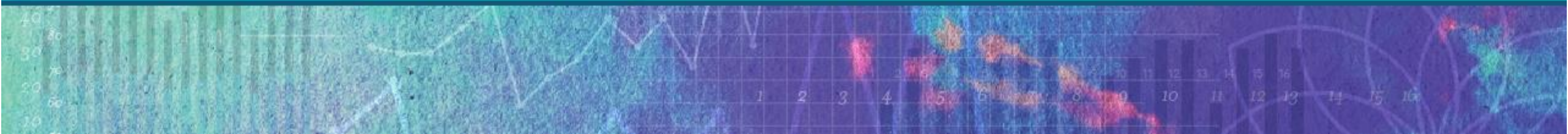
ANALYST BOOT CAMP

ANALYST BOOT CAMP

ALGEBRA & STATISTICS

MARCH 2025

MUIP00071



HOUSEKEEPING

- Actively participate; Do not multitask or do other things during class time.
- Close other applications (including email).
- Be prepared to hear your name.
- Use the “raise your hand button” when you have questions.
- We will use a “parking lot” for questions that are out of scope for this class.

- I will frequently ask “does everyone follow that” because I can’t see your faces.

PRE-TEST

- 2 X 2 + 2 X 2 =
- Rebase the series to the last number: 5, 7, 15, 12, 6
- What are the first 2 terms of the MA3 for the series above:
- The average is a measure of the _____ of a set of data.
- The standard deviation is a measure of the _____ of a set of data.
- In a Normal Distribution (mean=100, SD=15), what Z-scores include 95% of the data values?__
- Explain the Central Limit Theorem? _____
- Given a data set of 15 values should we use the z-score or t-score? __
- The most common level of significance used for Confidence Intervals is: __

COURSE GOALS

Develop an understanding of Algebra,& Statistical Analytical concepts.

Demonstrate that understanding by solving practical problems using Excel.

QUANTITATIVE ANALYSIS BASICS

- Order of Operations, Square Roots, Exponents
- Calculation of Average – Mean, Median, Mode
- Calculation of Standard Deviation
- Weighted Average
- Moving Average
- Growth Rate
- Indexing
- Rebasing

SYMBOLS

Arithmetic Symbols

Multiplication	$*$ $\times$ $\cdot$
Division	/ $\frac{a}{b}$ $\div$
Parentheses	( )
Brackets	[ ] { }
Exponents	x ^ 2, x^2
Subscripts	x_1 , x_2

Logical Symbols

$=$ Equal to

$\neq$ Not Equal to

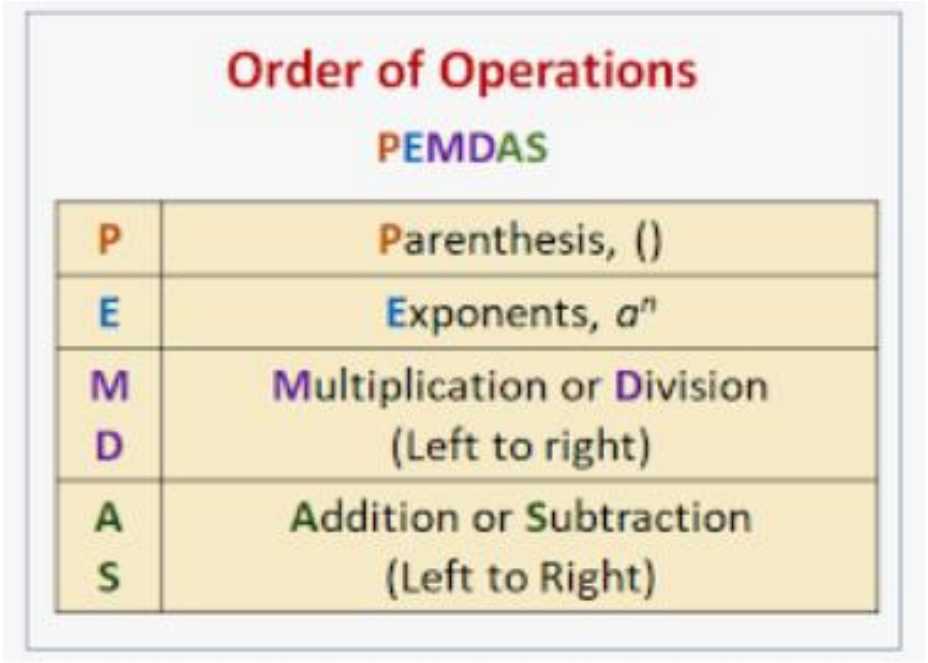
$\leq$ Less than or equal to

$\geq$ Greater than or equal to

$<$ Less than

$>$ Greater than

ORDER OF OPERATIONS (PEMDAS)



1. The Order of Operations matters because we can get different answers to the same problem.

2. Mathematicians agree to use PEMDAS.

3. Multiplication is not ranked before division, as M before D might suggest. They are ranked equal in the 3rd tier. We work on an algebraic expression from left-to-right within the tier.

4. The same rules hold for addition vs. subtraction in the 4th tier.

THE ORDER OF OPERATIONS MATTERS

$$2 \times 2 + 2 \times 2$$

Sequential:

2 x 2 + 2 x 2
4 + 2 x 2
6 x 2
12

PEMDAS:

2 x 2 + 2 x 2
4 + 2 x 2
4 + 4
8

If you have a calculator, which order does it use?

What answer do you get from the Windows calculator on your laptop?

This one is a hot topic on the internet. Google it : $8 \div 2(2 + 2)$

SQUARE ROOTS & EXPONENTS

Square Root (SQRT) – What number, multiplied by itself, is the number under the square root symbol?

1. # under symbol is > 1: $\sqrt{64} = \sqrt{8 \times 8} = 8$ $\sqrt{1.44} = \sqrt{1.2 \times 1.2} = 1.2$ Answer < #

2. # under symbol is between 0 and 1: $\sqrt{0.81} = \sqrt{0.9 \times 0.9} = 0.9.$ Answer > #

3. # under symbol is < 0: $\sqrt{-1}$ = ? Try this in excel. No answer.

Exponents – The count of how many times a number is multiplied by itself

$10^2 = 10 \times 10 = 100$ $2^4 = 2 \times 2 \times 2 \times 2 = 16$

CALCULATION OF AVERAGES

The **average** is a measure of **"central tendency"** of the data.

Example: [2, 2, 2, 4, 10]

There are 3 common types of averages, calculated as:

- The **MEAN**: Add up the data points and then divide by the number of data points.
  2+2+2+4+10 = 20 20/5 = 4 Mean = 4
- The **MEDIAN**: Arrange the numbers in order, then pick the middle number. Median = 2
- The **MODE**: The number that appears most often. Mode = 2. If there are no repeats, the mode does not exist.

CALCULATION OF STANDARD DEVIATION (SD)

Standard Deviation (SD) is a measure of how **"spread out"** the data is.
The calculation uses PEMDAS, Square Roots, and Exponents.

Ex 1. **(2, 2, 2, 4, 10)** has a mean of 4. SD is a measure of the **deviations** from 4.

SD = SQRT (((2-4)^2 + (2-4)^2 + (2-4)^2 + (4-4)^2 + (10-4)^2) / 5)
= SQRT ((2^2 + 2^2 + 2^2 + 0 + 6^2) / 5)
= SQRT ((4 + 4 + 4 + 0 + 36) / 5)
= SQRT (48/5) = SQRT (9.6) = 3.09

Ex 2. To compare, **(2, 3, 4, 5, 6)** also has a mean of 4, but the SD is:

SD = SQRT (((2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2) / 5)
= SQRT ((2^2 + 1^2 + 0^2 + 1^2 + 2^2) / 5)
= SQRT ((4 + 1 + 0 + 1 + 4)) / 5)
= SQRT (10/5) = SQRT (2) = 1.41

Both sets have the same mean = 4, but the first is more "spread out" and has a bigger SD.

EXCEL EXERCISES

Go to the Excel exercises in your workbook. The answers are to the far right on each tab.

Tabs

- OrdOps
- SqR_Exp
- Avg
- SD

WEIGHTED AVERAGE – EASY Example

Weighted Average gives more weight to the more important segments of data.

Unweighted GPA		
Course	**Credits**	**Grade (A=4,B=3, C=2,D=1)**
Chemistry	5	2
Math	4	3
Writing	4	3
Volleyball	1	4
Sum		**12**

Unweighted GPA = Sum(grades) / (# of Grades)

Unweighted GPA = 12/4 = 3.00

Same as all courses being 1 credit

Weighted GPA			
Course	**Credits**	**Grade (A=4,B=3, C=2,D=1)**	**Credits x Grade**
Chemistry	5	2	10
Math	4	3	12
Writing	4	3	12
Volleyball	1	4	4
Sum	**14**		**38**

Weighted GPA = Sum(credits x grade) / Sum (credits)

Weighted GPA = 38/14 = 2.71

Same as (2+2+2+2+2 + 3+3+3+3
(+3+3+3+3 + 4) / 14 = 38/14

WEIGHTED AVERAGE in GAINSHARE

Weighted Average gives more weight to the more important segments of data.

PL Direct has more importance than H/O because there are more Premiums.

PGR targets CR < 0.96 or Profit Margin > 0.04

Unweighted Gainshare (Hypothetical Data)		
Segment	**Premiums ($MM)**	**Combined Ratio (CR)**
PL - Direct	10	0.91
PL - Agency	8	0.98
CL	4	0.94
Homeowners	2	1.11
Sum		**3.94**

Unweighted GS = Sum(CR) / (# of CRs)

Unweighted GS = 3.94/4 = 0.99

Weighted Gainshare (Hypothetical Data)			
Segment	**Premiums ($MM)**	**Combined Ratio (CR)**	**Premiums x CR**
PL - Direct	10	0.91	9.10
PL - Agency	8	0.98	7.84
CL	4	0.94	3.76
Homeowners	2	1.11	2.22
Sum	**24**		**22.92**

Weighted GS = Sum(Premiums x CR) / (# of CRs)

