STRINGFEST ANALYTICS

EXCEL STATISTICS FOR BUSINESS ANALYTICS

George Mount

Data Analyst & Educator at Stringfest Analytics

George works as an independent analyst and data analytics educator with the goal to help clients manage their data so they think more creatively. He serves as a technical expert and lead curriculum developer for Thinkful's data analytics program and is the instructor of the DataCamp course "Survey and Measure Development in R."

George blogs about data, innovation, and career development at georgejmount.com. He holds a master's degree in information systems with a certificate of achievement in quantitative methods from Case Western Reserve University

COURSE OBJECTIVES

- Explore a dataset for research questions
- Check assumptions and build hypotheses
- Test formally for a difference in means between two groups
- Make compelling business recommendations using inferential statistics

WHY WOULD WE DO THIS IN EXCEL?

"You get to look at the data every step of the way, building confidence while learning the tricks of the trade."

-- John Foreman

FOLLOWING ALONG

- Each section is a sub-folder
- Demos = follow along with me
- Drills = try it yourself
  - Refresh your memory with the demo notes

1. EXPLORATORY DATA ANALYSIS IN EXCEL

What the hey is EDA?

Preliminary understanding of variables

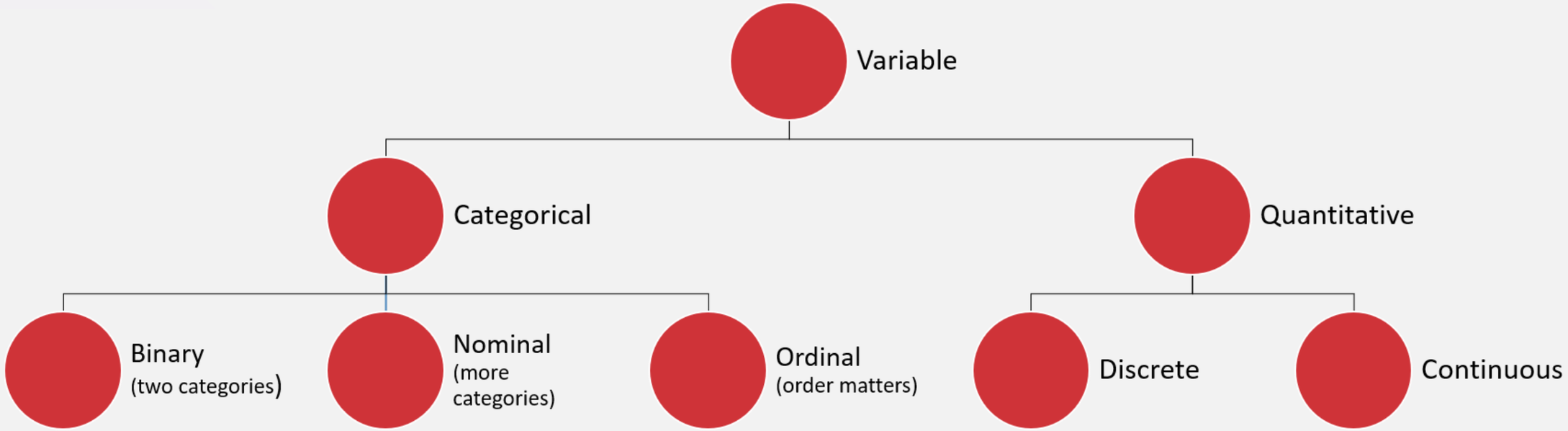
- Classify them
- Summarize them
- Visualize them
- Check assumptions before using them

What is a variable?

- Classical statistics: at least a 90%/10% split

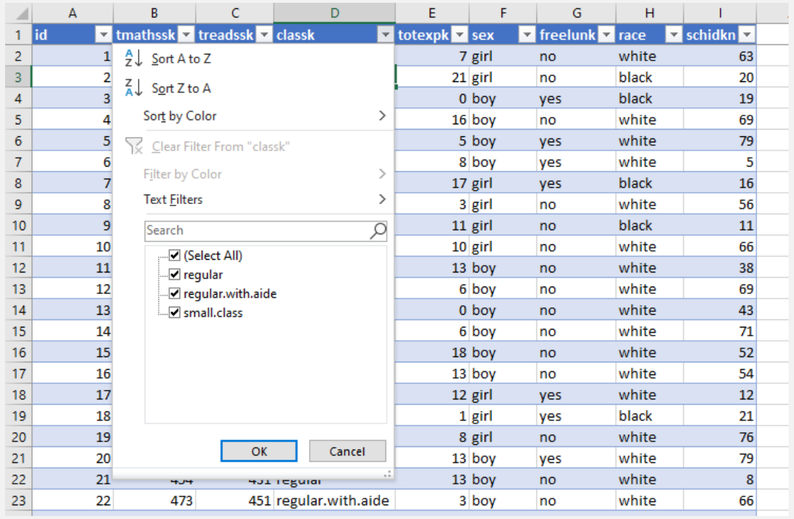
VARIABLE TYPES AS RECIPE INGREDIENTS

WHAT'S ON THE TABLE?



DRILL

- File: `star.xlsx`
- What type is each variable?
- (Don't be afraid to poke at the data!)

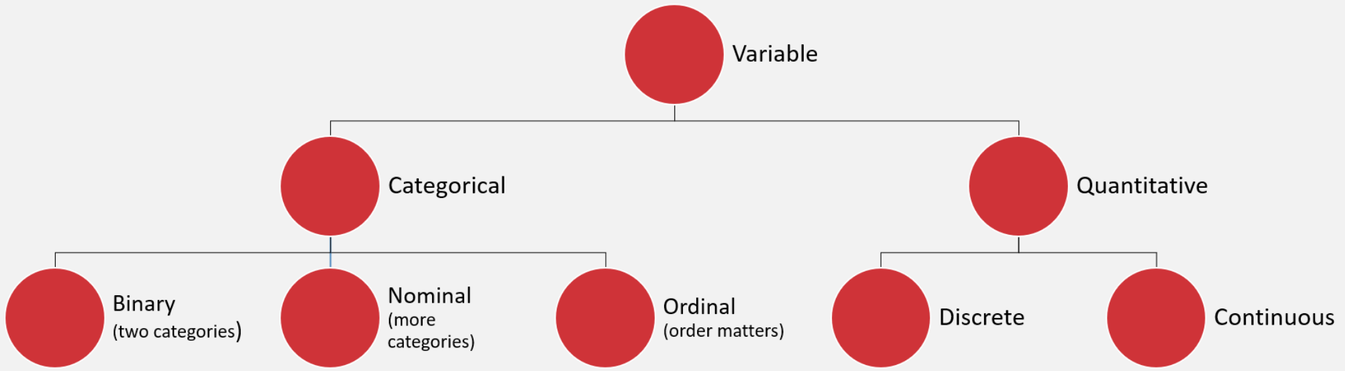


id	tmathssk	treadssk	classk	totexpk	sex	freelunk	race	schidkn
1				7	girl	no	white	63
2				21	girl	no	black	20
3				0	boy	yes	black	19
4				16	boy	no	white	69
5				5	boy	yes	white	79
6				8	boy	yes	white	5
7				17	girl	yes	black	16
8				3	girl	no	white	56
9				11	girl	no	black	11
10				10	girl	no	white	66
11				13	boy	no	white	38
12				6	boy	no	white	69
13				0	boy	no	white	43
14				6	boy	no	white	71
15				18	boy	no	white	52
16				13	boy	no	white	54
17				12	girl	yes	white	12
18				1	girl	yes	black	21
19				8	girl	no	white	76
20				13	boy	yes	white	79
21				13	boy	no	white	8
22	473	451	regular.with.aide	3	boy	no	white	66

