Excel Statistics

240 CenSARA

Instructor: Steve Hiebsch
Email: steve.hiebsch@gmail.com

Course's Agenda

TIME	DAY 1	DAY 2	DAY 3
8:30 – 8:50	Set up Check-In	Check-In & Review Day 1	Check-In & Review Day 2
8:50 – 9:40	Introductions Pre-Test	Module 3 Central Tendency	Module 5 Linear Regression
9:40 – 10:35	Module 1 Excel Basics	Module 3 Practice	Module 5 Practice
10:35 – 10:50	Break	Break	Break
10:50 – 11:40	Module 2 Describing Data	Module 4 Est & Confidence Interval	Module 6 Lognormal Distribution
11:40 – 12:30	Module 2 Practice	Module 4 Practice	Module 6 Practice
12:30 – 12:50		Review	Post-Test
12:50 – 1:00	Q&A over day	Q&A over day	

Introductions

- Introduce yourself
  - Name and position
  - Education and/or work background
  - Something more personal about yourself e.g. hobbies, special talent, something you accomplished of which you are proud, something nobody knows about you, etc. (optional)

Start Pre-Test

- Open Microsoft Word file named "0 Pre-Test Excel…" on the Website for this course. Then follow the instructions.

Pre-Test 240CenSARA

Purpose of Class

- Emphasis is on learning to use Microsoft Excel for statistical purposes.
  - While we will have some amount of statistical learning, this is *not* the main emphasis.
  - We will use some environmental type data, but again that is not the primary purpose.
- Primary Purpose is to prepare student for some classes where there is need to use Excel in specific Statistical Analysis.

Module 1: Using Microsoft Excel Worksheets

After completing this module, the student will be able to:

1. Understand why Excel is useful as a statistical tool
2. Define what is meant by a worksheet and a workbook
3. Enter data into a worksheet
4. Create formulas and solve problems with a worksheet using basic arithmetic functions
5. Edit data that is in a worksheet
6. Edit data by using right-click functions, copy with cross hairs, F4 key cell ($) modifiers

Module 2: Describing Data - Graphical Presentations

After completing this module, the student will be able to:

1. Use Excel to create common graphic presentations as pie chart, bar charts, simple histograms, line charts, and scatter plots.
2. Edit and modify charts

Module 3: Describing Data – Measures of Central Tendency

After completing this module, the student will be able to:

1. Explain the characteristics and uses of measures of central tendency
2. Explain the characteristics and uses of measures of dispersion
3. Use Excel functions to calculate the arithmetic mean, median, mode, and standard deviations
4. Use Excel's Analysis ToolPak add-in to find measures of central tendency and dispersion

Module 4: Estimation & Confidence Interval

After completing this module, the student will be able to:

1. Define point estimate
2. Define level of confidence
3. Use Excel to calculate a confidence interval for a population when the sample size is 30 or larger
   AND when the sample size is 30 or less.

Module 5: Linear Regression

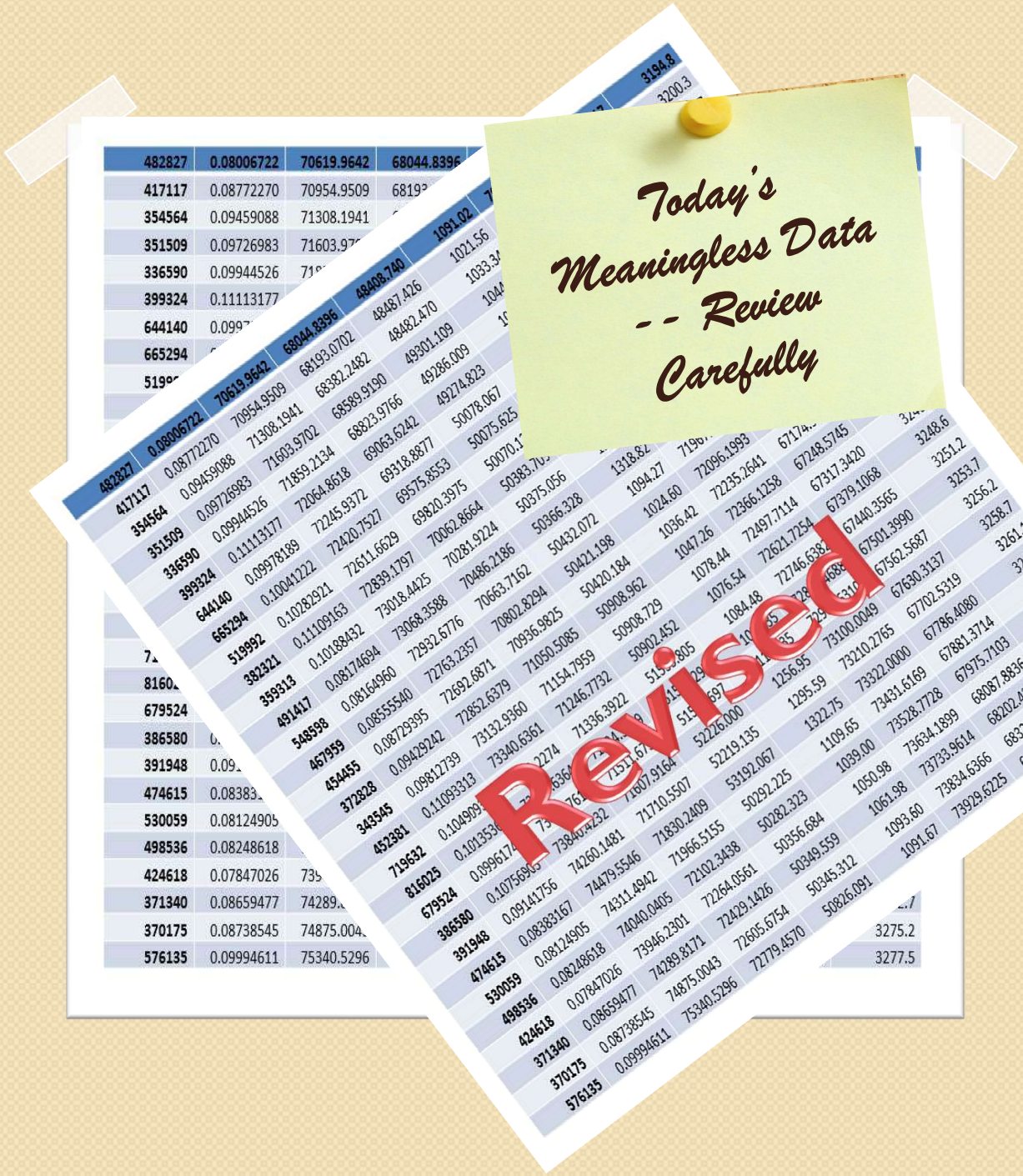
After completing this module, the student will be able to:

1. Explain linear regression
2. Use Excel to draw a scatter diagram
3. Use Excel to find a least squares regression line
4. Use Excel and the least squares regression equation to predict the value of a dependent variable based on an independent variable

Module 6: Lognormal Function

After completing this module, the student will be able to:

1. Explain lognormal distribution
2. Use Excel Lognormal functions to create a lognormal distribution probability



Module 1: Using Microsoft Excel Spreadsheets

What is a Spreadsheet?