Weighted GS = 22.92/24 = 0.96

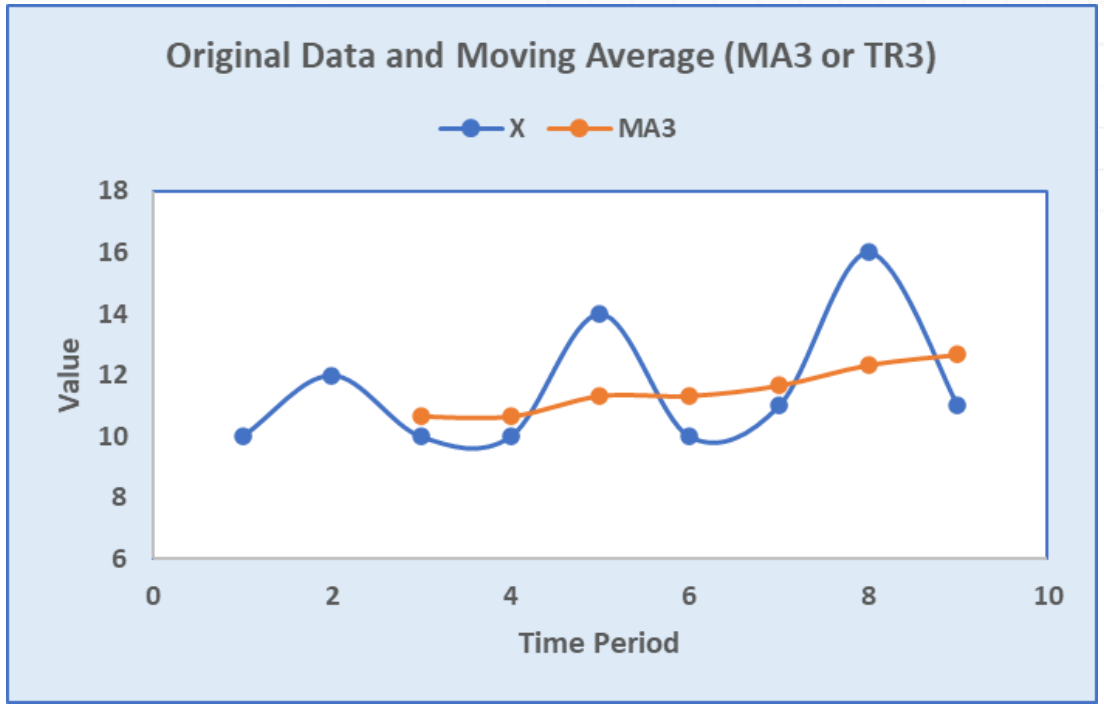
Often people want to use the "average of averages" because it sounds good. But they are assuming all segments have the same weight and that is generally not a valid assumption.

MOVING AVERAGE
MAn or TRn where n = # of data points

The two purposes of a Moving Average are to

- smooth out fluctuations in the data and
- remove seasonality (usually MA12 or TR3) to show trend

Clusters of 3. Second term high.			
Time Pd	X	MA3	
1	10		
2	12		
3	10	10.7	average of time 1-3
4	10	10.7	average of time 2-4
5	14	11.3	average of time 3-5
6	10	11.3	
7	11	11.7	
8	16	12.3	
9	11	12.7	



GROWTH RATE

Growth Rate is the percent change of a quantity **from start to end of a time period**. It can be positive or negative. There are 2 ways to calculate it. You should be familiar with both.

PIFs increase from **2000** in Year 1 to **2600** in Year 2. What is the Annual Growth Rate?

Method 1: (End-Start)/Start **(2600 - 2000) / 2000** = 0.30 x 100% = 30%

Method 2: (End/Start) - 1 **(2600/2000) -1** = 1.30 -1 = 0.30 x 100% = 30%

PIFs decrease from **2000** in Year 1 to **1600** in Year 2. What is the Annual Growth Rate?

Method 1: (End-Start)/Start **(1600 - 2000) / 2000** = -0.20 x 100% = -20%

Method 2: (End/Start) - 1 **(1600/2000) -1** = 0.80 -1 = -0.20 x 100% = -20%

GROWTH RATE

Caveat: It really doesn't make sense to calculate the growth rate if the starting and ending values are of opposite signs (one is positive and one is negative), but people still do it.

GEICO has a Loss of 2600 in Year 1, followed by a Profit of 2000 in Year 2.
What is the Annual Growth Rate?

Method 1: (End-Start)/Start **[2000 - (-2600) / (-2600)]** = 4600/(-2600) = (-1.77) x 100% = **(-177%)**

Paradox: Their financial picture improved, but the (-177%) is associated with deterioration, not improvement

Guideline: To avoid the paradox, both starting and ending values should be positive. If not, indicate N/A

INDEXING

Divide all the numbers in a series by the **first number**. This sets the first number equal to 1.00
Then multiply the newly created numbers by 100. This creates an **Indexed Series.**

Example: Economic series (CPI, PPI)

Q1	Q2	Q3	Q4
87	**95**	**108**	**120**
1.00	1.09	1.24	1.38
100	109	124	138

Why do we create an index? Because it is easy to interpret:

Q2 is 9% higher than Q1
Q4 is 38% higher than Q1

REBASING AND RELATIVITIES

Rebase: Divide the base (reference number) by itself. This sets the Base = 1.00
Relativities: The ratios of other numbers to the base.

Many reports at PGR use “Loss Ratio Relativities” or “Pure Premium Relativities”. Here is an example where we re-base to the Total Combined Ratio (CR) and calculate relativities.
Combined Ratio is (1 – Profit Margin) and is a metric used in insurance.

Example

Segments	Profit Margin	Combined Ratio (CR) = (1 - PM)	Math	Relative CR Relative to 0.96
SAM	-5%	1.05	1.05/0.96	1.09
DIANE	1%	0.99	0.99/0.96	1.03
WRIGHTS	10%	0.90	0.90/0.96	0.94
ROBINSONS	14%	0.86	0.86/0.96	0.90
TOTAL	**4%**	**0.96**	**0.96/0.96**	**1.00**

Indexing is basically re-basing to the first term of the series of numbers.

EXCEL EXERCISES

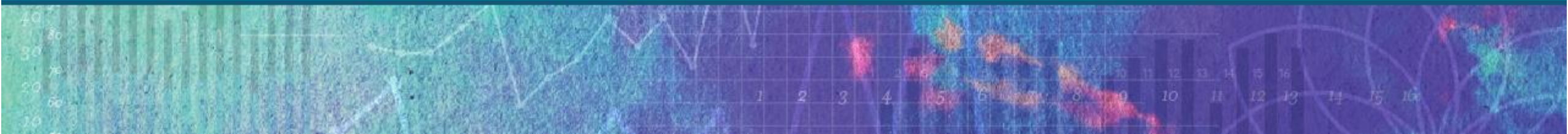
Go to the Excel exercises. Answers are to the far right on each tab.

Tabs

- WtdAvg
- MovAvg
- Growth Rate
- Index
- Rebase

ANALYST
BOOT CAMP

STATISTICS



STATISTICS BASICS

- The fundamental relationship of a Sample to a Population
- Descriptive Statistics
  - Center and spread of the data
  - Outliers
  - Visualizations

- Inferential Statistics: infer population parameters from sample parameters
  - Correlation & Causation
  - Linear Regression
  - Types of Distributions
  - Normal Distribution
  - Z-scores & Standard Normal Distribution
  - Central Limit Theorem

KEY CONCEPT: POPULATION VS.SAMPLE

Population: An entire group of people or objects of interest
Sample: A subset of the population. Hopefully representative of the population.

Samples are used because we can't get the entire population, or it is too expensive to get.

Examples:

- Population: All the data analysts in the world. Sample: all data analysts in Ohio
- Population: All Major League Baseball players. Sample: Cleveland Guardians players

  - Think of some examples yourself and tell us about them

ARE THESE POPULATIONS OR SAMPLES?

1. A teacher gives an exam to their pupils. The teacher wants to summarize the results of the exam.

   Population. The teacher is only interested in these specific pupils' scores and nobody else.

2. A researcher has recruited males aged 45 to 65 old for an exercise study to investigate risk markers for heart disease (e.g., cholesterol).

   Sample. A researcher investigating health related issues will not simply be concerned with only the participants of their study; they will want to show how their sample results can be generalized to the whole population (in this case, all males aged 45 to 65 years old).

3. A census is taken every 10 years in the US to gather information on demographics, etc.

A census is close to a population but misses some people because it is difficult to find them.

KEY CONCEPT: SAMPLE BIAS

Sample Bias is a systematic misrepresentation of the population data and is to be avoided.

Examples:

- Using temperatures in June to represent the entire year.
- 1948 Presidential election prediction based on a phone survey when many people did not have a phone
- Statistical theory tells us survey answers are plus/minus 3% if there are 1100 respondents
  - Then why are election surveys so bad at predicting the winner? Answer: the answers to the surveys are from a biased set of respondents.
- Think of some examples and share them.

EXTREME BIAS – SURVIVORSHIP BIAS

A famous case of "survivorship bias" happened during WWII. American bombers were suffering significant losses during missions over Germany.

The Air Force was deciding where to put more protective armor on the planes. They studied the damaged planes and found that the fuselage (body) had the most bullet holes. So, they decided that was the area that needed to be reinforced.

But a statistician, Abraham Wald, reasoned differently. He thought the sample of planes with fuselage holes were the ones that returned. The ones that didn't return likely had bullet holes in a different place – the engine and wings.

The armor was installed over the engines. That increased the % of bombers that successfully returned.

DESCRIPTIVE STATISTICS

Techniques for summarizing a set of data

in ways that people can easily interpret.

TYPES OF DATA

- **Numerical** variables
  - Example 1: The number of miles on the odometer of a car.
  - Example 2: The age of an insured customer.
- **Categorical** variables are non-numeric.
  - Example 1: The make of a car (Chevy, Tesla, Ford, etc.)
  - Example 2: The age in groupings: (0-19, 20-39, 40-59, 60-79, 80+)
  - Example 3: Survey Scores (5/4/3/2/1). These are often treated as numeric.

MEAN, MEDIAN, AND MODE IN DEPTH

Numeric values use **Mean** and **Median** as measures of **"central tendency".**

For this data, the median is the best measure of the "central tendency" because there are a few atypically large project times that greatly influence the mean.

IT Project Time	
	50
	80
	90
	100
	130
	160
	180
	190
	400
	700
Average	208
Median	145

However, if the goal is to measure the **improvement** in project time and the improvement is due to the large values being reduced, the median may not pick up the improvement. This is particularly true for small to medium sized data sets.

IT Project Time	
	50
	80
	90
	100
	130
	160
	180
	190
	400
	700
Average	208
Median	145

IT Project Time	
	50
	80
	90
	100
	130
	160
	180
	190
	200
	350
Average	153
Median	145

Cut these 2 times in half.

Improvement
-55
0

ANALYST BOOT CAMP

MEAN, MEDIAN, AND MODE IN DEPTH

Categorical variables use the **mode** to describe "**central tendency**".

The mode of the car companies is Audi. The mode of the survey is 2.

Brand
Audi
Audi
Audi
Audi
Audi
Audi
BMW
BMW
Porsche
VW
VW
VW
Mercedes
Mercedes

GuyGo Insurance Co. Survey	
1=Satisfied 2=Neutral 3=Not satisfied	
Indicate your score for:	**Score**
The company	2
Your Business Leader	2
Your Manager	1
Our mascot	3

Strictly speaking, survey answers are categorical, but people often treat them as numbers and compute average scores.

MEAN, MEDIAN, AND MODE IN DEPTH – KHAN

Khan Academy

Statistics Intro: Mean, median, & mode

To access Khan Academy Lessons:
Control-Click the link to access the material everywhere in the presentation.

STANDARD DEVIATION IN DEPTH

The average is a measure of the “central tendency” of a set of data.

The **standard deviation** is a measure of how **compact** or **spread-out** the data is.

