GENERAL ASSEMBLY

STATISTICS FUNDAMENTALS - PART II

LEARNING OBJECTIVES

- Correlation and Causation
- Linear Regression with statsmodels
- Hypothesis Testing, P-values and Confidence intervals

LESSON #4

PRE WORK & REVIEW

LAST LESSON REVIEW

- Git
- Basic stats: Mean, STD, Correlation
- Normal distribution
- Skewness, box-cox transformation
- Categorical Data, one-hot-encoding

QUESTIONS?

ANY QUESTIONS FROM LAST CLASS?

QUESTIONS FROM EXIT TICKET

STATISTICS FUNDAMENTALS - PART II

SO MANY PYTHON SCIENTIFIC LIBRARIES

PYTHON LIBRARIES

Data formats

- Numpy: array manipulation
- Pandas: Dataframe manipulation

Models:

- Statsmodel: statistically-oriented approaches to data analysis, with an emphasis on econometric analyses
- Scipy: The ancestor commits
- Scikit: machine learning. name stems from the notion that it is a "SciKit" (SciPy Toolkit), a separately-developed and distributed third-party extension to SciPy. commits

and more to come

- NLTK
- Gensim
- scikit-image

  ....

BEFORE WE BEGIN

ALLEN DOWNEY

ALLEN DOWNEY

THINK STATS

CORRELATION - CAUSATION

CAUSATION AND CORRELATION

Which is true?

X and Y are very correlated

- Then X causes Y or Y causes X
- Can't say!

CAUSATION AND CORRELATION

Not the same thing

- Calculating Correlation: easy
- Demonstrating and Quantifying Causation: Causal Inference: Not so easy

Examples of correlation obviously without causation http://www.tylervigen.com/spurious-correlations

CAUSALITY

CAUSAL INFERENCE

So how do you identify causality (in presence of correlation)?

http://egap.org/methods-guides/10-strategies-figuring-out-if-x-caused-y

=> However most common strategy is to find not causality but correlation through linear regression.

Statistics and economics usually employ pre-existing data or experimental data to infer causality by **regression methods**.

Works under **VERY** strong assumptions

LINEAR REGRESSION

LINEAR REGRESSION

- Find the best line that fits the samples

NOTEBOOK 1: LINEAR REGRESSION

Local Online

LINEAR REGRESSION

Anscombe's quartet https://en.wikipedia.org/wiki/Anscombe%27s_quartet

MULTILINEAR REGRESSION

see Notebook Local / Online

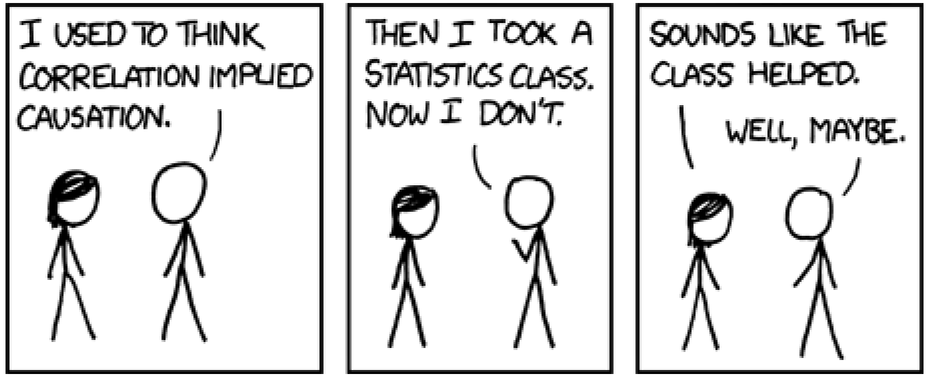
BACK TO CAUSALITY

REGRESSION AND CAUSATION:

For regression coefficients to have a causal interpretation we need both that

- the linear regression assumptions hold: linearity, normality, independence, homoskedasticity
- and that all confounders of, e.g., the relationship between treatment A and Y be in the model.