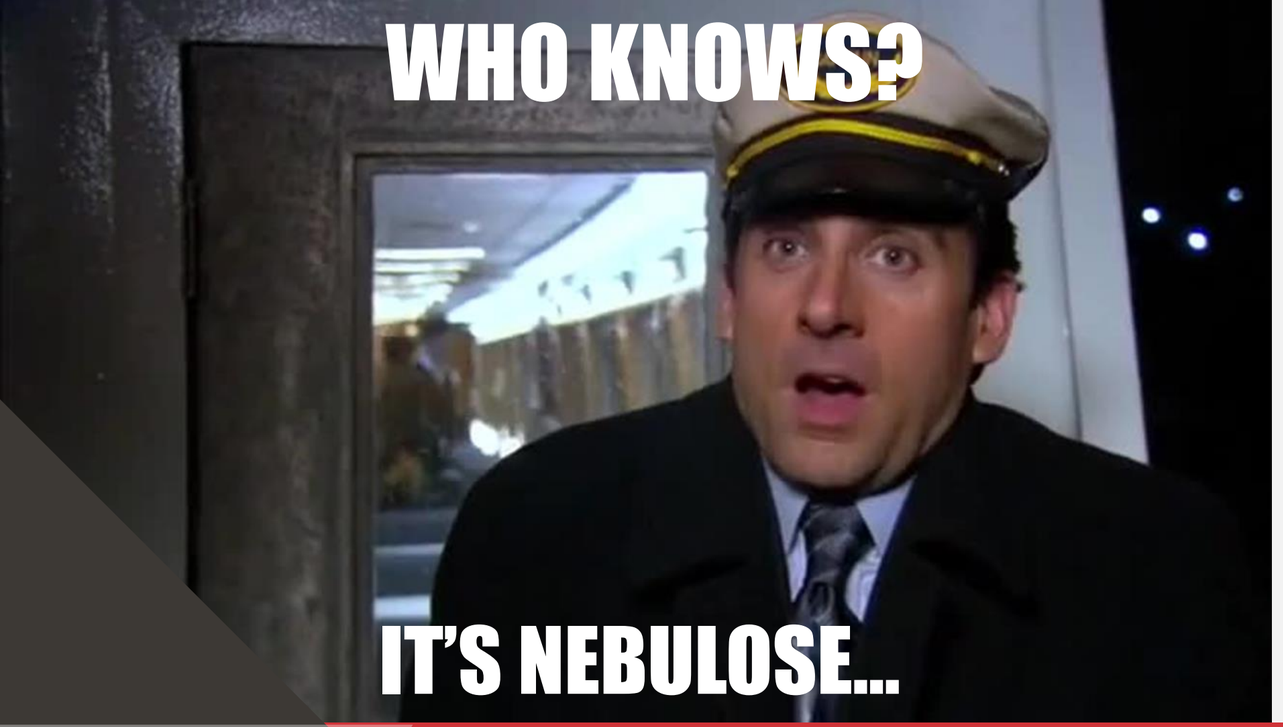
Variable	Description	Type?
`tmathssk`	Total math scaled score	
`treadssk`	Total reading scaled score	
`classk`	Type of class	
`totexpk`	Years of total teaching experience	
`sex`	Sex	
`freelunk`	Qualified for free lunch?	
`race`	Race	
`schidkn`	School indicator	

DRILL

- File: `star.xlsx`
- What type is each variable?
- (Don't be afraid to poke at the data!)
  - (Filtering for now...)



Variable	Description	Type?
`tmathssk`	Total math scaled score	Continuous
`treadssk`	Total reading scaled score	Continuous
`classk`	Type of class	Nominal
`totexpk`	Years of total teaching experience	Discrete
`sex`	Sex	Binary
`freelunk`	Qualified for free lunch?	Binary
`race`	Race	Nominal
`schidkn`	School indicator	Nominal



VARIABLE TYPES ARE CONTEXTUAL

https://www.theanalysisfactor.com/level-of-measurement-not-obvious/

Summarizing a variable: frequencies

- Used for categorical variables
- How many of each value do we have? (Counts)
- Use PivotTables

DRILL

- Demo: `wages.xlsx`
  - (Don’t forget about the demo notes)

Summarizing a variable: descriptive statistics

- Used for continuous variables
- Central tendency: what value is the variable centered around?
  - Mean: “average”
  - Median: “middle”
  - Mode: “most common”

DRILL

- You are consulting for a non-profit on fundraising strategies
- Donation data is shown to the right
- Which measure of central tendency (mean/median/mode) should be tracked?

Variable
10
10
20
30
100

DRILL

- You are consulting for a non-profit on fundraising strategies
- Donation data is shown to the right
- Which measure of central tendency (mean/median/mode) should be tracked?
- **No one statistic rules them all**

Variable
$10
$10
$20
$30
$100

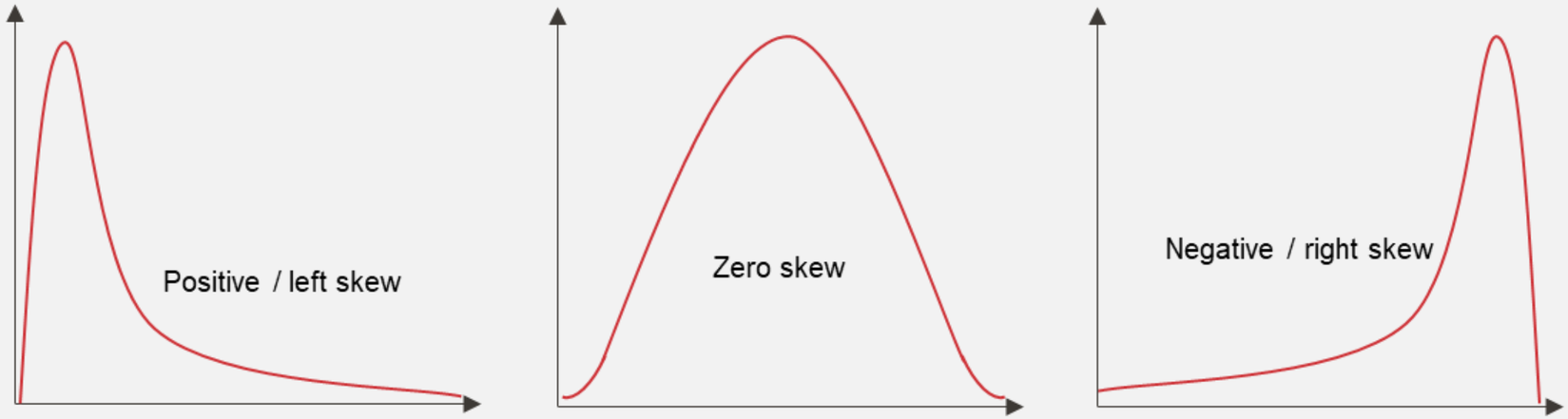
Statistic	Value
Mean	$34
Median	$20
Mode	$10

Summarizing a variable: descriptive statistics

- Variability: how spread is the variable from the center?
  - Range: What are the min/max?
  - Variance: Relative to the mean, how far away does the data fall?
  - Standard deviation: Square root of the variance
  - Standard error: accuracy with which a sample distribution represents a population (standard deviation / sample size)