There are 2 Excel calculations, depending on if the data is a population or sample

(Population) Standard Deviation = **stdev.p**

Sample Standard Deviation = **stdev.s**

For theoretical reasons**, stdev.s** is generally a little larger than **stdev.p.**

STDEV ( ) uses **stdev.s( )** to be conservative**.**

STANDARD DEVIATION IN DEPTH – KHAN

Khan Academy

Standard Deviation

Sample Standard Deviation

DISPLAYING DATA

Khan Academy

Displaying & Describing

ANALYST
BOOT CAMP

OUTLIERS

Outliers are data values that are significantly different from the other values

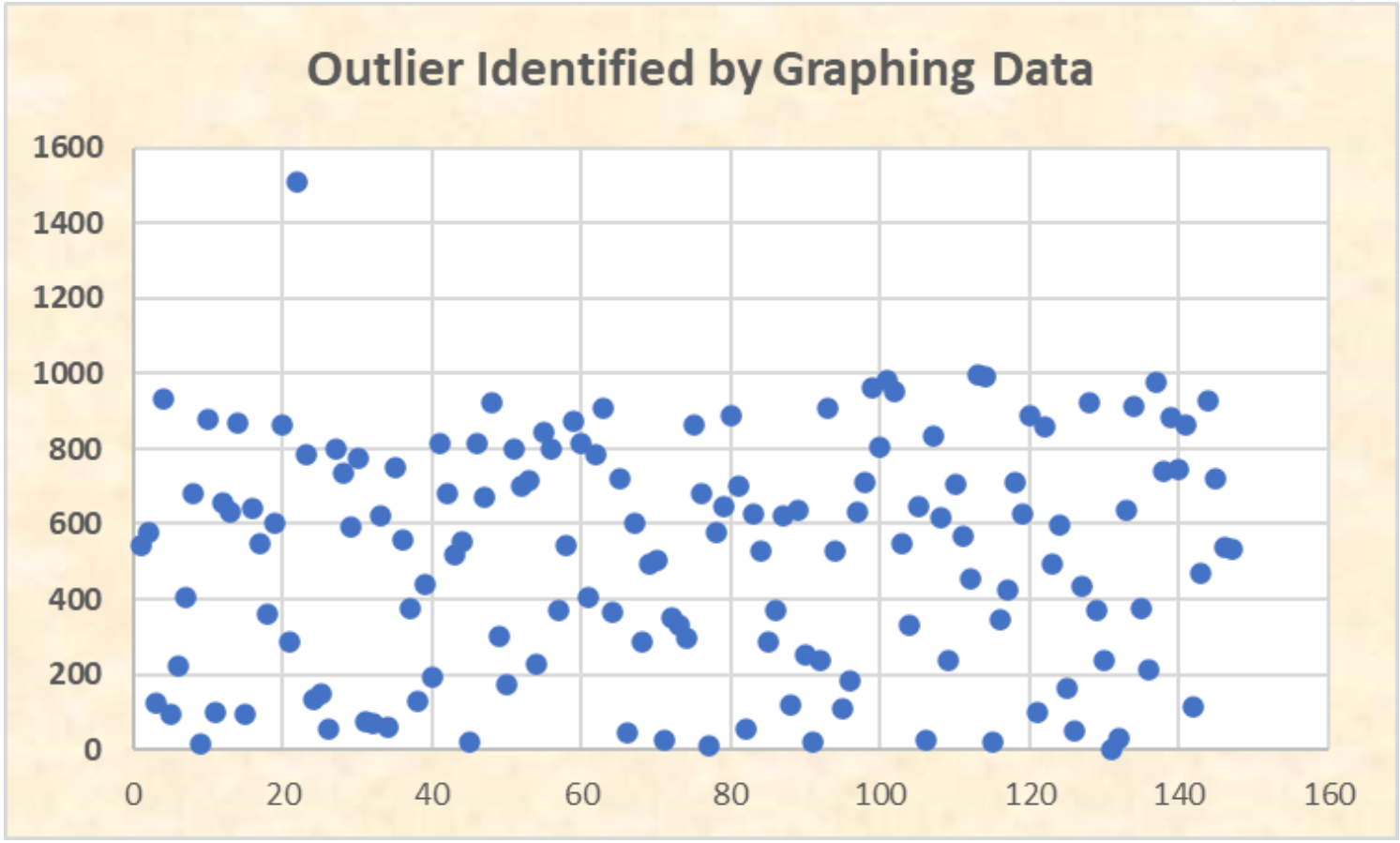
Example: 0, 0, 1, 1, 2,100, 3, 2

- Outliers can dramatically affect the calculations of $\bar{x}$ and s.
  with the outlier, $\bar{x}$ = (0+ 0 + 1 + 1 + 2 + 100 + 3 + 2) / 8 = 13.6
  without the outlier, $\bar{x}$ = (0+ 0 + 1 + 1 + 2 + 3 + 2) / 7 = 1.3

- You can only remove outliers if you have a valid reason. Example from Claims:
  - 100 could represent the real # of cars involved in a pile up (keep it)
  - 100 could be a “fat fingered” data error (change or remove it).

- PGR Pricing Indications: we remove Large Losses initially, then redistribute them.
- In practice, it is difficult to identify outliers in big datasets. Graphing might help..

OUTLIERS



ANALYST
BOOT CAMP

OUTLIERS– KHAN

Khan Academy

Outliers

ANALYST
BOOT CAMP

INFERENTIAL STATISTICS

This is when the fun begins.

Estimate **population** parameters by **sample** parameters.

- Parameters of a Population (Greek letters)
  - **μ** Mu Population Mean
  - **σ** Sigma Population Standard Deviation
  - **N** Number of observations in the Population
  - y = **α**x + **β** Slope and y-intercept of the Population

- Parameters of a Sample
  - **$\bar{x}$** x-bar Sample Mean
  - **s** Sample Standard Deviation
  - **n** Number of observations in the Sample
  - y = **m**x + **b** Slope and y-intercept of the Sample

CORRELATION

Two variables are **correlated** if there is a relationship or connection between them.

The statistical notion of correlation is the **correlation coefficient.**

- The population correlation coefficient is Greek letter **rho ρ**.
- The sample correlation coefficient is **r**.

- r is calculated in excel as r = correl(range1,range2). It is between -1 and +1.
  - r tells us the sign (direction) of the relationship but not how strong the relationship is.
  - r-squared tells us how strong the relationship is.
    - It is the % of the variation in y is explained by the variation in x.

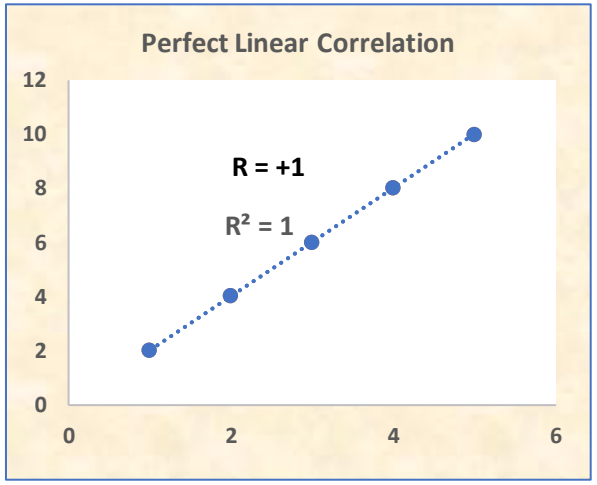
See the next slide for examples

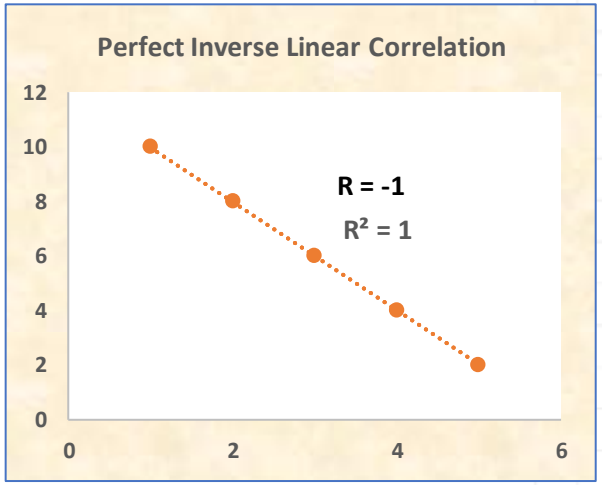
CORRELATION SCENARIOS

r = +1 r-square =1.0

Exact positive relationship between x (horizontal axis) and y (vertical axis).

All the variation in y is due to the variation in x.





r = -1 r-square = 1.0

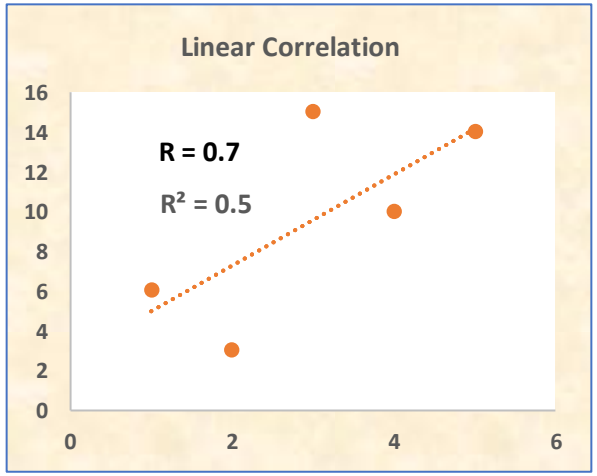
Exact inverse relationship between x (horizontal axis) and y (vertical axis).

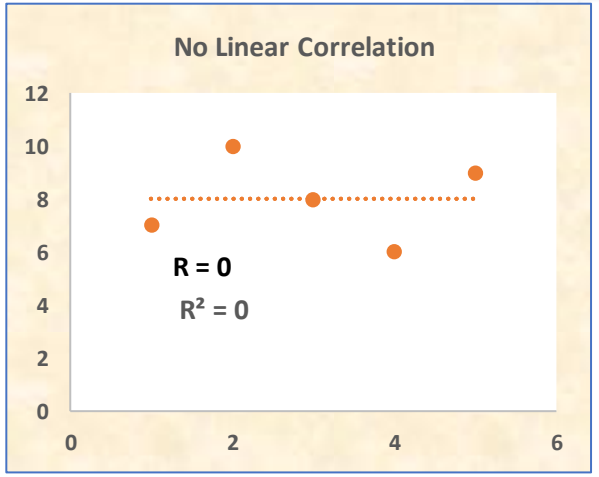
All the variation in y is due to the variation in x.

r = 0.7 r-square = 0.5

Positive relationship between x (horizontal axis) and y (vertical axis).

50% of the variation in y is due to the variation in x.