XKCD


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

LINEAR REGRESSION ASSUMPTIONS

Linearity: the relationship between the covariates and target to be linear test with scatter plots

Normality: Normal distribution QQ plot

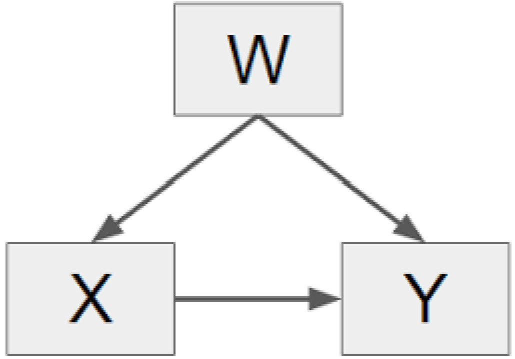
Independence: no or little multicollinearity between variables Correlation matrix

Homoscedasticity: for a given variable the low and high range have the same statistical properties Chunk data and Check Variance

Confounders

CONFOUNDING

External (hidden, unknown) variable which influences X and Y influences their correlation.



see http://www.statisticshowto.com/confounding-variable/

CONFOUNDING

Ex:

- Relationship between **ice-cream consumption** and number of **drowning deaths** for a given period

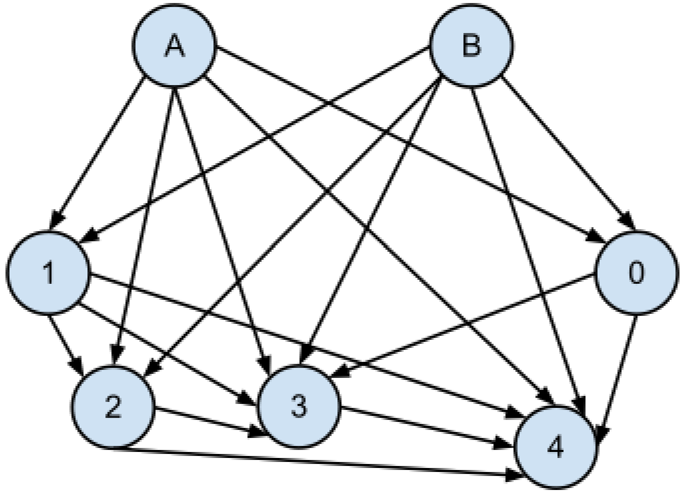
```
Confounding: ?
```

- Relation between **Exercise** and **Weight loss**

```
Confounding: ?
```

DAGS TO EXPLORE CONFOUNDERS

DIRECTED ACYCLIC GRAPH



Read more about it:

- Directed Acyclic Graphs (DAGs)
- DAGs - wikipedia

CONFOUNDER INFLUENCE

Notebook Confounder influence http://www.r-bloggers.com/how-to-create-confounders-with-regression-a-lesson-from-causal-inference/

STRATEGIES FOR CAUSATION: EXPERIMENT DESIGN

IN THE REAL WORLD

- Bias can be eliminated with random samples.
- Introduce control variables (keep the variable constant) to control for confounding variables. For example, you could control for age by only measuring 30 year olds.
- Within subjects designs test the same subjects each time. Anything could happen to the test subject in the “between” period so this doesn’t make for perfect immunity from confounding variables.
- Counterbalancing can be used if you have paired designs. In counterbalancing, half of the group is measured under condition 1 and half is measured under condition 2.