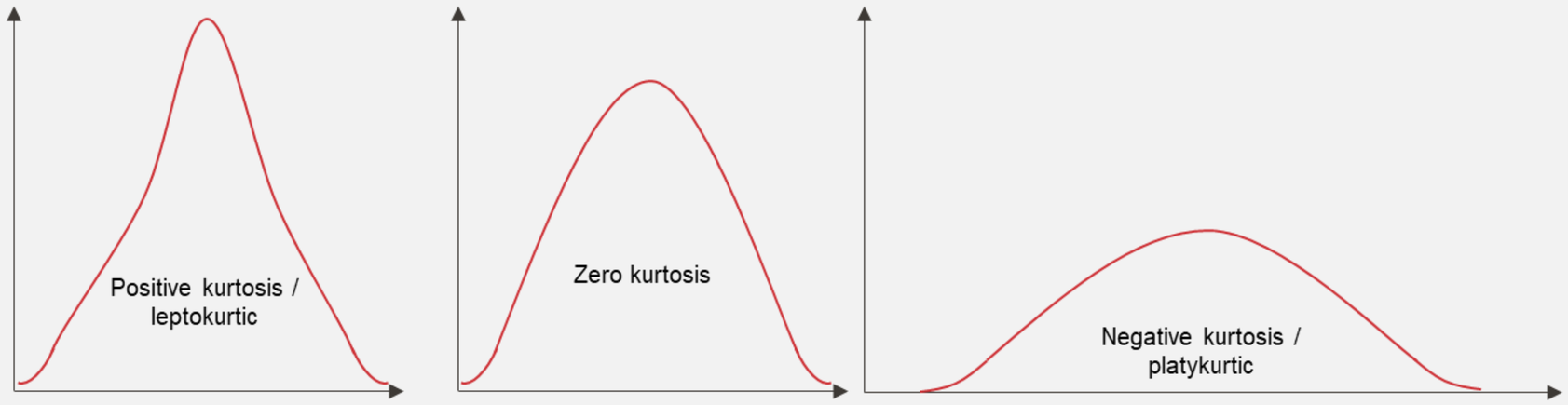
Summarizing a variable: descriptive statistics

- Distribution: How is the data distributed?
  - Skewness: How lopsided is the data?



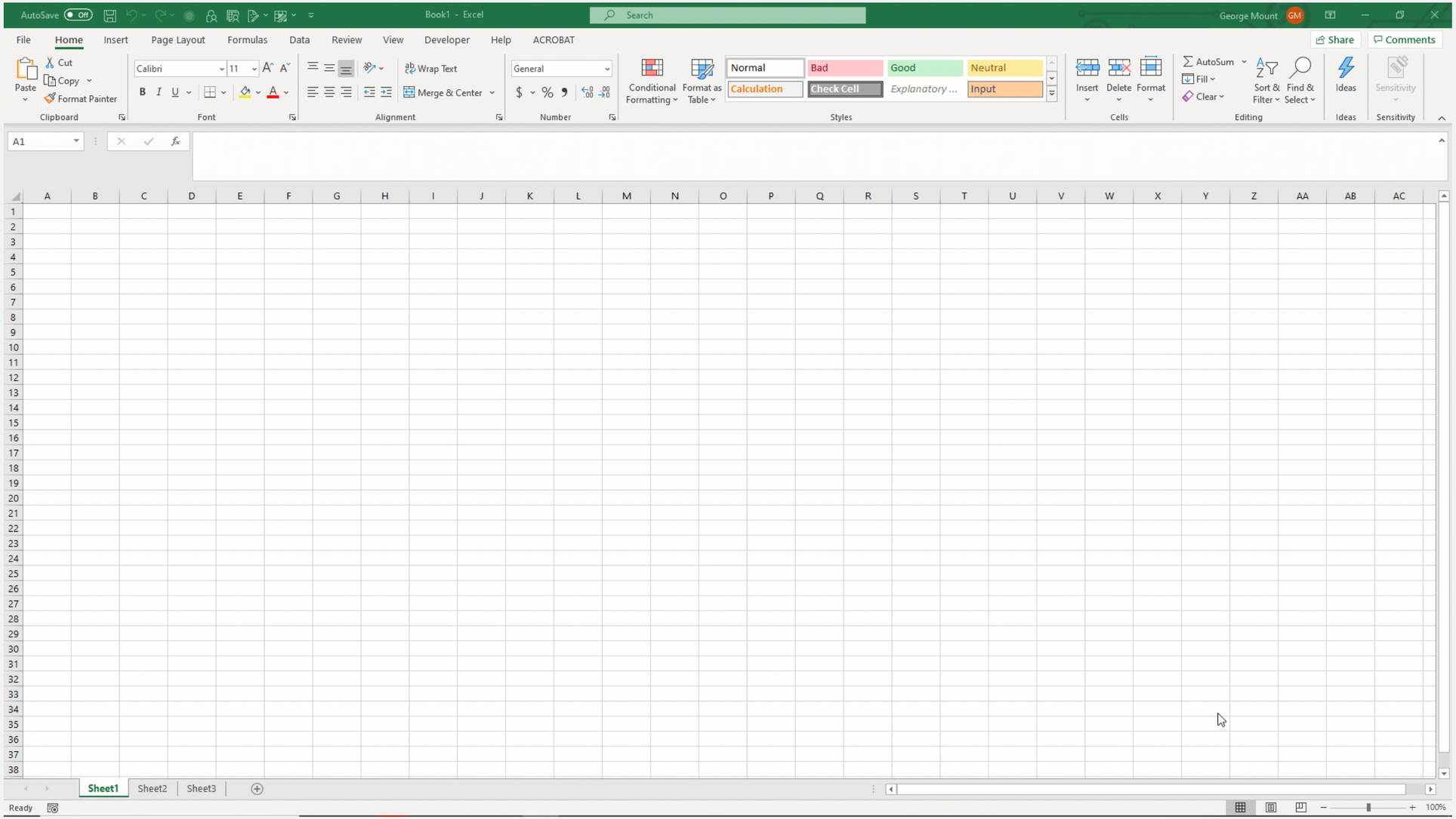
Summarizing a variable: descriptive statistics

- Distribution: How is the data distributed?
  - Kurtosis: How jagged is the data?