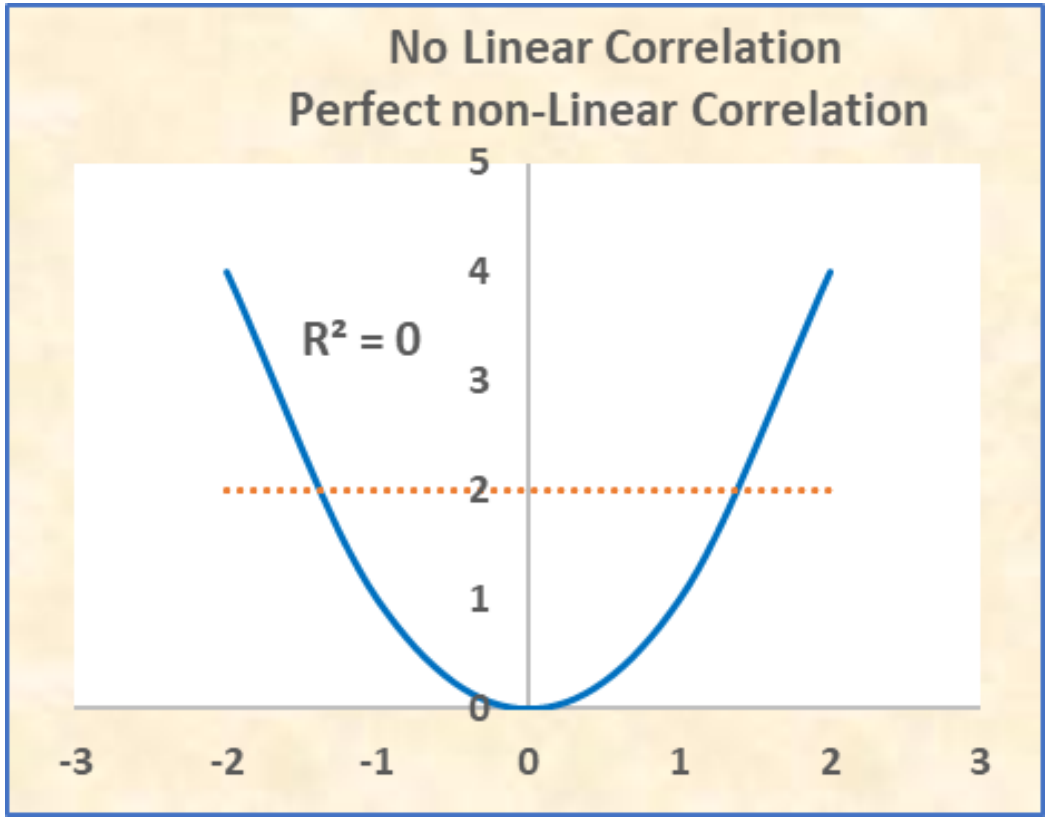


R = 0.0 r-square =0.0

No linear relationship

ANALYST BOOT CAMP

NONLINEAR CORRELATION SCENARIOS



Just because r=0 & r-square = 0 does not mean there is no relationship between x and y.

There is **no Linear relationship** because the orange line has slope, r, and r-square =0.

But there is a **perfect non-Linear relationship** on the parabola.

CORRELATION VS. CAUSATION

There may be a **correlation** between two variables, but **that does not** necessarily mean that one **causes the other**.

Example 1: High correlation, no causation:

Ice Cream sales and baseball homeruns.

Example 2: High correlation with causation.

Weight gain in a rat and amount of food eaten.

PGR charges lower premiums to drivers with high credit scores. This is correlation, but the argument is that people who are responsible with financial decisions are generally more responsible drivers.

CORRELATION VS. CAUSATION – KHAN

Khan Academy

Correlation

Correlation & Causation

EXCEL EXERCISES

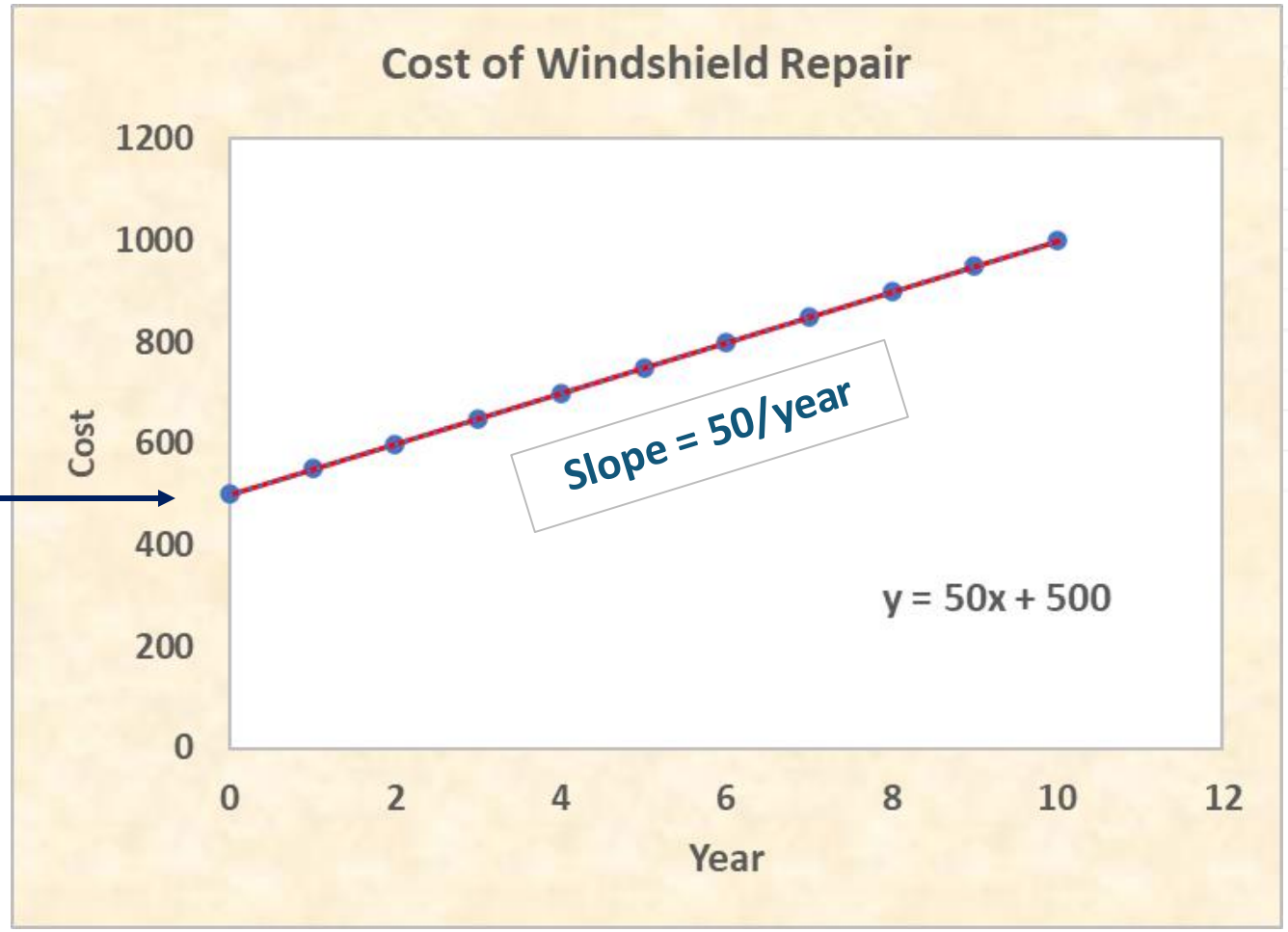
Go to the Excel exercises in your workbook. The answers are to the far right on each tab.

Tabs

- MeanMed
- Display
- Outlr1
- Outlr2
- Corr1
- Corr2

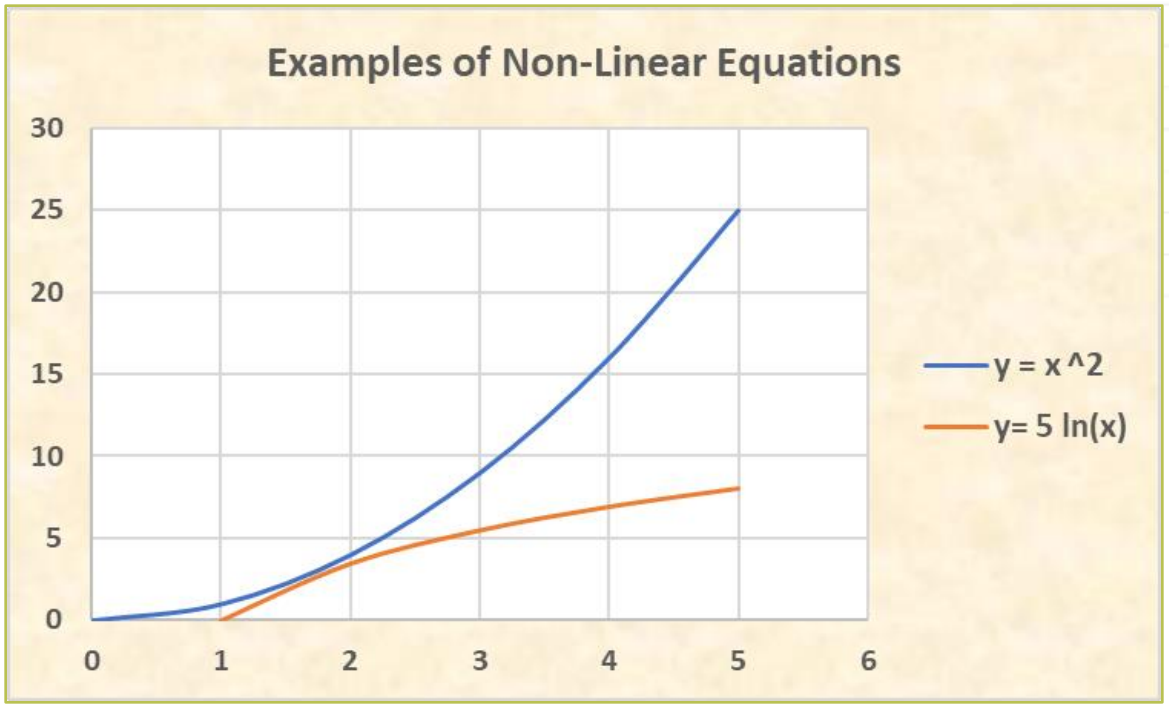
A LINEAR EQUATION IS A STRAIGHT LINE

y intercept = 500



NONLINEAR EQUATIONS ARE NOT STRAIGHT LINES

Examples: y = x^2 y = 5*ln(x)



LINEAR EQUATIONS (TRENDS) CALCULATIONS

A **linear equation** has the form **y = mx + b**.

b is the **y-intercept**: where the line crosses the *y* axis when *x* is 0.
m is the **slope**: the amount y increases when x increases 1 unit.

Method 1 to find the equation y = mx + b: Slope/intercept method

Today's cost to repair a windshield is $500. The cost increases $50 each year.
How much will a windshield cost to repair 6 years from now?