LINEAR REGRESSION FOR CAUSAL INTERPRETATION

NOTEBOOK 2: LINEAR REGRESSION FOR CAUSAL INTERPRETATION

Run regression sales ~ TV

LINEAR REGRESSION

Another excellent notebook on linear regression on the DataRobot blog

Datarobot blog post on linear regression Notebook on linear regression

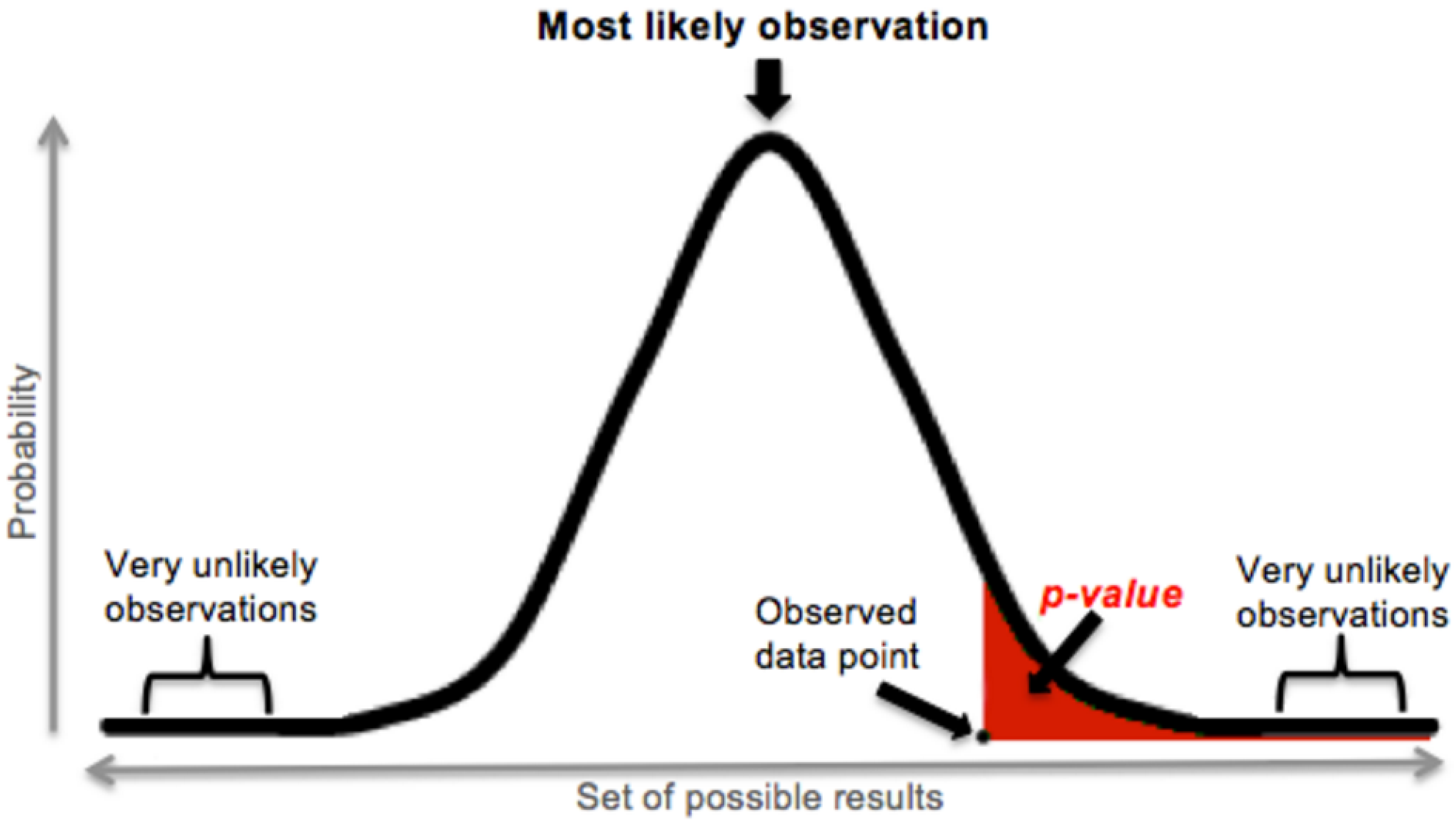
HYPOTHESIS TESTING

HYPOTHESIS TESTING

A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data

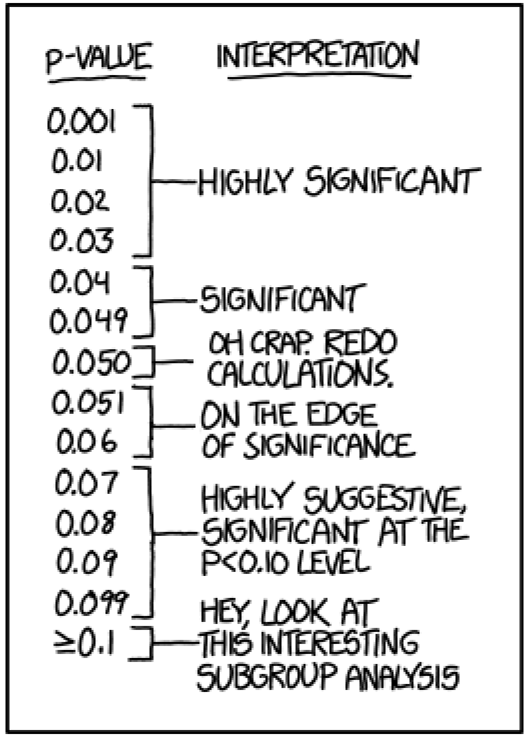
Helps determine if the result is a fluke due to sampling for instance or a real thing sampling errors