- **A computer program** that stores data in a tabular format.
- A computer program with features and/or capabilities that include
  - Calculations of formulas
  - Production of charts and graphics
  - Data analysis tools capable of handling large quantities of data.

What is a Spreadsheet?

- Spreadsheet packages are used to help the user understand and solve numerical problems
  - Used in almost every field of business, government, and academia.
- Microsoft Excel is popular spreadsheet package
- Others
  - VisiCalc – one of the original
  - Lotus 1-2-3 – first big package (still available)
  - Open Office – originally by Oracle
  - WordPerfect Office – Quattro Pro
  - Microsoft – Works (in 2009 out of production) became Microsoft – Office Starter Edition
  - Online Spreadsheet – Google Docs

Basics of Excel

- Spreadsheet packages use worksheets of columns and rows to view the data.
- A collection of worksheets make up a workbook.
  - Old default in MS Excel was 3 worksheets
- A *spreadsheet* program, such as Excel, is used to create *workbook* files that contain one or more *worksheets* which contains tabular data.

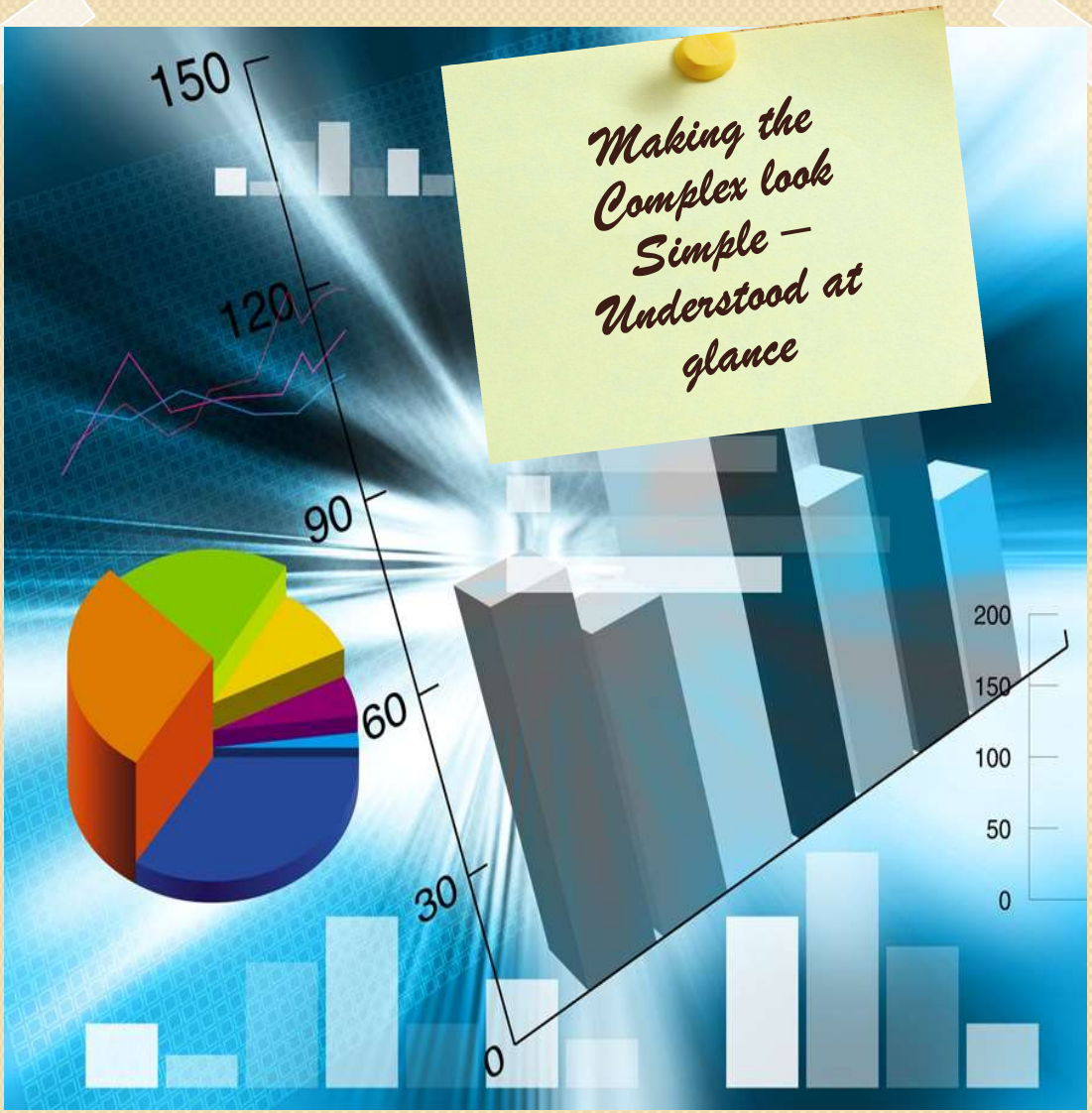
Excel as a Statistical Tool

- Useful functions in spreadsheet package
  - Enter data that is related
  - Develop relationship using math & statistical tools
  - Look at results when changes occur
- Present data in "more easily" understandable manner
  - Charts
  - Tables
  - Graphs
- Help in decision making

Excel Basics

- Entering Data in Excel discussed in this class.
  - Fill – series data
  - Create formula
  - Copy – Icons vs. cross hair vs. Key strokes
  - Locking on "Key Cell" with F4 function key
    - Use of $ in formula
  - Naming an array for repeated future use
- Creating formulas and solve problems
  - Sum function vs. keying in plus symbol
  - Product function vs. keying in multiplication symbols
  - Power function vs. keying in ^ for exponential
  - Count function vs. counting cells and values
- Edit data that is in a worksheet