When x=0, y=$500, so the **y-intercept is b** = $500.
The cost increases $50 each time x increases by 1, so **m = $50**.

y = $50x + $500

In year 6: y = ($50)(6) + $500 = $300 + $500 = $800

LINEAR EQUATIONS (TRENDS) CALCULATIONS

Method 2 to find the equation y = mx + b: Two Point Method

Given: (2, 600) and (6, 800) are two points on a line. Find the equation of the line.

x=2, y=600 and x=6, y=800

1. Compute the slope: (change in y) / (change in x) = (800-600)/(6-2) = 200/4 = 50

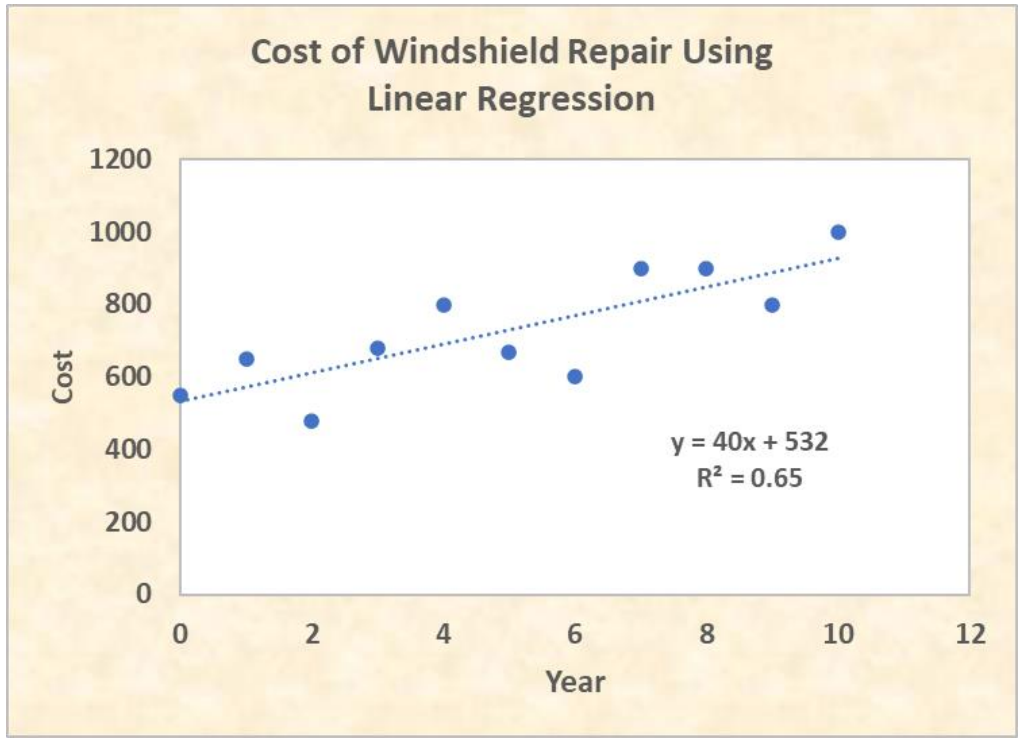
So, **y = mx + b** **y = 50x + b**

2. Put either point into the equation to find b: 600 = 50(2) + b 600 = 100 +b b = 500

y = 50x + 500

LINEAR REGRESSION

In Algebra, if all the points are **on** a straight line, the equation is **y = mx+ b**



In Statistics, the data points are **scattered** about a straight line. We use Linear Regression to get the formula y = 40x + 532, which is an estimate of the true straight line.

Formally, just as we estimate **μ** and **σ** by
$\bar{x}$ and **s**

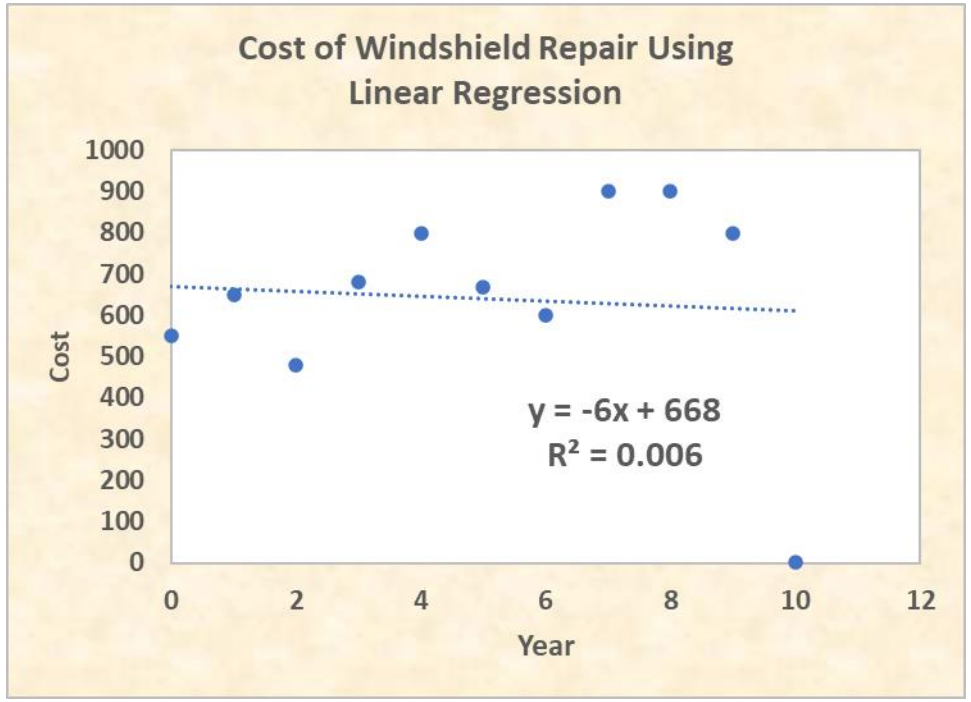
In **Linear Regression** we estimate y = **α**x + **β** by
y = **m**x + **b**

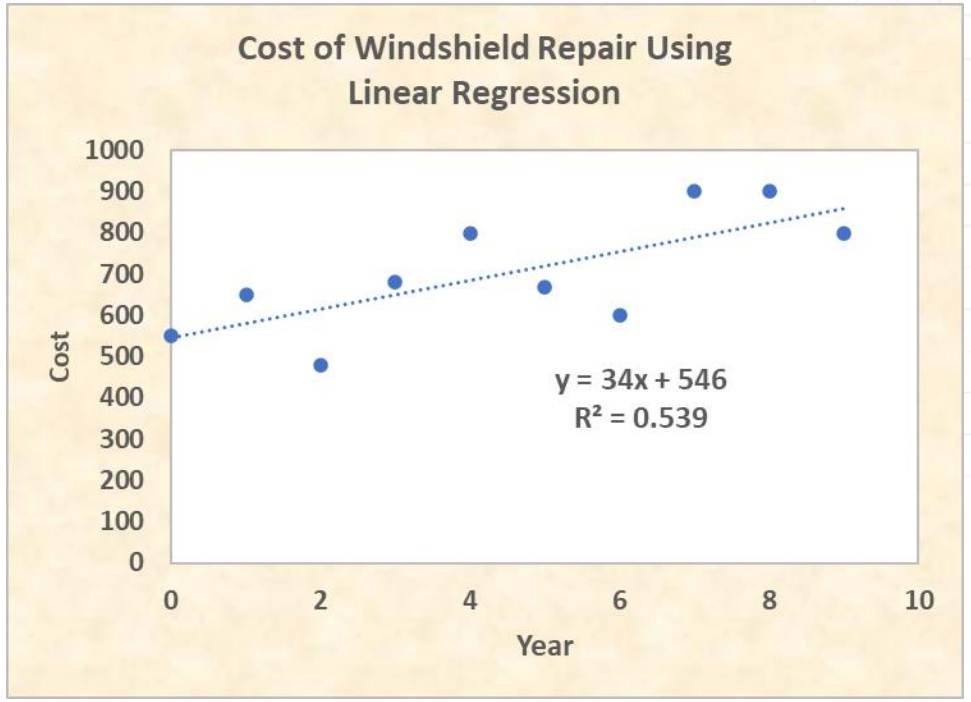
LINEAR REGRESSION

Khan Academy

Linear Regression

LINEAR REGRESSION - OUTLIERS





If two analysts get different answers, the first place they look is their SQL. But the reason could be the way they handle outliers. The right-hand graph has positive slope because it eliminated the outlier. The left has a negative slope because it did not eliminate the outlier

Outliers can influence the interpretation of a linear regression, even as far as finding a positive or negative slope.

ANALYST BOOT CAMP

EXCEL EXERCISES

Go to the Excel exercises in your workbook. The answers are to the far right on each tab.

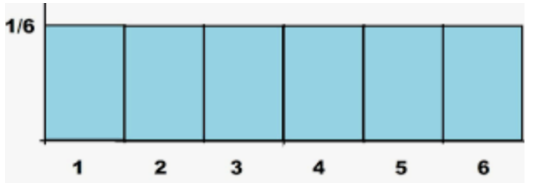
Tabs

- LinEq
- LinReg
- Outlr_LinReg

STATISTICAL DISTRIBUTIONS OF DATA

The notion of random variable and statistical distributions:

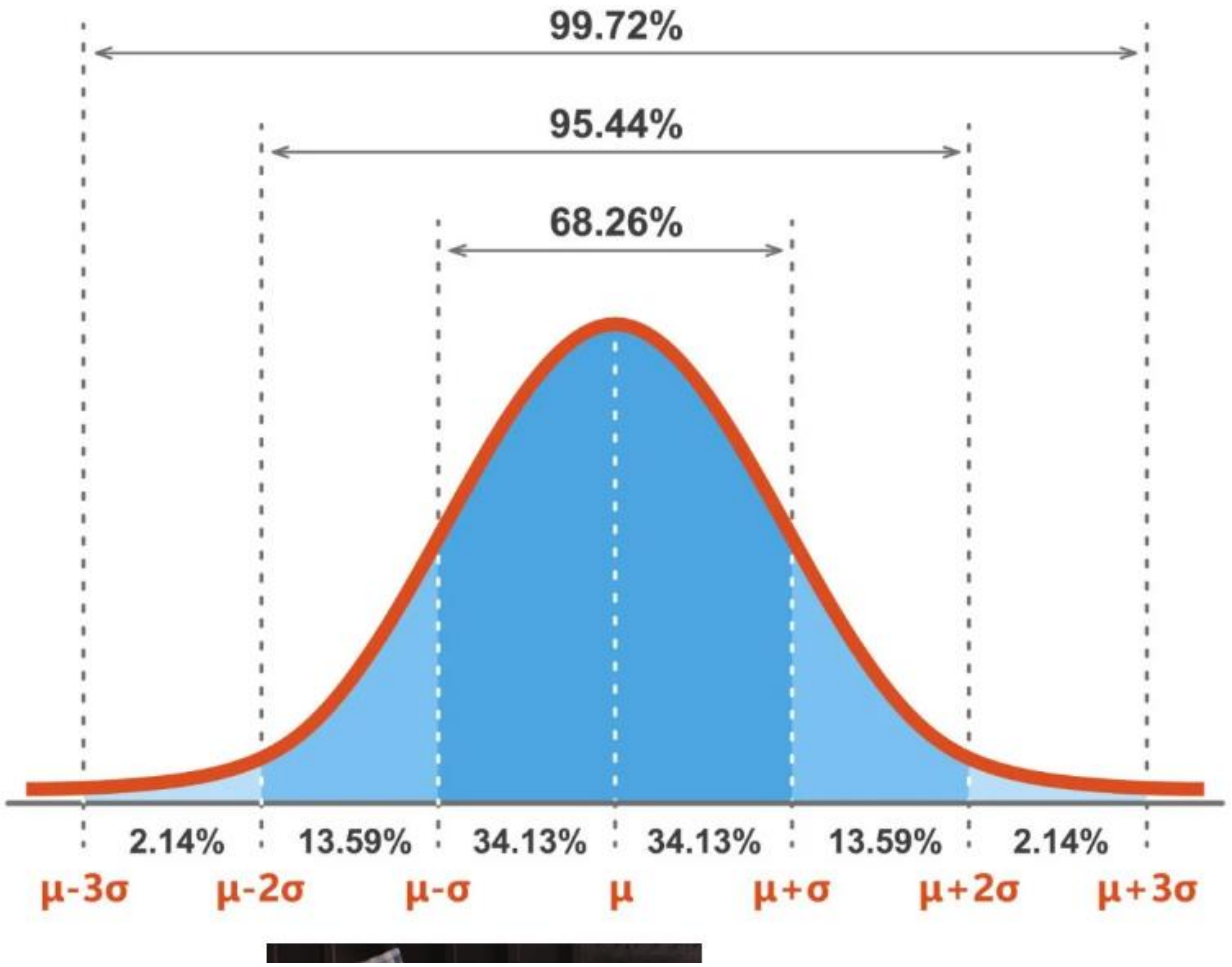
1. The value of a variable (X) is not fixed, it is random. (ex. roll a die).
2. But the value is not totally random. It has a smallest and largest possible value {1, 2, 3, 4, 5, 6}.
3. The X-values can be displayed on a horizontal number line.
4. The Y–value associated with the X-value is the probability of the X-value occurring. For the die, the probability of each X is 1/6.
5. This is a Uniform Distribution



6. The sum of all the probabilities is 1.0.

   1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1.0

NORMAL DISTRIBUTION – A SPECIAL CASE



Many things in the real world are **normally distributed** (ex. IQ, hand size).

The equation of the normal distribution is

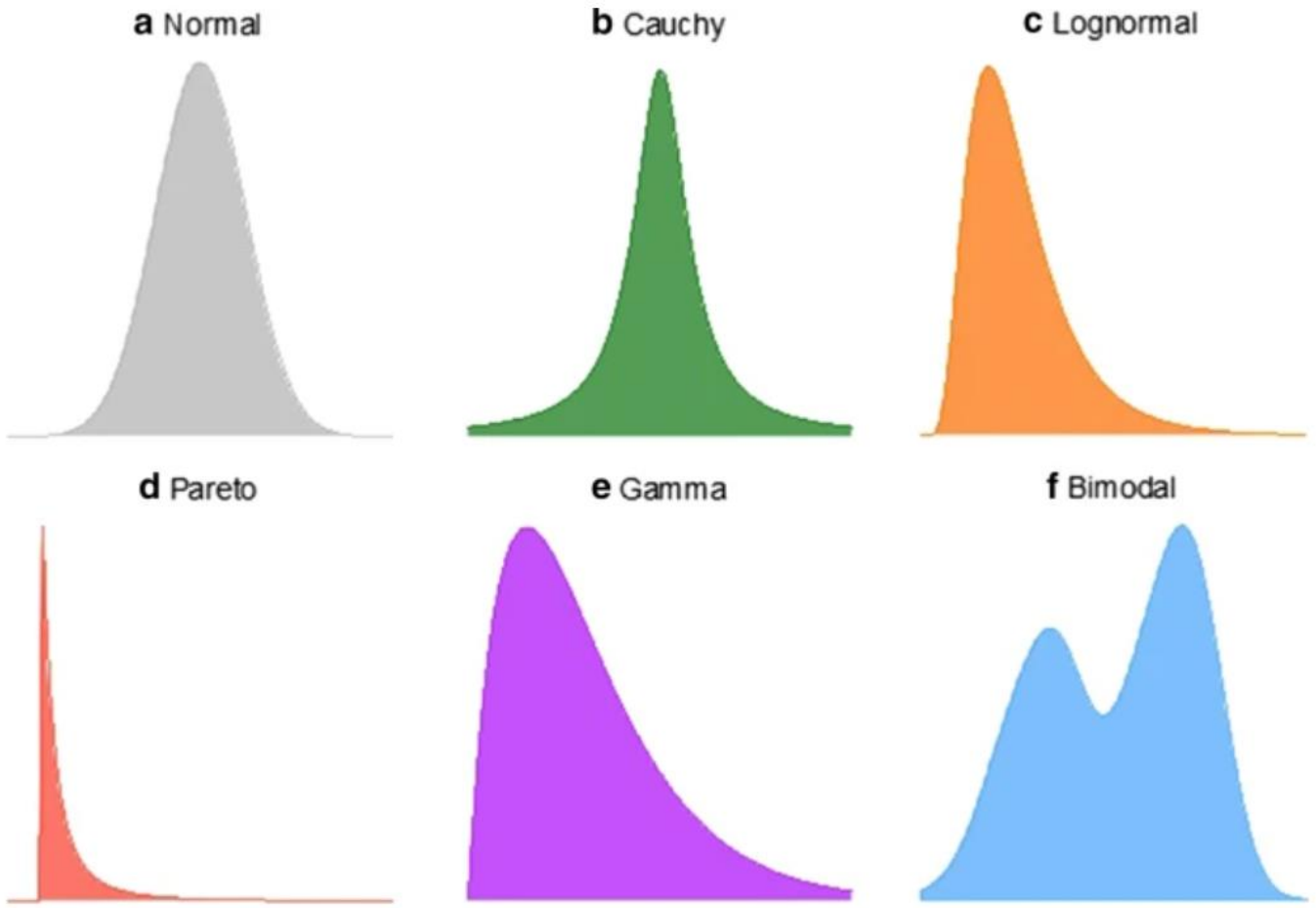
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The total area under the curve is 1.00. Furthermore:

- 68% of the area is within 1 SD of the mean,
- 95% within 2 SD,
- 99.7% within 3 SD.

So, if we have a normal distribution and we know the mean & SD, then we have an excellent idea of the possible values of the random variable.

STATISTICAL DISTRIBUTIONS OF DATA



We pay a lot of attention to Normal distributions, giving the impression it is the only one that ever occurs.

But many real-world data sets are not Normally distributed.

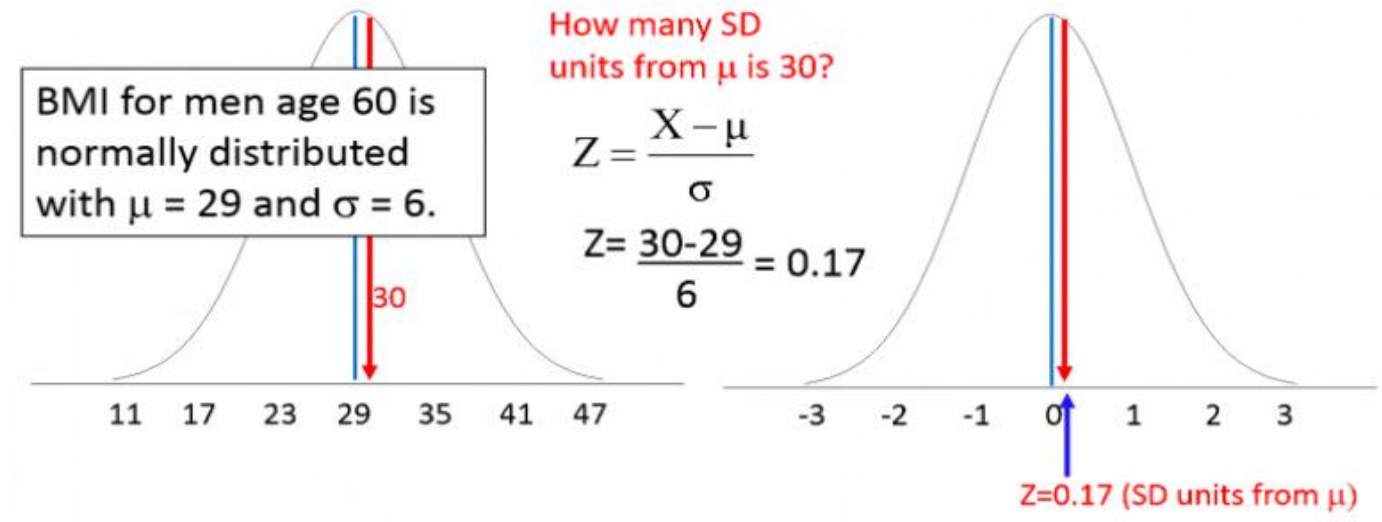
Can you give examples of data that has one of the distributions to the left?

NORMAL DISTRIBUTIONS

Khan Academy

Normal Distributions

Z-SCORES CONVERT ANY NORMAL DISTRIBUTION TO A STANDARD NORMAL DISTRIBUTION (MEAN=0 SD=1)

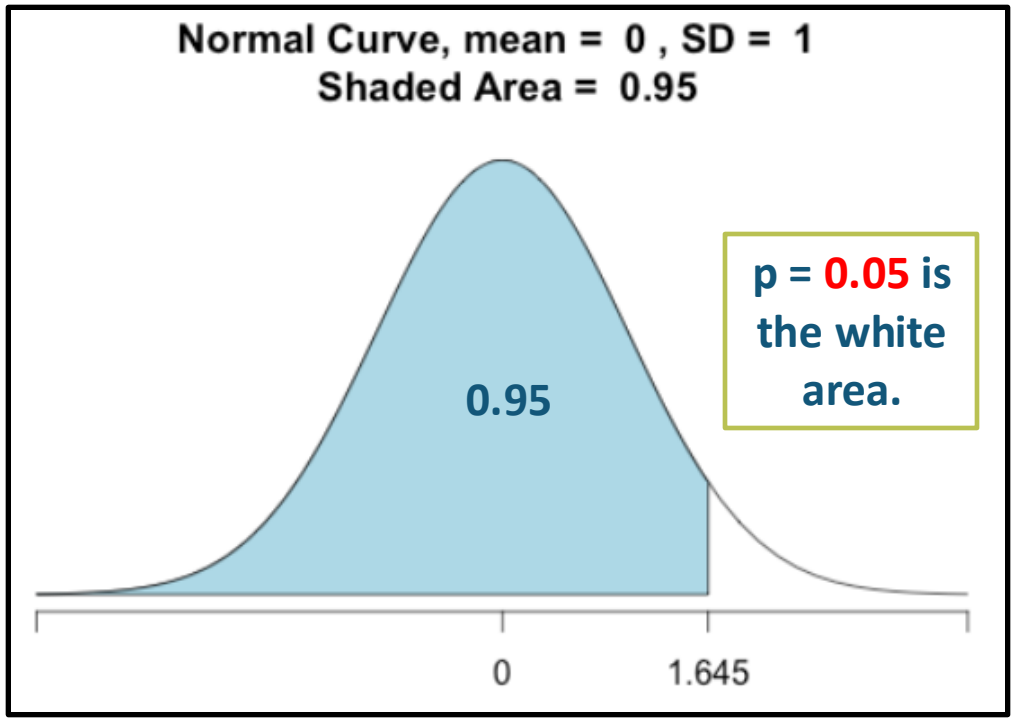


We re-frame (convert) the problem because there are published tables for the standard normal distribution.