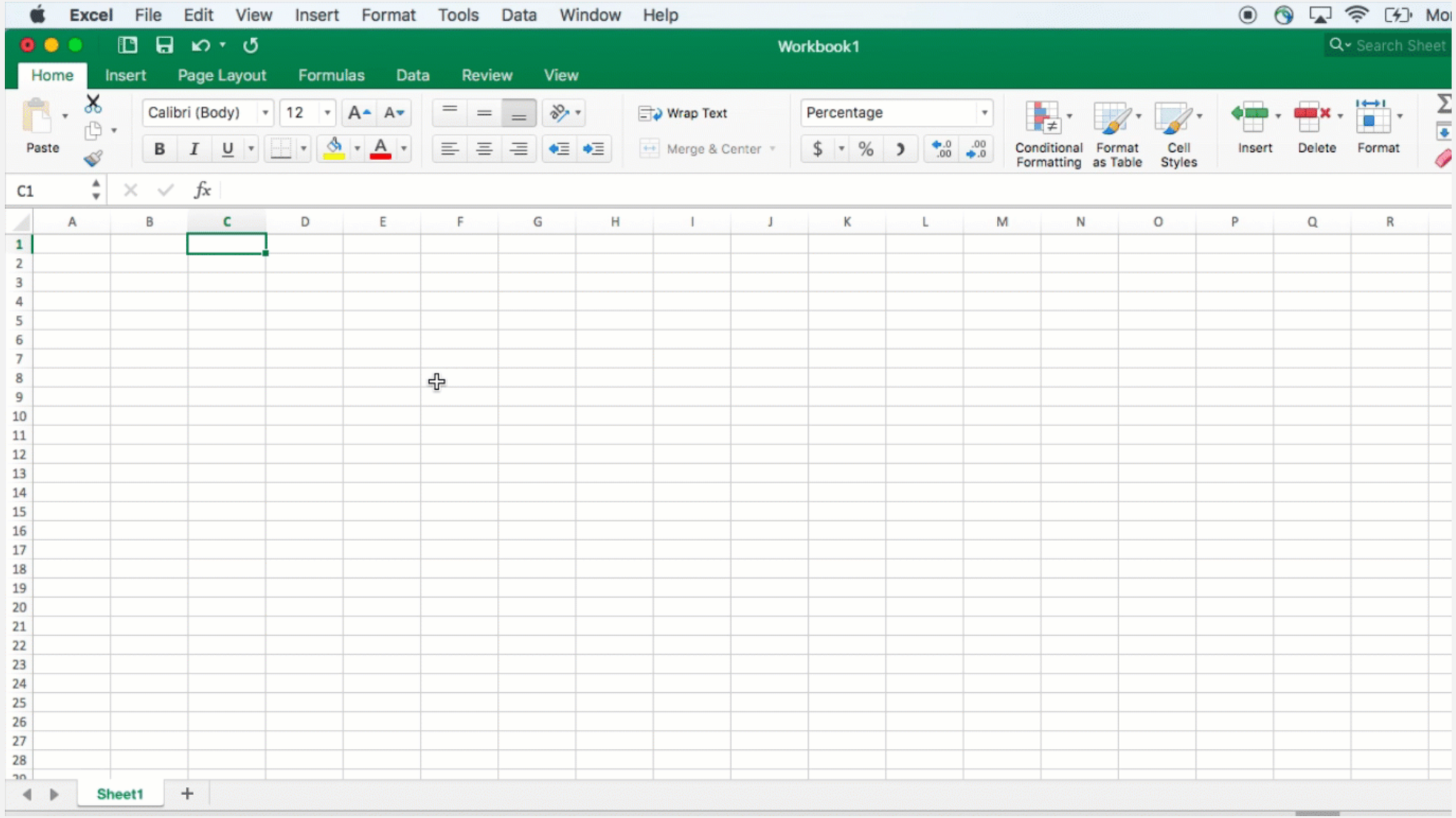


HAVE YOU INSTALLED THE DATA ANALYSIS TOOLPAK?

ON WINDOWS:


AutoSave
Book1 - Excel
Search
George Mount
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Developer
Help
ACROBAT
Share
Comments
Paste
Cut
Copy
Format Painter
Clipboard
Calibri
11
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Normal
Bad
Good
Neutral
Calculation
Check Cell
Explanatory ...
Input
Styles
Insert
Delete
Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
Ideas
Sensitivity
A1
Sheet1
Sheet2
Sheet3
Ready
100%

ON MAC:


Excel File Edit View Insert Format Tools Data Window Help
Workbook1
Search Sheet
Home Insert Page Layout Formulas Data Review View
Paste
Calibri (Body)
12
Wrap Text
Merge & Center
Percentage
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
C1
Sheet1

DEMO

- File: `wages.xlsx`
- Summarize `wage` with the Analysis ToolPak

Summarizing a variable: visualization

- Histogram
  - What does the data's distribution "looks like?"
  - How many data points lie with each bin?

DEMO

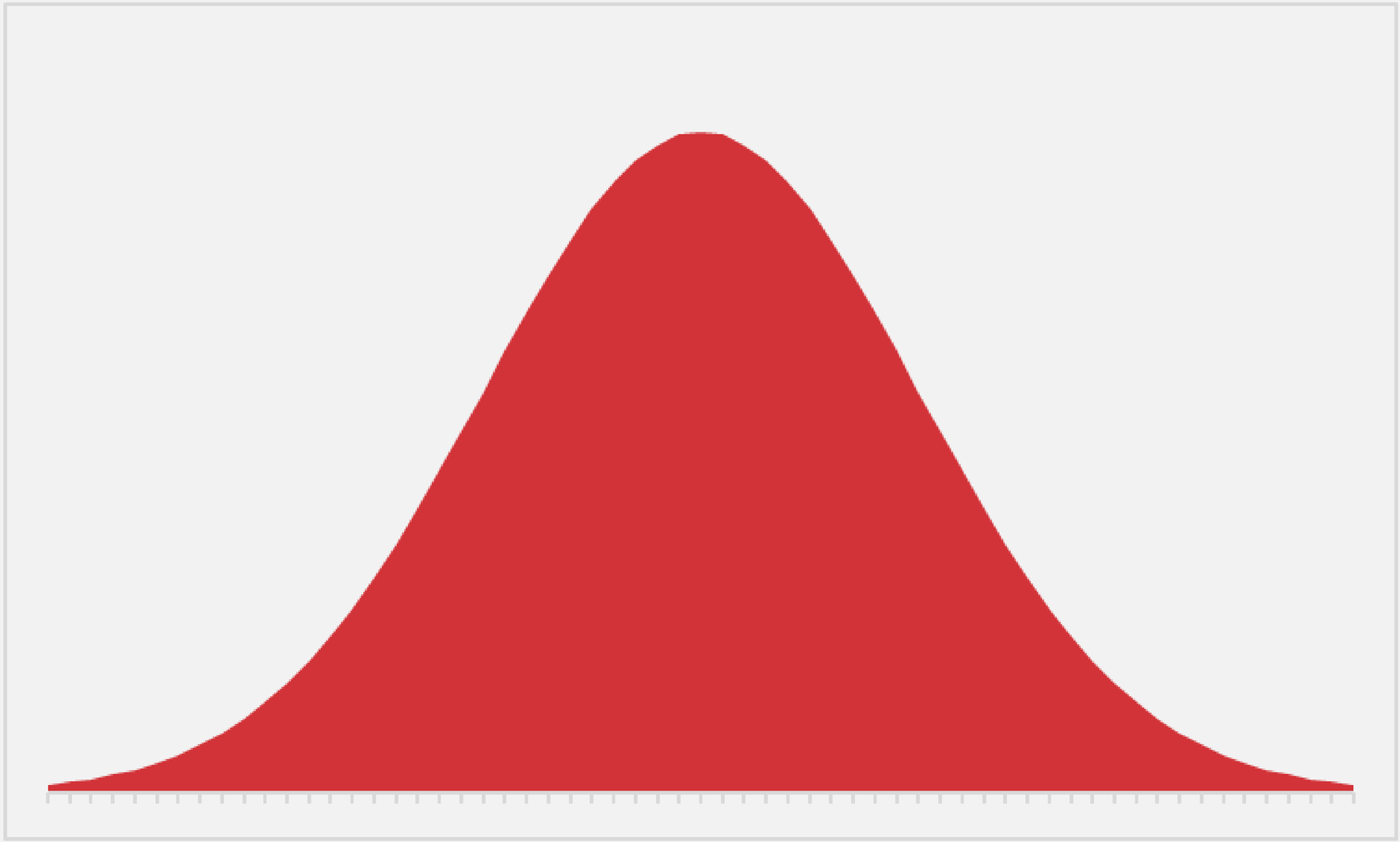
- File: `wages.xlsx`
- Visualize `wage` with a histogram

DRILL

- File: `housing-descriptive.xlsx`
- Explore this dataset:
  - Choose a categorical variable and run its frequencies
  - Choose a continuous variable, run its descriptives and plot a histogram