Go to Worksheet – Basic Excel



Module 2: Describing Data: Graphical Presentations

Good Graphical Presentations

Graphics can aid a presentation's understandability

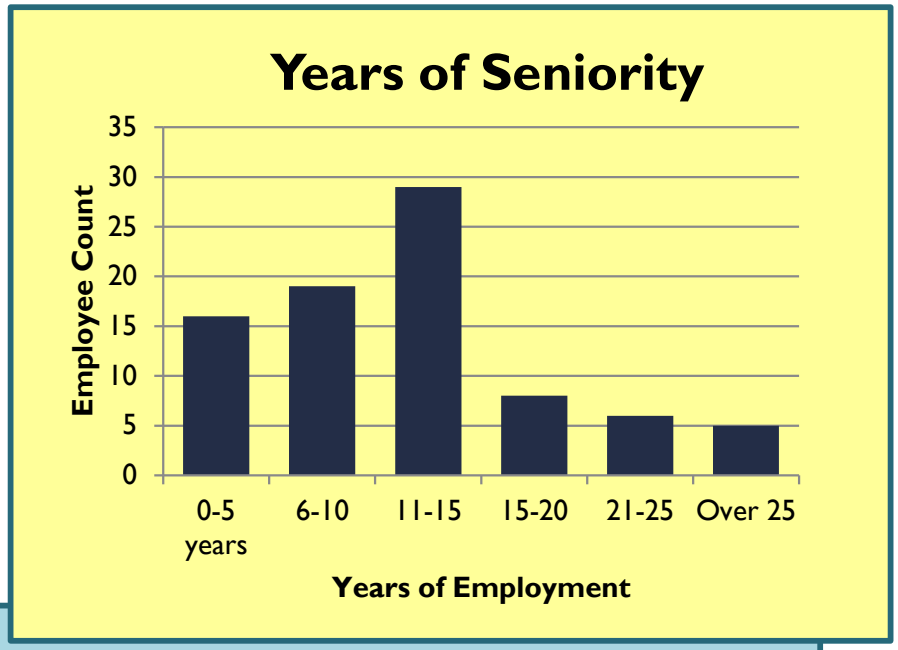
- Use Excel to create common graphic presentations
  - pie chart,
  - bar charts,
  - simple histograms,
  - line charts, and
  - scatter plots.
- Edit and modify charts

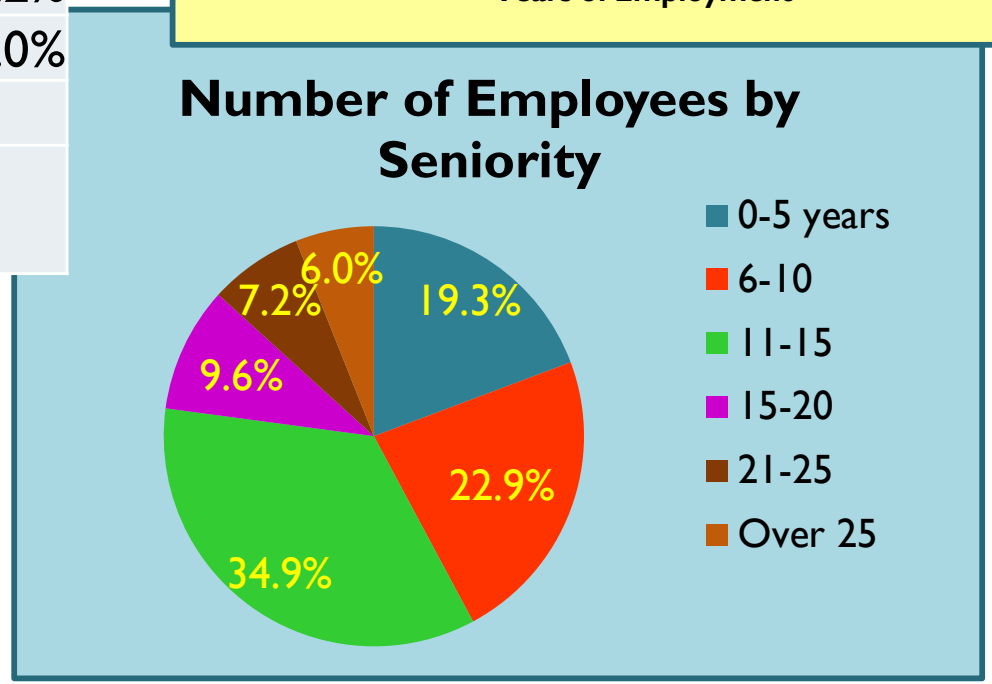
Bar Chart vs. Pie Chart

- **BAR CHART -** A graph in which the classes are reported on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are proportional to the heights of the bars.
  - Microsoft Excel calls this a “Column Chart”
- **PIE CHART -** A chart that shows the proportion, percent or relative frequency that each class represents of the total number of frequencies.

Graphics made easy with Excel

Seniority of Employees	Frequency	Relative Frequency
0-5 years	16	19.3%
6-10	19	22.9%
11-15	29	34.9%
16-20	8	9.6%
21-25	6	7.2%
Over 25	5	6.0%
Total Employees	83	





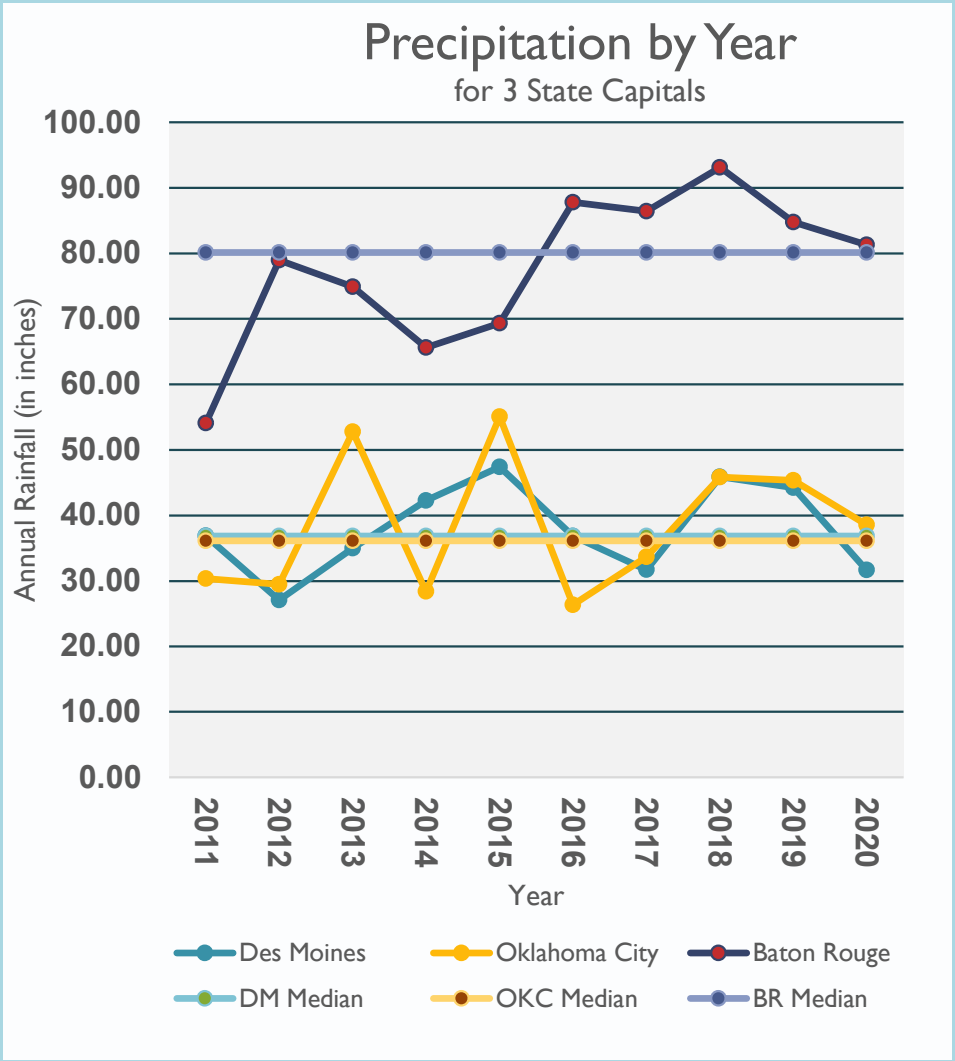
Other Graphics

- Histogram
  - Histogram vs. Bar Charts
- Line Charts – Good for showing changes in data over time.
- Scatter Plots - similar to line graphs
  - Both vertical & horizontal access to plot data points.
  - Show if one variable is related to another – called **correlation** .