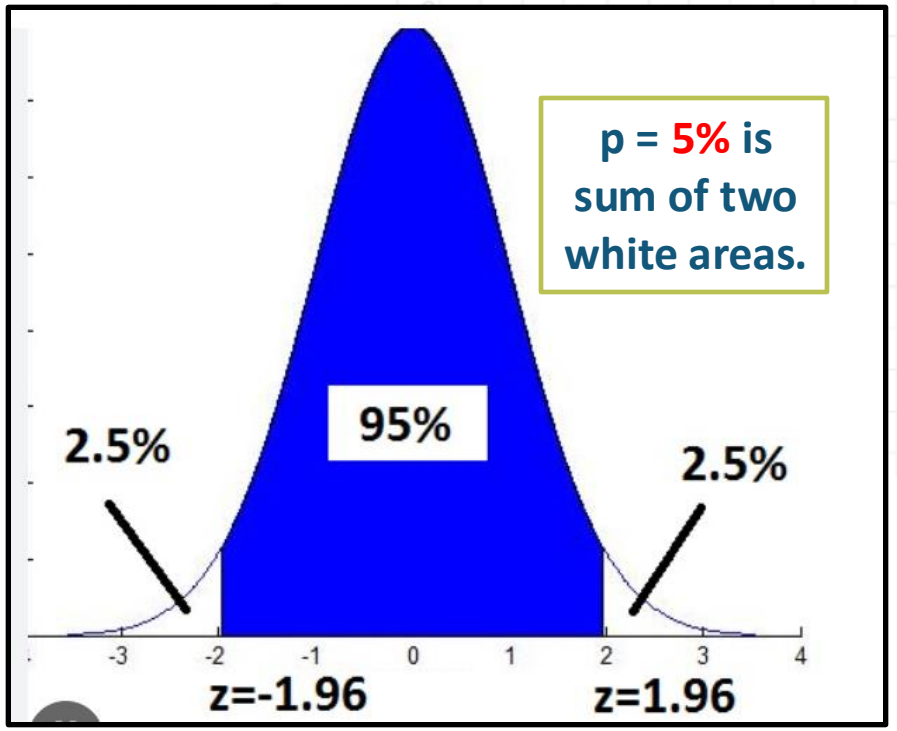
The figure on the left is a normal distribution, mean=29, standard deviation=6. What is the probability a BMI is less than 30? This is the area to the left of 30 since the total area = 1.00

The figure on the right re-frames the problem by using a Z-score to transform to a ”standard normal distribution” which has mean=0, standard deviation = 1. We want the area to the left of 0.17.

Z-SCORES FOR THE STANDARD NORMAL DISTRIBUTION WHEN P = 0.05



1.645 is the 1-tail Z-score for **p=0.05**



+1.96, -1.96 are the 2-tail Z-scores for **p=5%**

Note that 1.96 ~ 2.00

ANALYST BOOT CAMP

Z SCORES – KHAN

Khan Academy

Z Score Introductions

ANALYST
BOOT CAMP

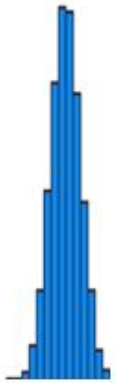
THE CENTRAL LIMIT THEOREM

Normal

Exponential

Triangular

Uniform



X-bar

The **Central Limit Theorem (CLT)** allows us to make a powerful statement about **the distribution of the mean (x-bar) of any shaped probability distribution** (top graphs).

Take many samples of size n >30 from the original distribution. Plot the means of those samples on a bar chart. That bar chart will be approximately normally distributed. And it will be skinnier than the original distribution because

the SD(mean) = SD(original distribution) divided by the square root of n.

We will do a class exercise using many samples of size 30 from the uniform distribution and plotting the means of those samples.

CENTRAL LIMIT THEOREM

Khan Academy

Central Limit Theorem

ANALYST
BOOT CAMP

A/B TESTING & CENTRAL LIMIT THEOREM

In A/B testing, we usually have a very large sample size. For example, in an A/B presentation I saw, the sample size was 64,000. The SQRT(64,000) is 253.

- The SD(mean) will be extremely small if the SD(sample) is divided by 253 and the Confidence Interval will be extremely small.
- There is also the issue of representativeness. Now if we are selecting alternatively 1 A, 1 B etc. this helps to assure representative A and B samples.
- So, for all practical purposes, the A/B test results can be treated as “exact numbers” and you don’t need to compute confidence intervals around the averages.

EXCEL EXERCISES

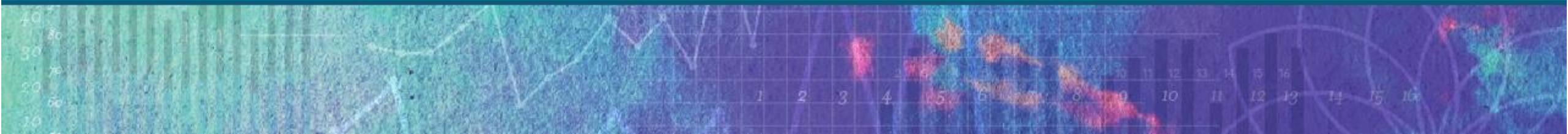
Go to the Excel exercises in your workbook. The answers are to the far right on each tab.

Tabs

- NormDist
- Z-score
- CLT1
- CLT2

ANALYST
BOOT CAMP

ADVANCED STATISTICS



ADVANCED STATISTICS TOPICS

- Hypothesis Tests for large samples: z-score
- Confidence Intervals
- Hypothesis Test for small samples: t-score

HYPOTHESIS TESTING

General Logic of Hypothesis Testing (Logic of Opposites)

We hope to demonstrate that something is meaningful. To do so, we assume it is not meaningful and proceed to refute that assumption. Since there are only two possibilities, the alternative, "it is meaningful", must be true.

Example:

1. We assume the Null Hypothesis: The true mean is = 0.
   The Alternative Hypothesis is: The true mean is not =0.
2. We allow for a 5% chance that we might be wrong. From a previous slide, the value of Z such that 5% of the area is to the right of Z is 1.645. That is the *critical Z*.

3. Calculate the Z-score for the given data's mean, assuming the true mean is zero.
4. If calculated Z is > *critical Z*, that event is "highly unlikely to occur", so the true mean must not be =0.
5. If calculated Z is < *critical* Z, we can't conclude that the true mean is not =0.

ONE-TAIL HYPOTHESIS TEST FOR THE MEAN

Example: The Countrywide Cost of a Total Loss claim is $21,000. A sample of 100 Total Loss claims from Virginia has a sample mean of $23,170 and a sample SD of $7,000. Is Virginia's true mean greater than $21,000 at the p=0.05 level of significance?

1. Null Hypothesis: VA true mean = 21,000.
   Alternative Hypothesis: VA true mean > 21,000.
2. This is 1-tail because we want to know if it is "greater than". Critical Z is 1.64.

3. SD(mean) = [7000/SQRT(100)] = 7000/10 = 700
4. The calculated Z-score is [23,170-21,000] / [SD(mean)] = 2170 / 700 = 3.1
5. Since 3.1 > 1.64, it is "highly unlikely" VA true mean is 21,000. We conclude the VA true mean is >21,000.

What If n=25? SD(mean)=7000/(sqrt(25)) = 1400. Z-score=2170/1400=1.55. Cannot reject the null hypothesis

TWO-TAIL HYPOTHESIS TEST FOR THE MEAN

Example: The average annual PTO time for the last 5 years in Claims has been 3.0 weeks. A sample of 100 PTO records for the current year-to-date has a sample mean of 2.75 weeks and a sample SD of 2 weeks. Is this year's true mean different than the historical 3 weeks at the p=0.05 level of significance?

1. Null Hypothesis: the true mean this year is = 3
   Alternative Hypothesis: the true mean this year is not = 3.
2. This is 2-tail because "different from" could be positive or negative. The Critical Zs are -1.96 and +1.96.

3. SD(mean) = [2/SQRT(100)] = 2/10 = 0.2
4. The calculated Z-score is [2.75-3] / [SD(mean)] = -0.25/0.2 = -0.125
5. Since -0.125 > -1.96, we cannot reject the null hypothesis that this year's true mean is 3 weeks..

HYPOTHESIS TESTING

Khan Academy

Hypothesis Testing

ANALYST
BOOT CAMP

95% CONFIDENCE INTERVALS

We are 95% confident the true mean is within that interval.

Example: Roadway Service Cost in Pennsylvania

What is a 95% CI for the true mean if the Sample Mean = 41, Sample SD = 15, and the sample size = 25?

1. The sample mean is normally distributed. A 95% interval is the Sample Mean plus/minus 1.96 x SD(mean).

2. The SD(mean) is SD(sample)/SQRT(n) = 15/SQRT(25) = 15/5 = 3

3. A 95% CI in which the true mean lies is 41-(1.96)(3) to 41+(1.96)(3) = 35.1 to 46.9

If a 95% confidence interval for the mean does not include zero, that is equivalent to a Hypothesis Test with p=0.05 rejecting the null hypothesis that mean=0. In my opinion, it is easier to run and explain.

CONFIDENCE INTERVALS

Khan Academy

Confidence Intervals

ANALYST
BOOT CAMP

A/B TESTING RELATIONSHIP TO CONFIDENCE INTERVALS

In A/B testing, we usually have a very large sample size. For example, in the earlier A/B presentation, the sample size was 64,000. The SQRT(64,000) is 253.

