distributions

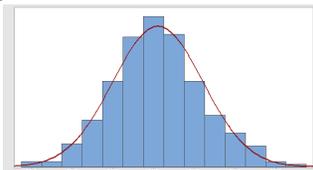
- There are many types of probability distributions.
 - Binomial, Poisson, Uniform, Normal, Beta, Exponential, Weibull, Gamma, etc.
- Probability distributions are either discrete or continuous. It depends on the random variable.

SIMPLER. FASTER. BETTER. LESS COSTLY.



Normal Distribution

- The **Normal Distribution** (Gaussian Distribution) is the classic bell-shaped curve which approximately describes many phenomenon in industry and science, and it is always the distribution of sample means from any distribution.



The Normal Distribution

- The “Normal” Distribution is a distribution of data which has certain consistent properties.
- These properties are very useful in our understanding of the characteristics of the underlying process from which the data were obtained.
- Many natural phenomena and man-made processes are distributed normally, or can be closely approximated by the normal distribution.

SIMPLER. FASTER. BETTER. LESS COSTLY. **LEANOhio**

The Normal Distribution

► A normal distribution can be described completely by knowing only the:

- mean and variance (or standard deviation)

What is the difference among these three normal distributions?

The Normal Curve and Probability

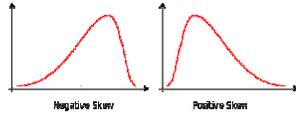
► The area under the curve can be used to estimate the probability of a certain measurement value occurring

Probability of sample value

Number of standard deviations from the mean

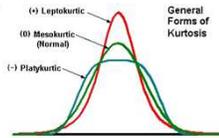
Skewness

- Skewness refers to a lack of symmetry. A distribution is skewed if one tail extends farther than the other. A value for skewness is included with the graphical summary:
 - A value close to 0 indicates symmetric data
 - Positive values imply a positive or right skew
 - Negative values imply negative or left skew



Kurtosis

- Kurtosis refers to how sharply peaked a distribution is. A value for kurtosis is included with the graphical summary:
 - Values close to 0 indicate normally peaked data.
 - Negative values indicate a distribution that is flatter than normal.
 - Positive values indicate a distribution with a sharper than normal peak.



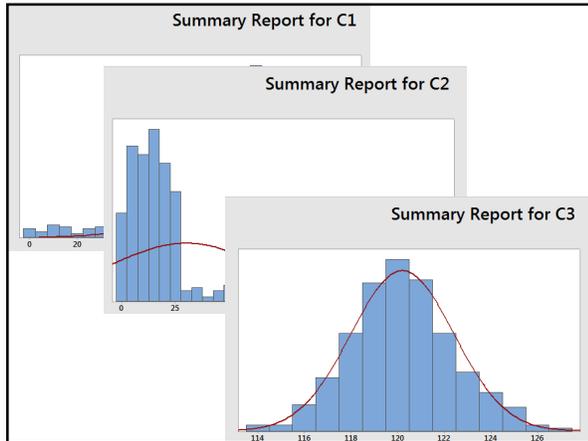
Shape Example

► Visualizing Skew and Kurtosis

Variable	Mean	StDev	Variance	Skew	Kurtosis
C1	100.64	25.97	643.84	-1.397	2.460
C2	29.982	36.404	1325.21	1.82	2.20
C3	120.21	2.21	4.89	.188	.102

► What would each distribution look like?

- | | |
|--|--|
| <p><u>Skew:</u></p> <ul style="list-style-type: none"> • Value close to 0 = symmetric data • Negative values= negative/left skew • Positive values= positive/right skew | <p><u>Kurtosis:</u></p> <ul style="list-style-type: none"> • Value close to 0 = normally peaked data. • Negative values= a flatter than normal peak. • Positive values= a sharper than normal peak. |
|--|--|



Four Key Characteristics of Continuous Data

Center	Spread (Variation)	Shape	Stability
<ul style="list-style-type: none"> • Mean • Median 	<ul style="list-style-type: none"> • Range • Standard Deviation • Variance 	<ul style="list-style-type: none"> • Skew • Kurtosis 	<ul style="list-style-type: none"> • Control Charts

Stability

- Is the process changing over time?
- Are there any trends, clusters, oscillations, etc?
- Run Charts and Control Charts will help determine if the process is stable.
 - This topic will be covered in more detail
- A stable process is a process which is free of assignable causes (in statistical control)

SIMPLER. FASTER. BETTER. LESS COSTLY. **LEANOhio**

Questions

SIMPLER. FASTER. BETTER. LESS COSTLY. **LEANOhio**