Line Charts

Precipitation by Year
for 3 State Capitals

Year	Des Moines	Okla. City	Baton Rouge
2011	36.88	30.37	54.10
2012	27.07	29.49	78.95
2013	34.99	52.79	74.89
2014	42.27	28.38	65.64
2015	47.39	55.07	69.33
2016	36.84	26.32	87.78
2017	31.71	33.67	86.43
2018	45.89	45.83	93.12
2019	44.22	45.36	84.76
2020	31.65	38.57	81.31
Average	37.89	38.59	77.63
Median	36.86	36.12	80.13



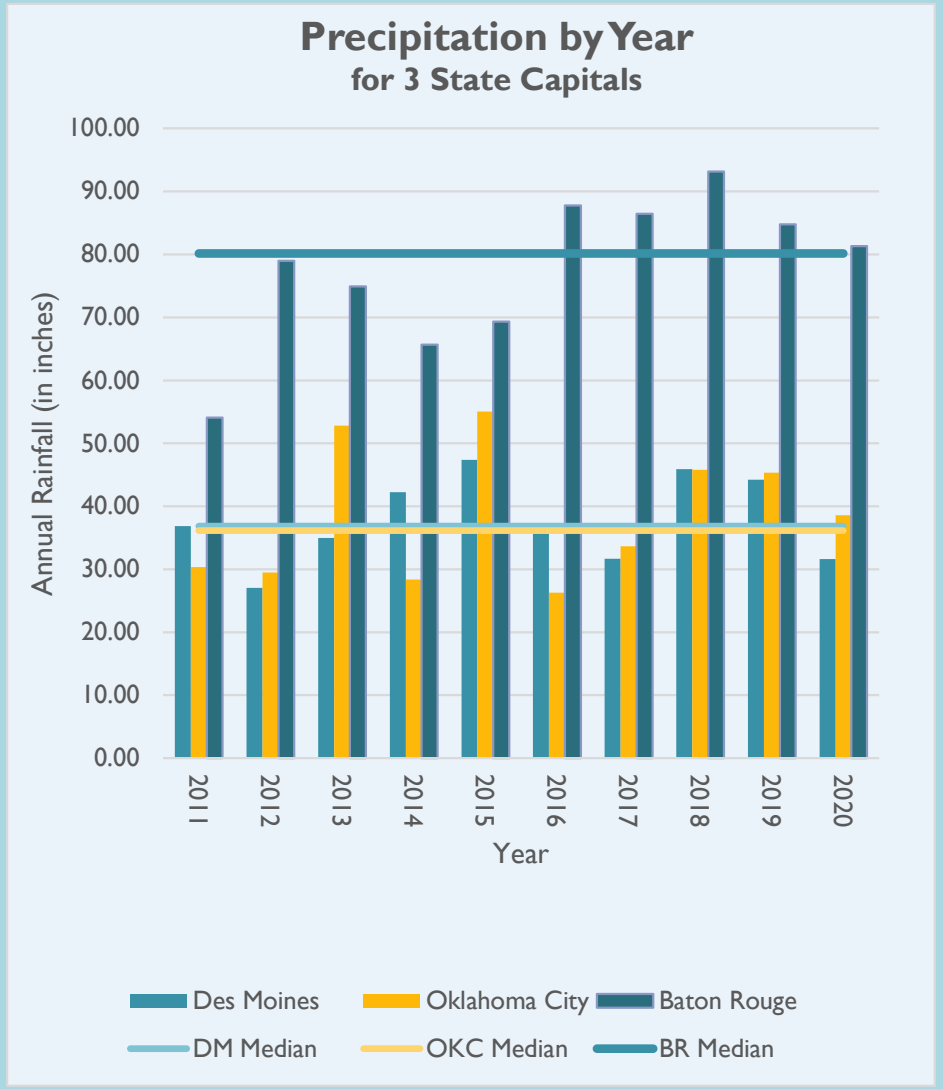
Source: https://www.noaa.gov

See precipitation data

Line Charts and Bar Charts

Precipitation by Year
for 3 State Capitals

Year	Des Moines	Okla. City	Baton Rouge
2011	36.88	30.37	54.10
2012	27.07	29.49	78.95
2013	34.99	52.79	74.89
2014	42.27	28.38	65.64
2015	47.39	55.07	69.33
2016	36.84	26.32	87.78
2017	31.71	33.67	86.43
2018	45.89	45.83	93.12
2019	44.22	45.36	84.76
2020	31.65	38.57	81.31
Average	37.89	38.59	77.63
Median	36.86	36.12	80.13