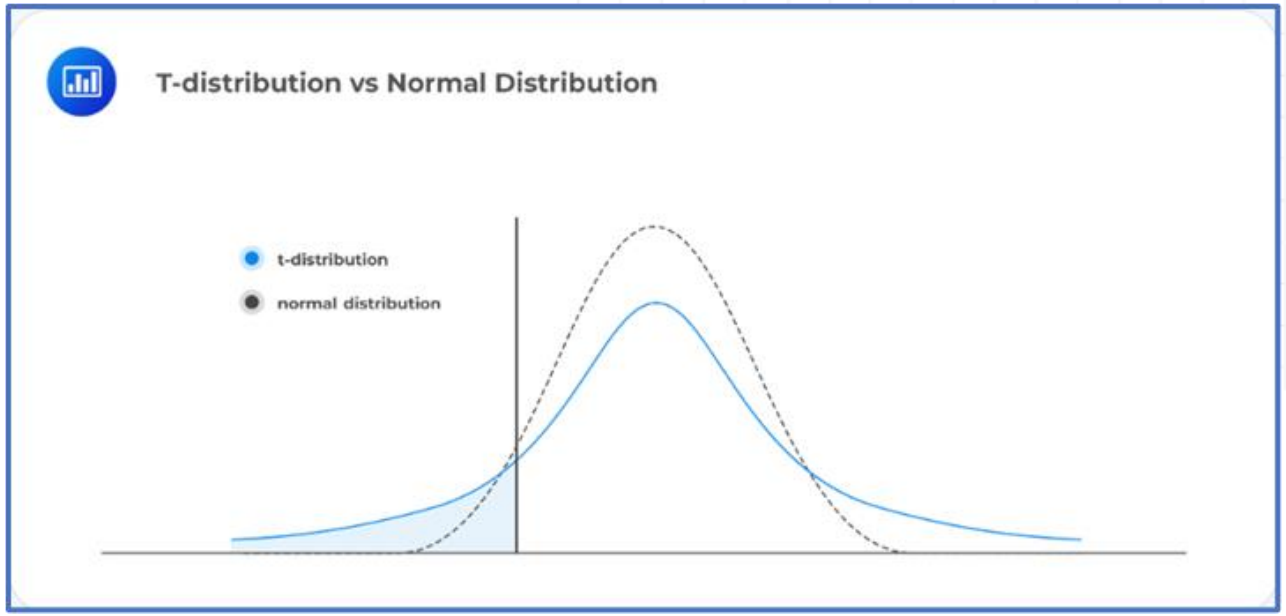
- The SD(mean) will be extremely small if the SD(sample) is divided by 253. The CI will then also be extremely small. So small that for all practical purposes you can treat the results as being “exact”.
- So, the values you got (below) can be treated as exact numbers when making the A vs. B comparison and you don’t need to compute confidence intervals around the averages.

  - 51.2% for the average of A
  - 51.7% for the average of B

Hypothesis Test for Small Data: t-Test

Critical t-values for p=0.05			
1 tail		2 tail	
df=(n-1)	value	df=(n-1)	value
1	6.314	1	12.706
2	2.920	2	4.303
3	2.353	3	3.182
4	2.132	4	2.776
5	2.015	5	2.571
6	1.943	6	2.447
7	1.895	7	2.365
8	1.860	8	2.306
9	1.833	9	2.262
10	1.812	10	2.228
11	1.796	11	2.201
12	1.782	12	2.179
13	1.771	13	2.160
14	1.761	14	2.145
15	1.753	15	2.131

Critical t-values for p=0.05			
1 tail		2 tail	
df=(n-1)	value	df=(n-1)	value
16	1.746	16	2.120
17	1.740	17	2.110
18	1.734	18	2.101
19	1.729	19	2.093
20	1.725	20	2.086
21	1.721	21	2.080
22	1.717	22	2.074
23	1.714	23	2.069
24	1.711	24	2.064
25	1.708	25	2.060
26	1.706	26	2.056
27	1.703	27	2.052
28	1.701	28	2.048
29	1.699	29	2.045
30	1.697	30	2.042
60	1.671	60	2.000
120	1.658	120	1.980
∞	1.645	∞	1.960



ANALYST BOOT CAMP

Hypothesis Test for Small Data: t-Test

A t-test is run the same way as a z-test but is used for small sample size.

Ex: Hailstone size in KS

A sample of size n=9 has a Sample Mean=6.4 inches and a sample SD=12. Is the true mean different from 4 inches (last year's value) at the p=0.05 level of significance?

1. From the t-table, the 95% critical value of t for a sample of size 9 [(n-1)=8 df] , is 2.306 (larger than z=1.96)

2. SD of the mean = 12/SQRT(9) = 12/3 = 4

3. t-score = (6.4 – 4)/4 = 2.4/4 = 0.60, which is less than 2.306.
   We cannot say the true mean is different from 4

EXCEL EXERCISES

Go to the Excel exercises in your workbook. The answers are to the far right on each tab.

Tabs

- HypTest
- ConfInt
- tDist

COURSE GOALS REVISITED

Demonstrate a basic knowledge of Analytics related Algebra concepts.

Demonstrate an understanding of basic Statistics knowledge.

POST-TEST

- 2 X 2 + 2 X 2 = 4 + 4 = 8
- Rebase the series to the last number: 5, 7, 15, 12, 6 0.83, 1.17, 2.5, 2.0, 1.0
- What are the first 2 terms of a MA2 for the series above: (5+7)/2, (7+15)/2 = 6, 11
- The average is a measure of the central tendency of a set of data.
- The standard deviation is a measure of how compact or spread out the data is
- Given a Normal Distribution with mean=100 and standard deviation=15, what are the Z-scores that include 95% of the data values? [70, 130]
- Explain the Central Limit Theorem? No matter what the distribution of a set of data, the mean of that set of data has a normal distribution.

Advanced

- Given a data set of 15 values, and we want to conduct a hypothesis test on the mean, should we use the z-score or t-score? __
- The most common level of significance used for Confidence Intervals is: ____

ALGEBRA ON A PAGE

Multiplication $*, \times, \cdot$

Division $/, \frac{a}{b}, \div$

Parentheses ( )

Brackets [ ], { }

Exponents x ^ 2, x^2, x ** 2

Subscripts x_1, x_2, x_3

Logical Symbols

$=$ Equal to

$\neq$ Not Equal to

$\leq$ Less than or equal to

$\geq$ Greater than or equal to

$<$ Less than

$>$ Greater than

Rebasing & Relativities

The act of dividing by a "base" or "reference" level is known as **rebasing**. This process is often used to create easy to digest measures called **relativities**. It is common to see analyses report measures such as "Loss Ratio Relativities" or "Pure Premium Relativities".

Orders of Magnitude & Scaling

Explicitly display or represent the scaling base

When conducting analysis, be mindful of scaled variables

Linear Equation

$$y = mx + b$$

"Straight Line", explains a relationship.

… as x increases, y increases or decreases a specified amount based on the slope (m)

PIFs increase from 2000 in Year 1 to 2600 in Year 2. What is the Annual Growth Rate?

Method 1:	(End-Start)/Start	(2600 - 2000) / 2000	= 0.30	x	100%	= 30%
Method 2:	(End/Start) - 1	(2600/2000) -1	= 1.30 -	= 0.30	x 100%	= 30%

Weighted Average

Segment	Value	Weight	Wt x Value
A	5	10	50
B	2	8	16
C	1	2	2
D	4	5	20
Average	3.00	25	88
	Weighted Average		3.52
			88/25

Order of Operations

Parenthesis
Exponents
Multiplication/Division
Addition/Subtraction

Square: number multiplied to itself.

Exponents: apply multiplication *n* times.

ANALYST BOOT CAMP

STATISTICS ON A PAGE

Population:

An entire group of people or objects of interest

Sample:

A subset of the population that is studied used to make an inference about the population.

Bias

The sample is not representative of the population.

Inferential Statistics:

Make predictions about the future based on previous data.

Outliers:

Data values that fall outside the expected ranges

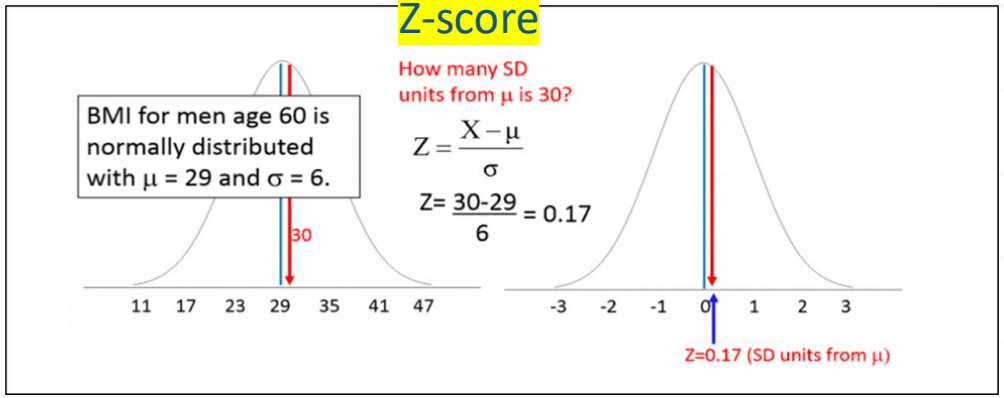
Mean, Median, Mode

tells us where the data is centered.

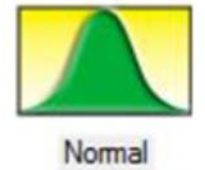
Standard Deviation

tells us how compact or spread out the data is.

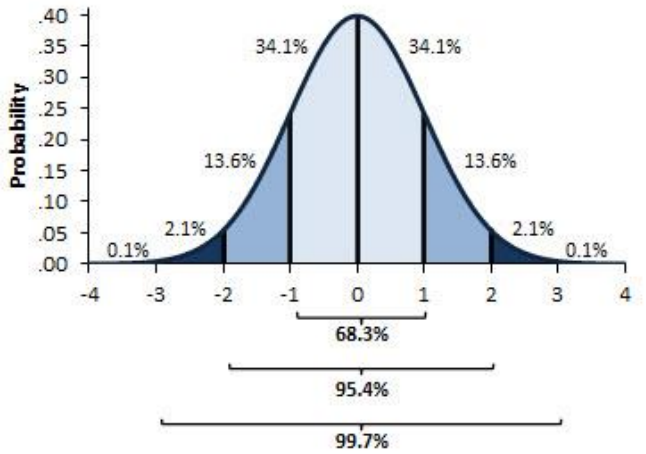
Z-score



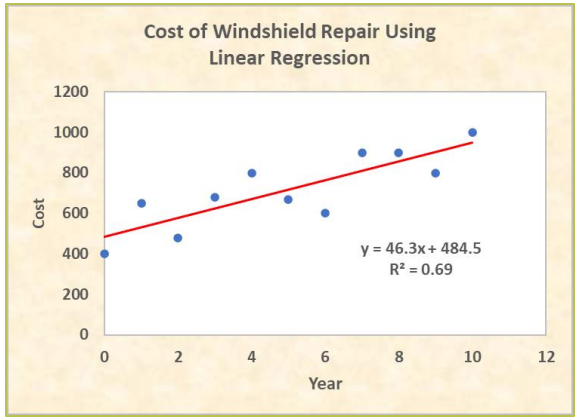
Central Limit Theorem

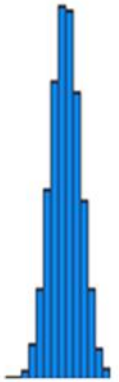


Normal Distribution



Linear Regression





X-bar

ANALYST BOOT CAMP