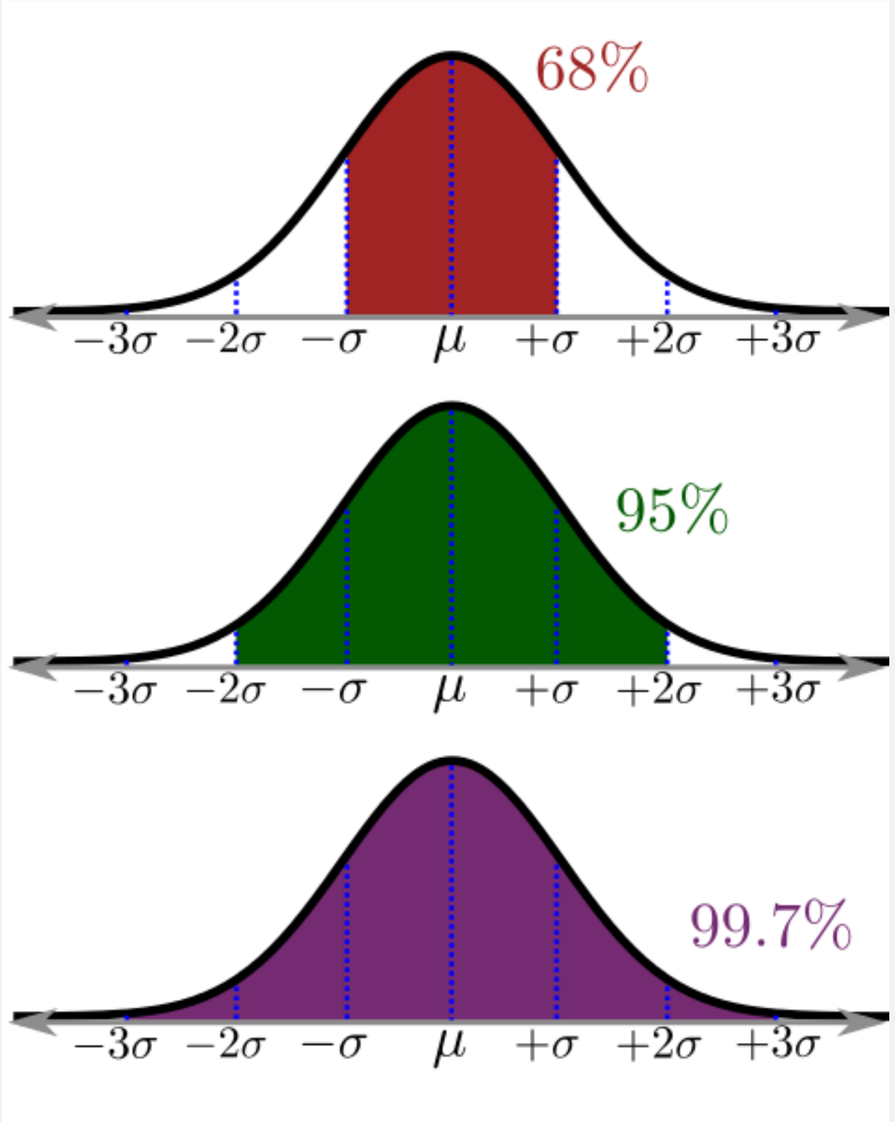
QUESTIONS?

EVER WANTED TO JUST BE NORMAL? STATISTICS IS YOUR CHANCE.



WHAT DOES IT MEAN TO BE NORMAL?

“Empirical rule”



https://www.katemarshallmaths.com/chapter-3-bell-shaped-distribution.html

SPOTTING NORMAL DISTRIBUTIONS

There are no normally distributed variables, but some are more normal than others:

1. Descriptives: skewness ± 8, kurtosis ± 20
2. Visualizations: does it “look” normal?
3. Inferentially: is it statistically distributed differently than normal?

https://www.katemarshallmaths.com/chapter-3-bell-shaped-distribution.html

STATROULETTE

- A roulette wheel returns values between 0 and 36.
- Let’s simulate a game of roulette in Excel

DEMO

- Simulate 500 rounds of a roulette spin.
- Plot the resulting frequency distribution.

DEMO

- Simulate 500 rounds of a roulette spin using `RANDBETWEEN(0,36)`
- Plot the resulting frequency distribution using a histogram
- Hit `F9` while in your workbook. What happens?

DEMO

- Now, simulate a roulette spin 100 times.
- Take the average spin.
- Do this 500 times.
- Plot the distribution of sample means

MAGIC… OR STATISTICS?

- Central limit theorem: the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, *if the sample size is large enough.*

How “large enough” is large enough?

- N = 30? 60? 100?
- It depends on how “normal” your sample is

MEET ME IN THE MIDDLE

- A roulette wheel returns values between 0 and 36.
- What is the average roulette spin given more and more spins?

DEMO

- large-numbers.xlsx

MAGIC... OR STATISTICS?

- Law of large numbers: the average of results obtained from trials become closer to the expected value as more trials are performed

QUESTIONS?

2. FOUNDATIONS OF INFERENTIAL STATISTICS

HOUSES WITH AIR CONDITIONING ARE BETTER THAN THOSE WITHOUT

HOUSES WITH AIR CONDITIONING ~~ARE BETTER~~ HAVE HIGHER SALES VALUES THAN THOSE WITHOUT

HOUSES WITH AIR CONDITIONING

~~ARE BETTER~~ HAVE HIGHER SALES

VALUES THAN THOSE WITHOUT,

ON AVERAGE

HOUSES WITH AIR CONDITIONING HAVE HIGHER SALES VALUES THAN THOSE WITHOUT, *ON AVERAGE, GIVEN THE SAMPLES WE HAVE*

THE AVERAGE VALUE OF ONE CATEGORY IS HIGHER THAN ANOTHER

(*Independent samples t-test*)

DEMO

- File: `housing-foundations.xlsx`
- Is there a difference in sales price for homes with air conditioning (`airco`)?
  - Find the descriptive statistics and plot the histograms for each category

DRILL

- File: `housing-foundations.xlsx`
- Is there a difference in sales price for homes with a full, finished basement (`fullbase`)?
- Save your work – you'll "build on" it later!

We have the data… just not *all* of the data

- There is a difference across 546 homes…
- … but what about *all* homes?
- Is our sample *reflective* of their populations?

… but that doesn’t stop your phlebotomist!

- Or your pollster
- Or your QA manager
- Or your real estate analyst 😉

GETTING RIGOROUS WITH A HYPOTHESIS

- Testable
- Falsifiable
- Burden of proof on the claimant

GETTING RIGOROUS WITH A HYPOTHESIS

$$H_o: \mu_1 - \mu_2 = 0$$

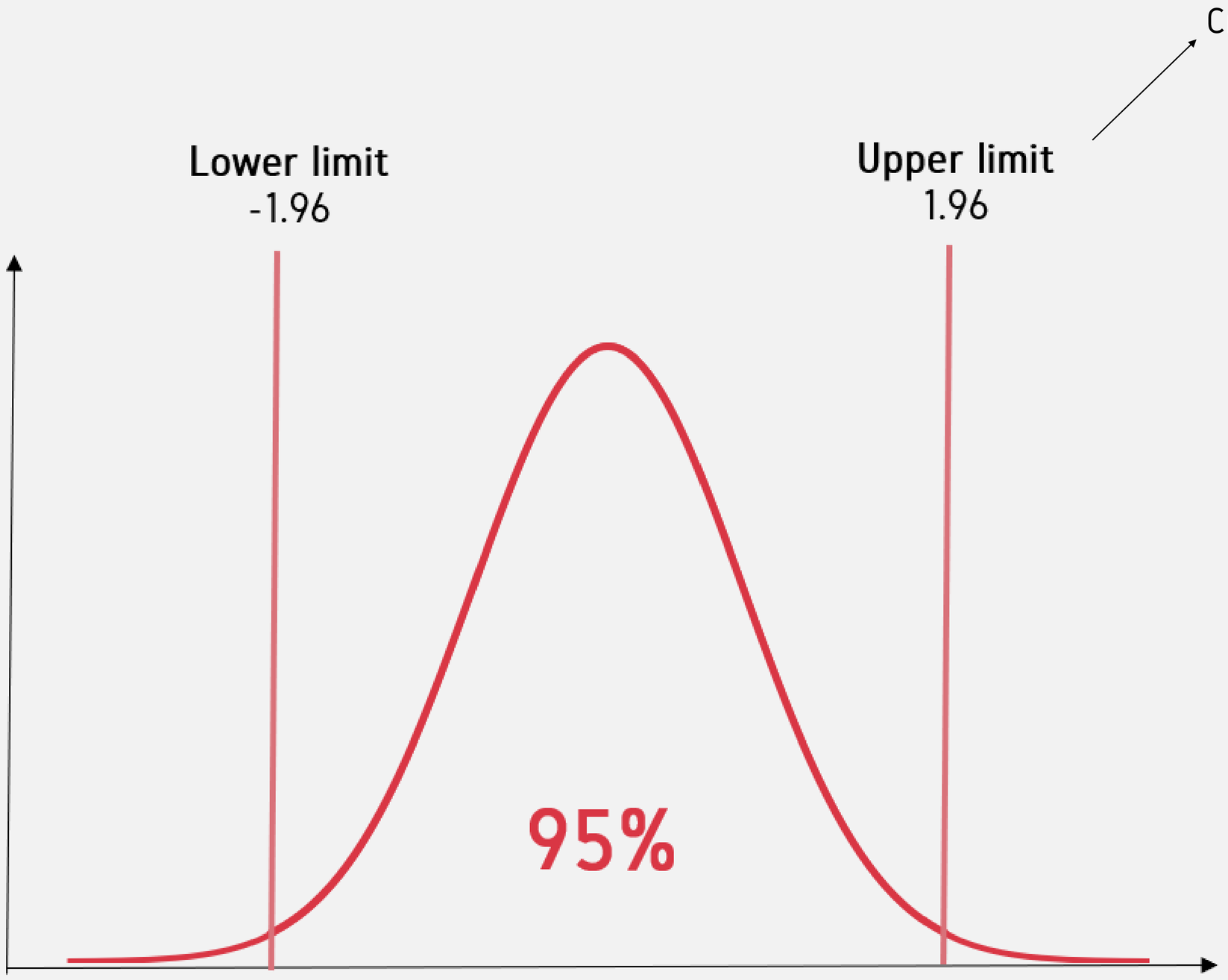
Null hypothesis: no difference in population means