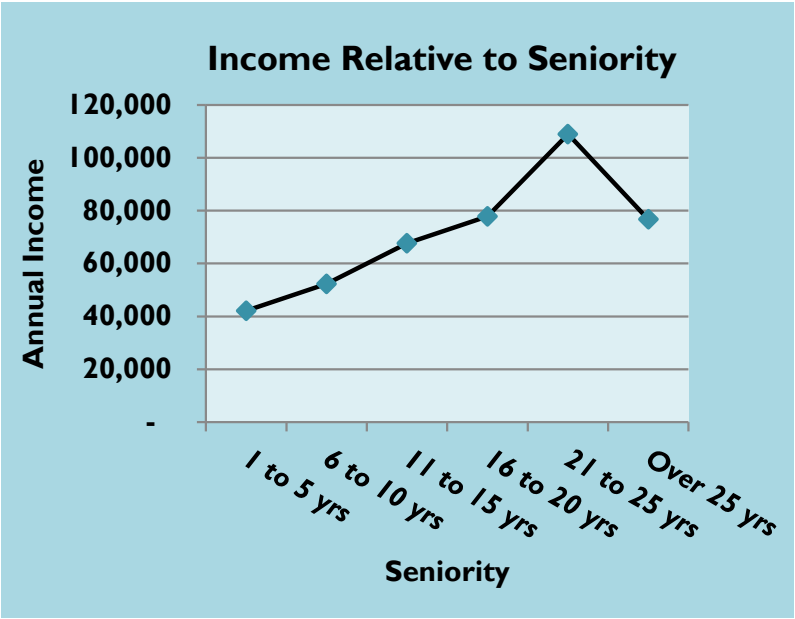


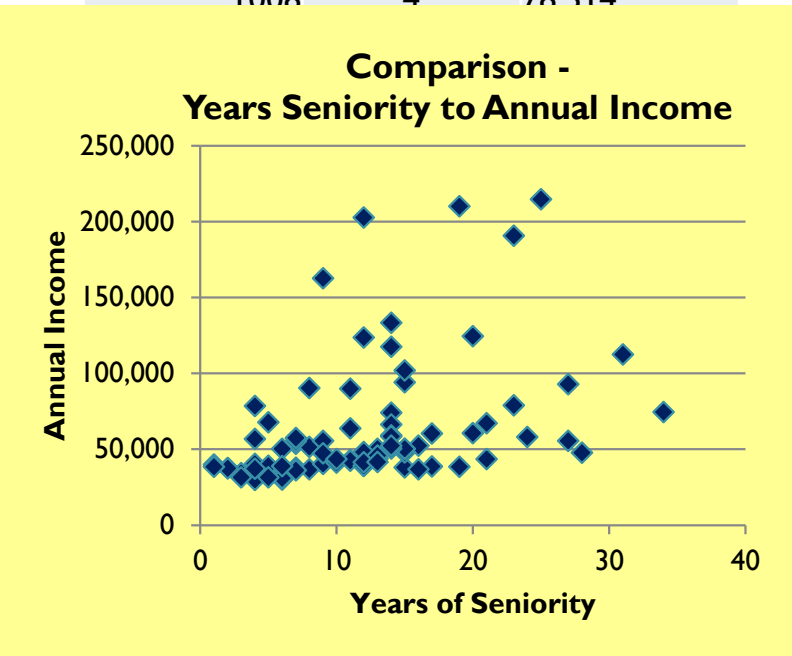
2 Graphics Build a Graph

Graphics Aid to Understanding

Seniority of Employees	Mid Point	Average Income
0-5 years	2.5	42,195
6-10	7.5	52,315
11-15	12.5	67,685
16-20	17.5	77,819
21-25	22.5	108,939
Over 25	30	76,724

Employee #	Seniority of Employee (years)	Income
1001	1	39,940
1002	5	67,949
1003	7	53,696
1004	15	38,003
1005	21	43,586
1006	23	190,876
1007	4	40,434
1008	4	78,514
1019	20	66,697
1020	15	101,961





Go to Worksheet - Graphics

Review of 1& 2

Worksheet - review

Module 3: Describing Data – Measures of Central Tendency

Meaning of Central Tendency

1. Explain the characteristics and uses of measures of central tendency
2. Explain the characteristics and uses of measures of dispersion
3. Use Excel functions to calculate the arithmetic mean, median, mode, and standard deviations
4. Use Excel's Analysis ToolPak add-in to find measures of central tendency and dispersion

Meaning of Central Tendency

- A single value that attempts to describe a set of data by identifying the central position within that set of data.
  - A measure of location
  - Sometimes referred to as point estimate
  - Most common is Average (Mean)
  - Others are Median and the Mode.

Dispersion around

- Measures of Dispersion help to know the spread around the central tendency.
  - Range – give difference between high & low
  - Standard Deviation – how clustered are the values around the central tendency.
    - Indication of likelihood of values relatively near central tendency or relatively far away.
- Knowing mean is wonderful, knowing dispersion gives understanding.

Statistical Function in Excel

- Average (mean) =AVERAGE(numbers)
- Median =MEDIAN(numbers)
- Mode =MODE(numbers)
- Standard Deviation (to include population vs. sample)

2007 and before	2010 and after
=STDEV(numbers)	= STDEV.S(numbers)
=STDEVP(numbers)	=STDEV.P(numbers)

Data Toolpak

- Using the Descriptive Statistic

	Column1
Mean	43670.14012
Standard Error	4006.646402
Median	29834.87583
Mode	#N/A
Standard Deviation	36502.28592
Sample Variance	1332416877
Kurtosis	6.66484503
Skewness	2.623861794
Range	164780.6589
Minimum	18906.44908
Maximum	183687.108
Sum	3624621.63
Count	83

Go to Worksheet – Central Tendency

Module 4: Confidence Interval

What's a Confidence Interval?

- Suppose you conduct a study and arrive at some statistics from your study, e.g. :
  - Sample Mean
  - Sample Standard Deviation
  - Sample Range
- How certain are you that your sample statistics represents the Population's parameters?
- Need to introduce some terminology.

Terminology – Point Estimate

- Point estimate is a single value *sample statistic* used to estimate a *population parameter*.

$\overline{X}$	for	μ
Sample Mean	for	Population Mean
S	for	σ
Sample Standard Deviation	for	Population Standard Deviation

Terminology – Confidence Interval

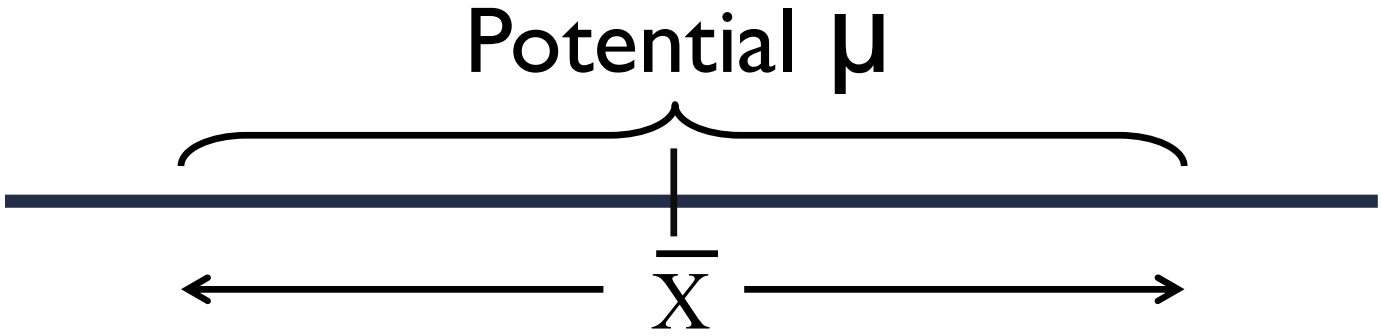
- Confidence Interval is a technique used to indicate the reliability of an sample's statistic as a estimator of the population parameter.
  - Samples do not replicate the population perfectly.
  - A Confidence interval is a range of values (interval) that likely contain the population parameter.

- **Confidence Interval = point estimate ± sample error**

Determinants of Confidence Interval

What determines the width of a confidence interval?

1. The **sample size**, *n*.
2. The **dispersion in the population**, usually σ estimated by *s*.
3. The desired **level of confidence**.