$$H_a: \mu_1 - \mu_2 \neq 0$$

Alternative hypothesis: some difference in population means

… but we don't *know* the population means…

- Central limit theorem to the rescue!
- *Infer* the population mean (which is normally distributed) given a sample mean…
- … Some uncertainty required.
  - Usually 5%


Critical value
Lower limit
-1.96
Upper limit
1.96
95%

EXPLICIT WARNING: MATH AHEAD

Calculating the test statistic

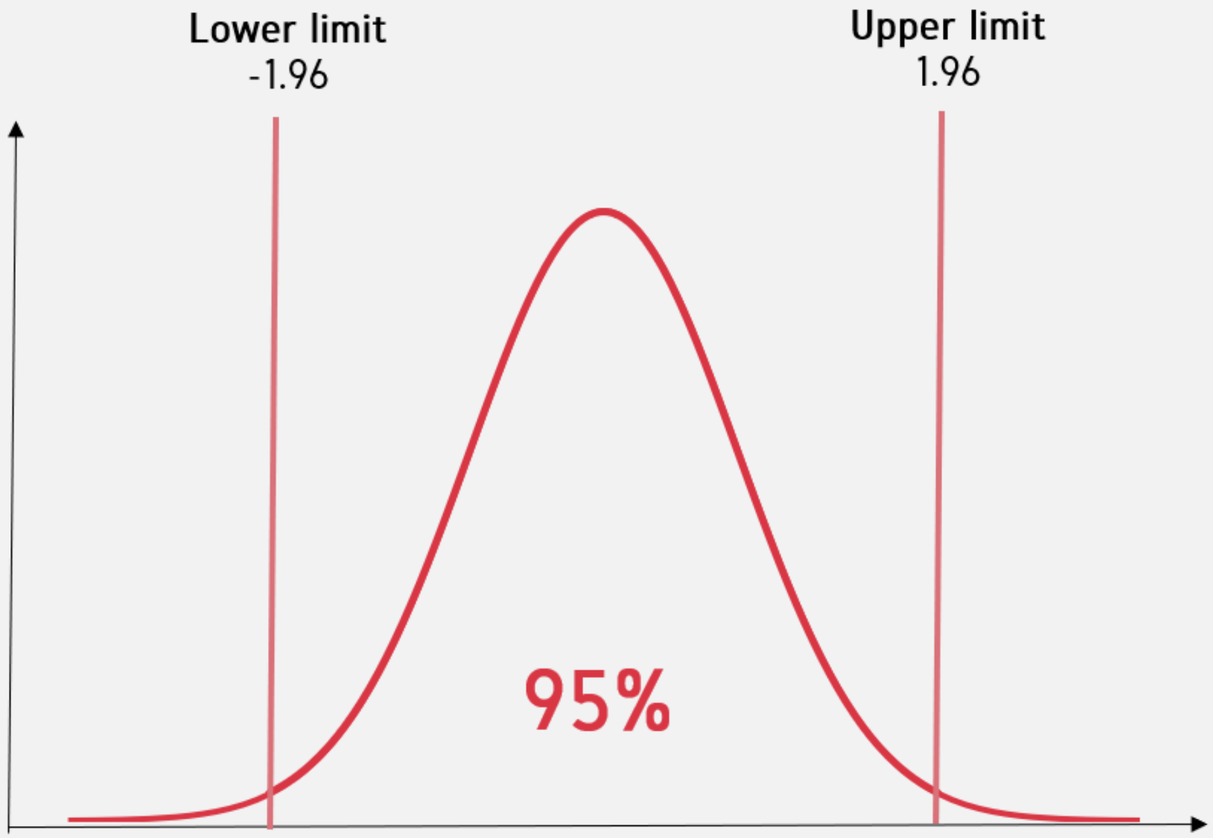
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}$ = sample mean

s = sample standard deviation

n = sample size

Does the test score exceed the critical value?



- **Yes:** Reject the null. There is likely to be a significant difference in means.
- **No:** Fail to reject the null. There is unlikely to be a significant difference in means.

Two-tail t-test

In statistics, nothing is proven, everything is falsifiable.

QUESTIONS?

3. T-TESTS FOR BUSINESS IMPACT

WARM-UP

- We want to test for a significant difference in sale price between houses with and without air conditioning.
- What are our hypotheses?

WARM-UP

- We want to test for a significant difference in sale price between houses with and without air conditioning.
- What are our hypotheses?

$$H_o: \mu_1 - \mu_2 = 0$$

Null hypothesis: there is no difference in the average sale price of homes with and without air conditioning.

$$H_a: \mu_1 - \mu_2 \neq 0$$

Alternative hypothesis: there is a difference in the average sale price of homes with and without air conditioning.

DEMO

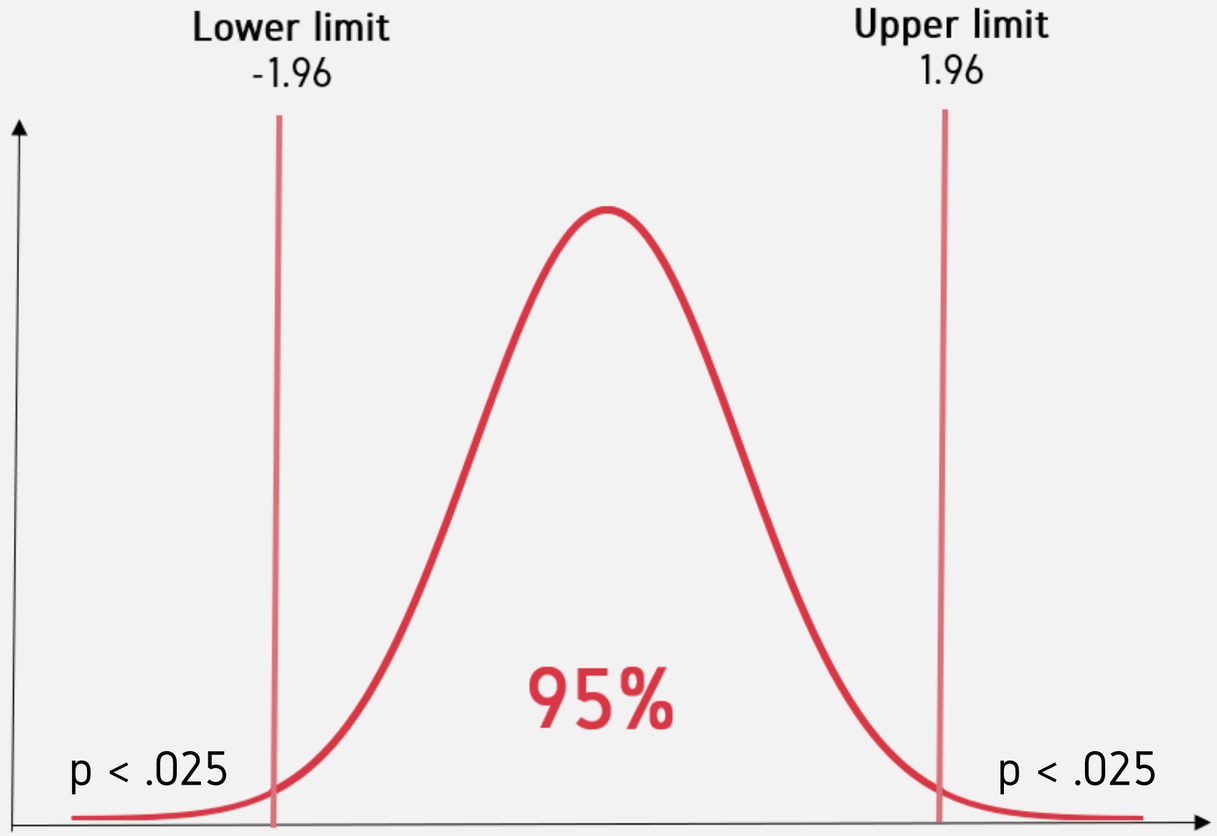
- File: `housing-inferential.xlsx`
- Is there a significant difference in sales price for homes with air conditioning?