How to Calculate

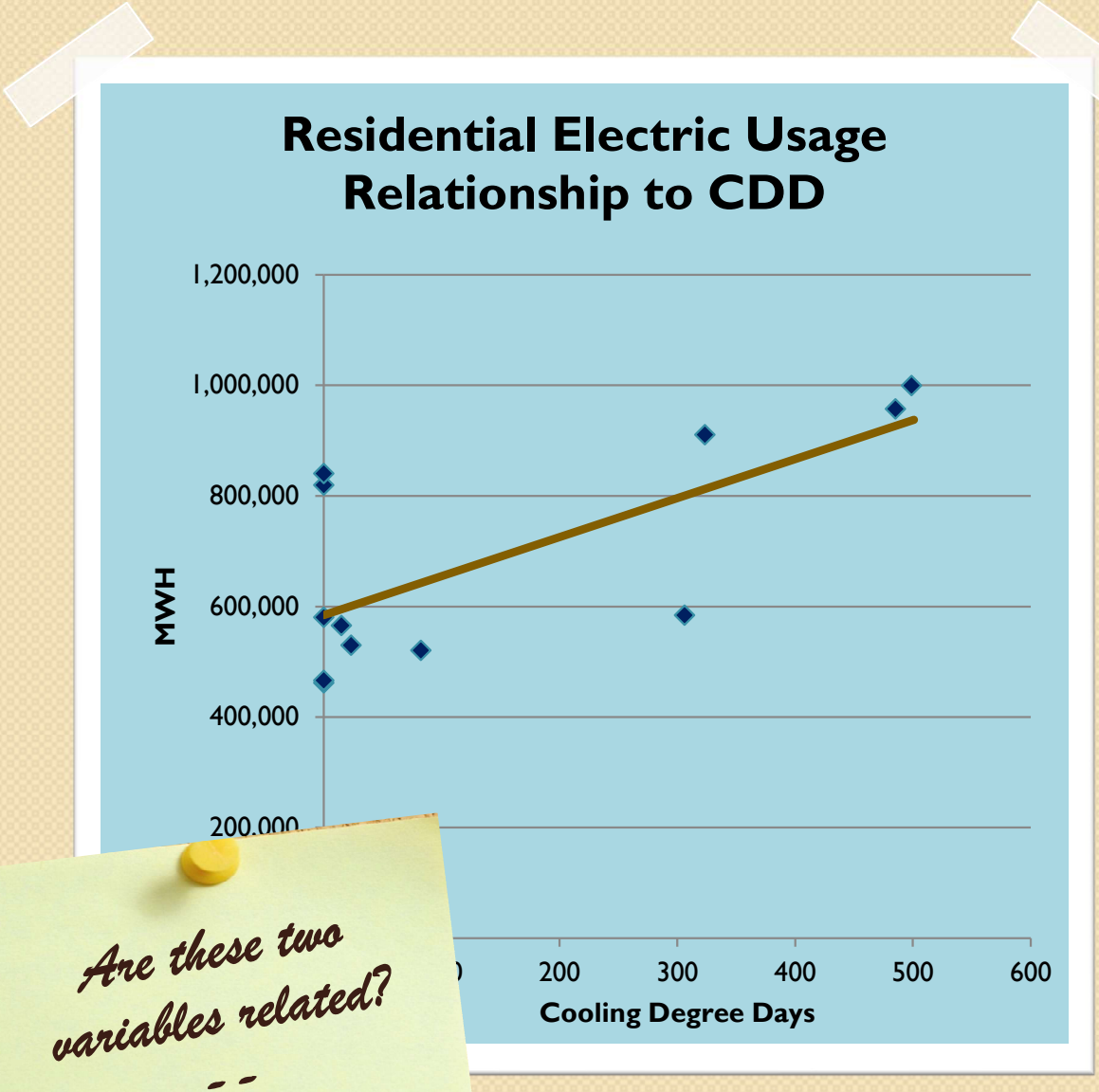
$$\text{C.I.} = \overline{x} \pm t \frac{s}{\sqrt{n}}$$

- C.I. = Confidence Interval
- $\overline{x}$ = sample mean
- t = t statistic
- s = sample standard deviation
- n = sample size

Go to Worksheet – Confidence Interval

Review of 3 & 4

Worksheet - review



Module 5: Linear Regression

Linear Regression

- Purpose is to develop a better understanding the relationship between two variable being studied.
  - Independent variable
  - Dependent variable
- Goal in Linear Regression:
  - Develop an equation to express the relationship b/t variables.
  - Use the equation to estimate the value of the dependent variable using the independent variable as a predictor.

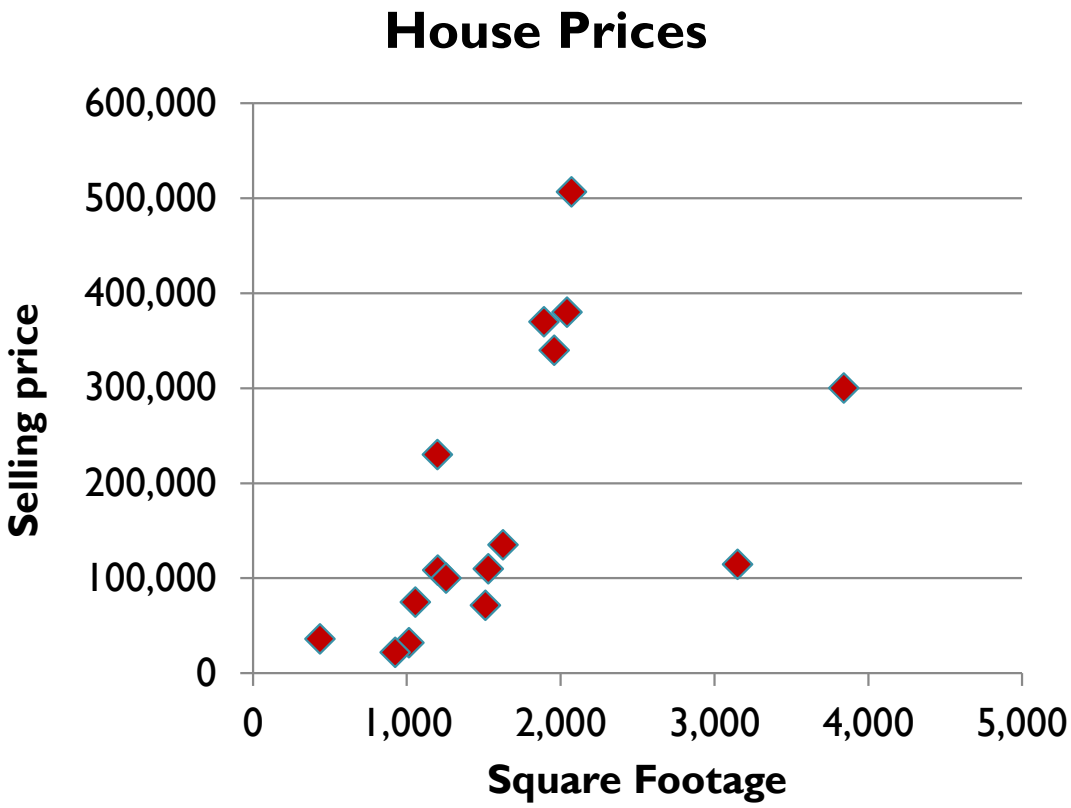
Terminology

- The **dependent variable (criterion)** is the variable being predicted or estimated.
- The **independent variable (predictor)** provides the basis for estimation.
- Examples :
  - Smoking cigarettes increases the likelihood of a person getting lung cancer.
    - Lung cancer is the dependent variable.
    - Smoking is the independent variable
  - Larger house have a higher selling price
    - Selling price is the dependent variable
    - Square footage is the independent variable

Start with Scatter Diagram

Remember back in Module 2

Square Footage	House Prices
3,840	300,000
1,510	71,500
435	36,000
3,150	114,500
1,626	135,000
1,014	32,000
1,529	110,000
1,056	75,000
1,201	108,500
1,892	370,000
2,070	506,666
2,041	380,000
1,959	340,000
1,199	230,000
1,254	100,000
9,24	22,000



Does there appear to be a relationship?

What is a Correlation

- Correlation is numerical measures used to express the strength of the relationship between two variables.
- Many examples of a relationship of correlations:
  - MPG and the car's weight.
  - Quantity of fuel burned and CO_2 emissions
  - Number of red squirrels in Ohio and Stock Market Prices
  - And from previous slide, the square footage of a house and selling price.
- Does correlation mean causation?

Coefficient of Correlation

- The **coefficient of correlation** (*r*) is a measure of the strength of the relationship between two variables.
  - Shows the direction and strength of the linear relationship between two interval or ratio-scale variables.
  - Correlations ranges from –1.00 to +1.00
    - Values of –1.00 or +1.00 indicate perfect or strong
    - Values close to 0.0 indicate weak correlation.
    - Negative values indicate an **inverse** relationship
    - Positive values indicate a **direct** relationship.

See precipitation data

Next build Regression model

The independent variable (*X*) is used to estimate the dependent variable (*Y*).

- A Regression Equation is an equation that expresses the linear relationship between two variables.

$$\hat{Y} = a + b\,X$$

- A math process called least squares technique is used to determine the equation.
  - Minimizes the sum of the squares of the vertical distances between the actual *Y* values and the predicted values of *Y, called Y hat.*

Least Squares approach

What determines good Regression Line?

Rules to follow:

- Logical relationship – relationship between X and Y variables is logical.
- Correlation Coefficient (r) is good
- Coefficient of Determination (r^2) is good
- b slope of the line (coefficient of X) is correct
- t value > 2.2 (rule of thumb)
- Standard error is low

Go to Worksheet – Regression Model

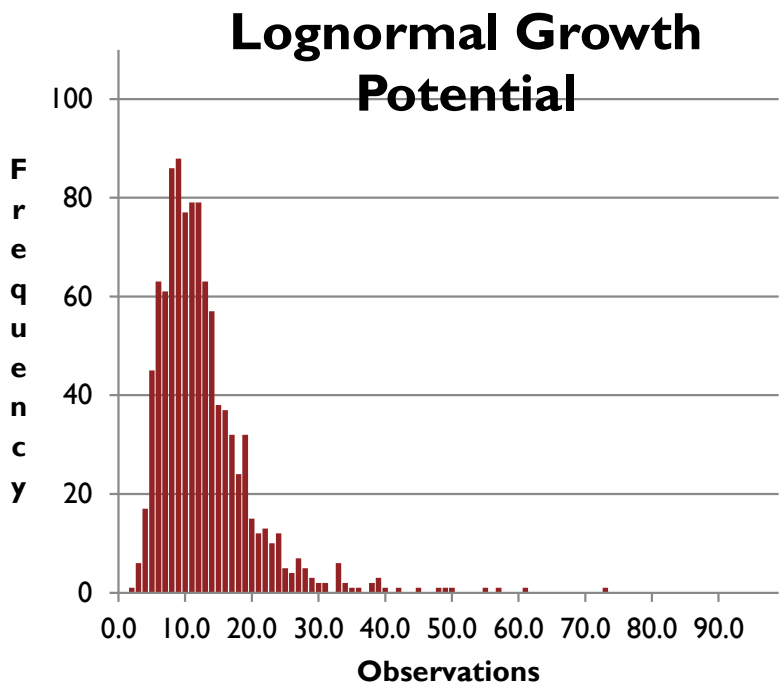
Module 6: Lognormal

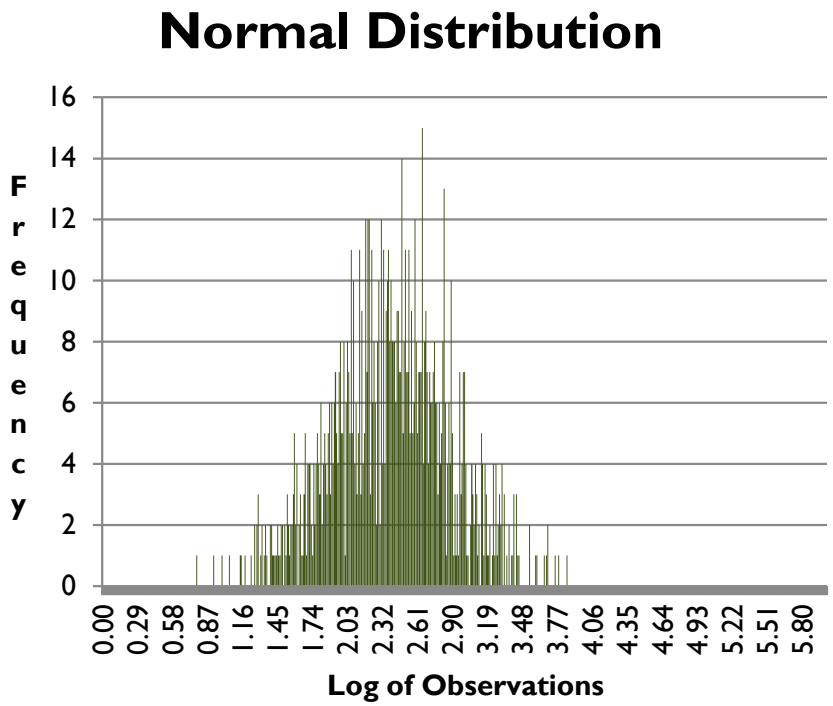
What is Lognormal?

- *Lognormal* – a distribution of a random variable for which the logarithm of the variable has a normal distribution.
- Positively skewed distributions are particularly common when:
  - Mean values are low,
  - Variances large
  - Values are not zero, and
  - Values cannot be negative

What is Lognormal?

- *Log-normal* – a distribution of a random variable for which the logarithm of the variable has a normal distribution.