DRILL

- File: `housing-inferential.xlsx`
- Is there a significant difference in sales price for homes with a full, finished basement?
- Save your results for later!

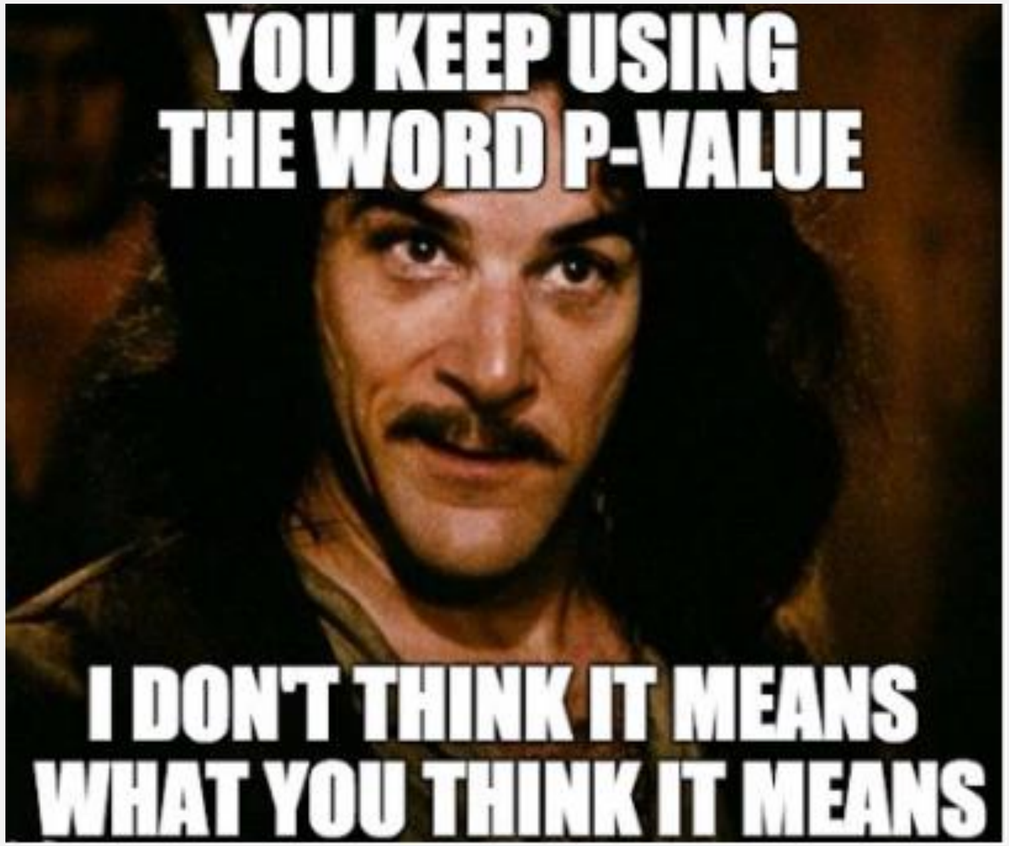
DEMO -- SUMMARY

1. Two-tailed independent samples t-test
2. 95% confidence = critical value of 1.96 = p-value of .05



A p-value is NOT a probability of making a mistake.

A p-value is the probability of obtaining an effect at least as extreme as the one in your sample data, *assuming the null hypothesis is true*.



SOLD: $200,000

SOLD: $200,000

If this were true in the population, you'd have found this in < .005% of samples due to random error.

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	85880.5896	59884.85
Variance	810167352.2	4.55E+08
Observations	173	373
Hypothesized Mean Difference	0	
df	265	
t Stat	10.69882732	
P(T<=t) one-tail	9.6667E-23	
t Critical one-tail	1.650623976	
P(T<=t) two-tail	1.93334E-22	
t Critical two-tail	1.968956281	

INTERPRETING P-VALUES

p-value	Interpretation
.02	Assuming the null is true, you would obtain the observed difference or more in 2% of samples due to random error.
.085	
.5	
.005	

INTERPRETING P-VALUES

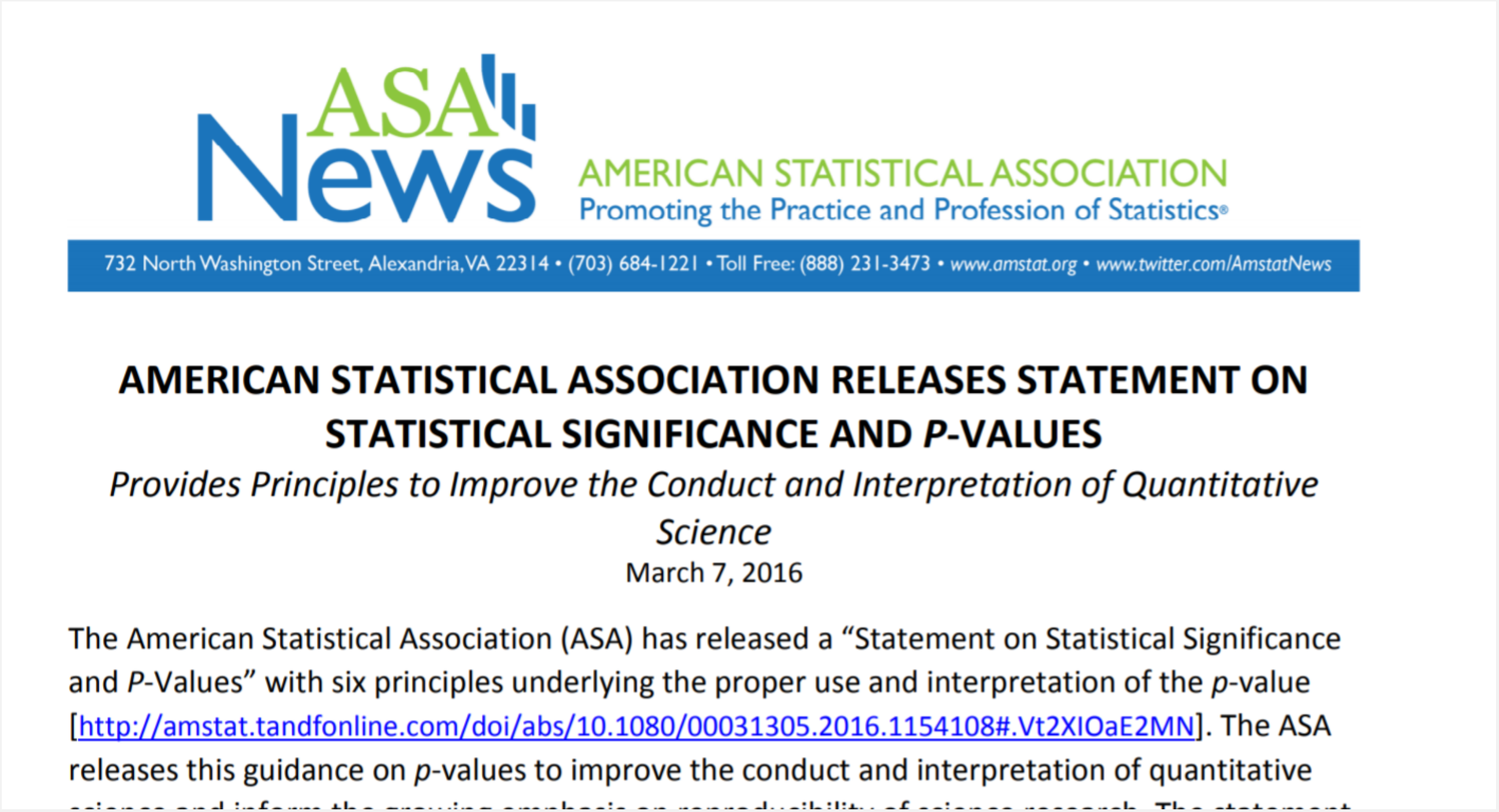
p-value	Interpretation
.02	Assuming the null is true, you would obtain the observed difference or more in 2% of samples due to random error.
.085	Assuming the null is true, you would obtain the observed difference or more in 8.5% of samples due to random error.
.5	Assuming the null is true, you would obtain the observed difference or more in 50% of samples due to random error.
.005	Assuming the null is true, you would obtain the observed difference or more in .05% of samples due to random error.

A p-value is NOT a substantive measure.

A p-value does not tell you *how much* of a difference there is, only that there is likely to be a difference.

The confidence interval will give us a substantive measure.

ONE STAT ~~TO RULE THEM ALL~~



ASA News

AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative

https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

Calculating the confidence interval

$$c.i. = (\overline{X_1} - \overline{X_2}) \mp Za_{/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Point estimate (difference in sample means)

Margin of error (Critical value * standard error of difference)

DEMO

- File: `housing-ci.xlsx`
- Calculate the confidence interval for difference in prices for homes with air conditioning

DRILL

- File: `housing-ci.xlsx`
- Calculate the confidence interval for difference in prices for homes with a full, finished basement

DEMO

- File: `margin-of-error.xlsx`
- Why do pollsters report a “margin of error?”
  - Why is it usually 2-3%?

PRESENTING T-TEST RESULTS FOR BUSINESS IMPACT

- Methods are not stories
- Translate to the universal denominator … $
  - “The results were significant at p < .005” or
  - “With 95% confidence, a house with AC sells for between $6K and $15K more.”
- When in doubt, visualize it

DEMO

- File: `housing-viz.xlsx`
- Visualize the difference in average sales price for homes with air conditioning

DRILL

- File: `housing-viz.xlsx` (continued)
- Statistically test and visualize the difference in average sales price for homes with gas for hot water heating, `gashw` **(Note: new variable!)**

It’s your world, the data is just living in it

t-Test: Two-Sample Assuming Unequal Variances

	yes	*no*
Mean	79428	67579.06334
Variance	923472100	698250450.3
Observations	25	521
Hypothesized Mean Difference	0	
df	26	
t Stat	1.915131244	
P(T<=t) one-tail	0.033268787	
t Critical one-tail	1.70561792	
P(T<=t) two-tail	0.066537575	
t Critical two-tail	2.055529439	

total sample size	546
mean difference	11848.93666
standard error of difference	6187.010263
Margin of error	12717.58173
Lower limit	-868.645073
Upper limit	24566.51839

- A p-value has wiggle room
- A confidence interval communicates what’s at stake
- A larger sample size can lead to more robust results

QUESTIONS?

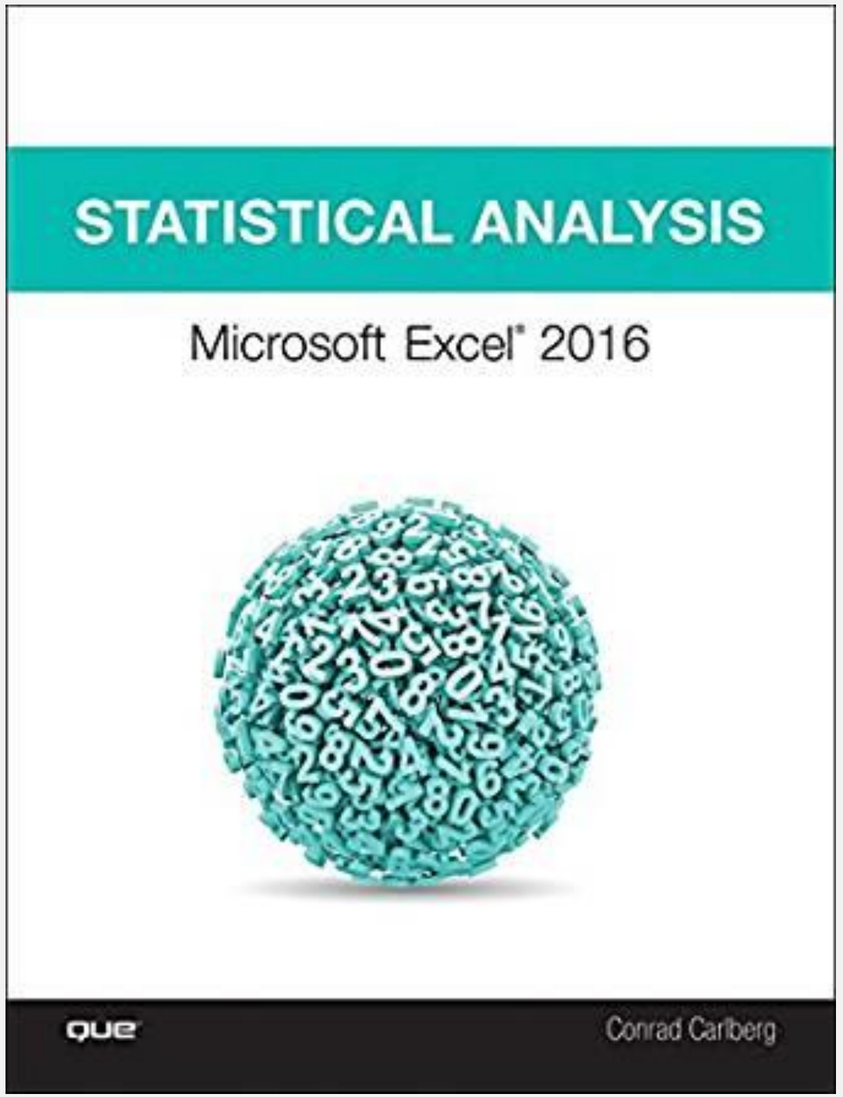
4. CONCLUSION

Future learning

- Check the `4-conclusion` folder for additional t-test practice
- Experimental design & A/B testing
- Testing multiple categories/non-normal categories/the same observations at multiple time points
- Linear regression

Statistical Analysis: Microsoft Excel 2016, by Conrad Carlberg

- On O'Reilly Learning at https://learning.oreilly.com/library/view/statistical-analysis-microsoft/9780134840437/



Data Smart: Using Data Science to Transform Information into Insight, by John Foreman

- On O'Reilly Learning at https://learning.oreilly.com/library/view/data-smart-using/9781118661468/

STRINGFEST ANALYTICS

LET'S TALK

LINKEDIN

linkedin.com/in/gjmount

EMAIL ADDRESS

george@stringfestanalytics.com

WEBSITE

stringfestanalytics.com

GITHUB

github.com/summerofgeorge

QUESTIONS?