Examples of Lognormal Distributions

- Skewed distributions often closely fit the log-normal distribution - examples:
  - Lengths of latent periods of infectious diseases,
  - Distribution of mineral resources in the Earth's crust
  - Inheritance of fruit and flower size
  - Return on equities in stock market
  - Survival rates of cancer patients
  - Failure rates in product tests.
  - Rainfall in Las Vegas

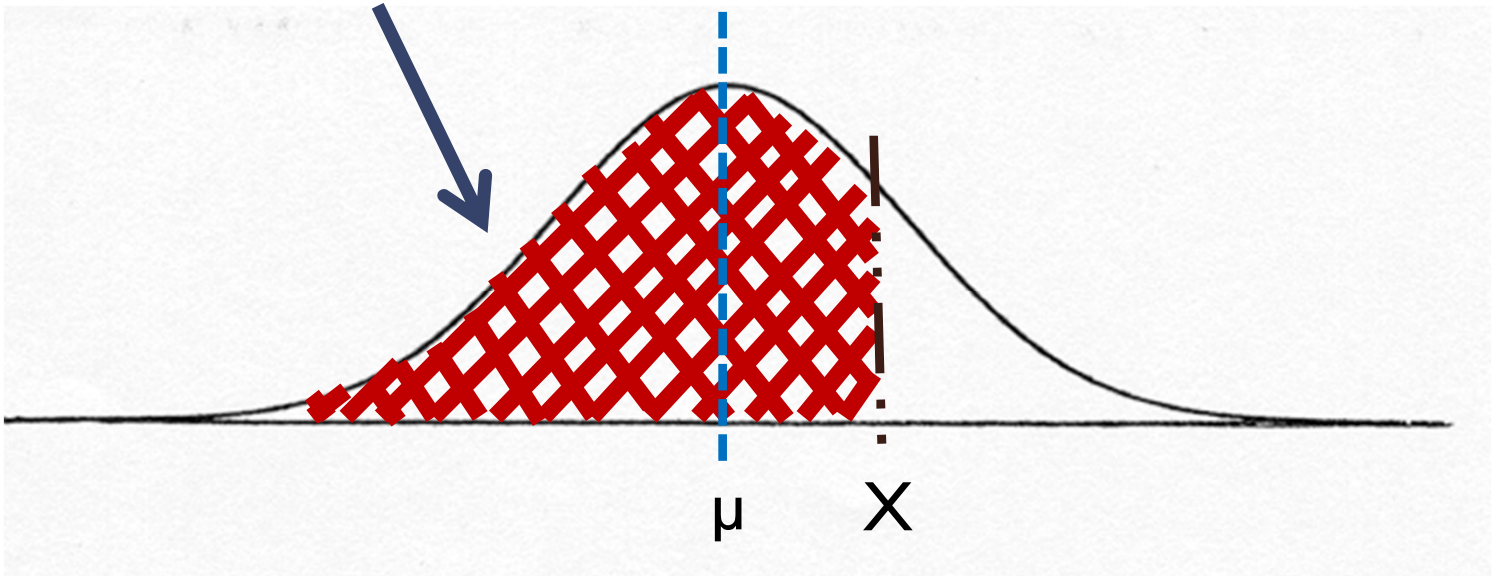
Normal Distribution

Before Excel:
1st Solve for Z
$Z=(X-\mu)/\sigma$
2nd Use a Z table to find the probability

- The Excel function is =NORM.DIST(X,μ,σ,Cumulative)

Where Norm Function gives probability of area < X:

- X = observation
- μ = Mean
- σ = Standard Deviation
- Cumulative is either true or false

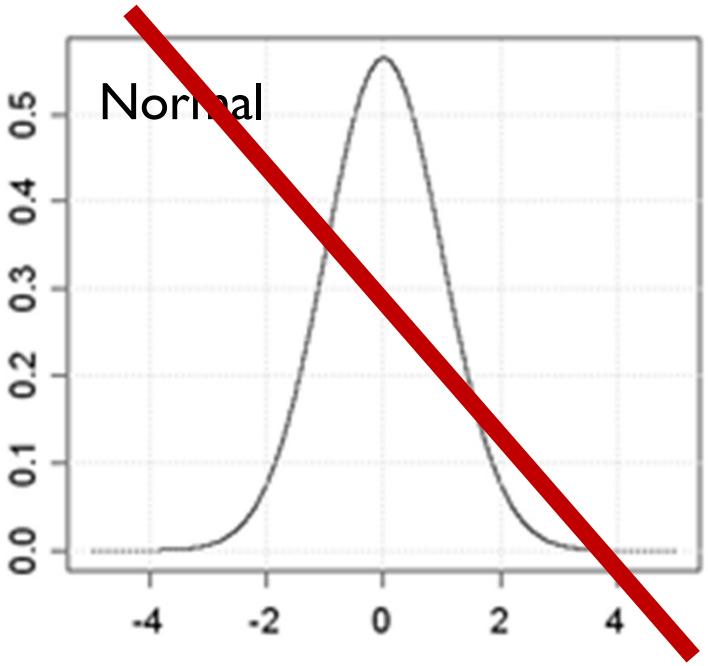


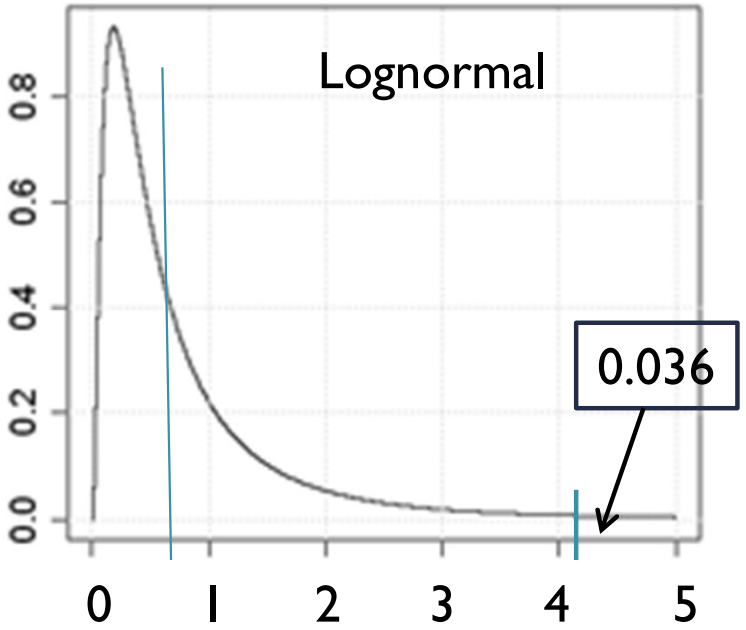
=NORM.S.DIST((LN(X) - Mu)/Sigma,Cumulative)

$Z=(LN(X)-\mu)/\sigma$

Lognormal Distribution

- But we're using function where LN(X) is normal
- Suppose you get monthly data for Rainfall in Las Vegas.
- You see it has the characteristics of Lognormal
- Mean is low and Variance is fairly large





μ = 0.3519 P(X) > 4

σ = 0.572

Go to Worksheet – Lognormal Worksheet

Review of 4-6

Worksheet – review 3

Excel Statistics

240 CenSARA

Instructor: Steve Hiebsch
Email: steve.hiebsch@gmail.com

Open file named, "7 Post-Test Excel …" and complete it

Post-Test 240CenSARA