P-VALUE



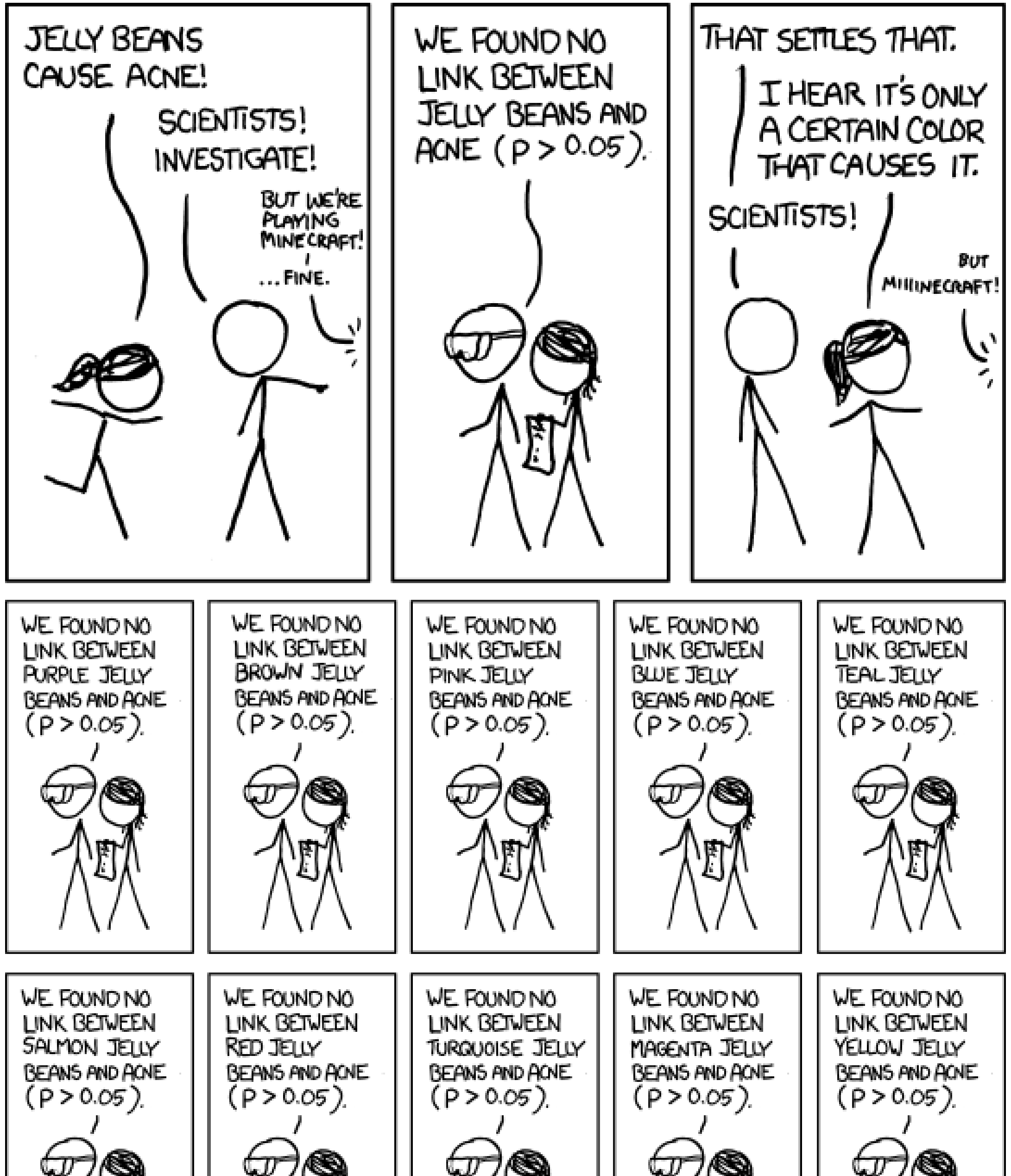
A *p-value* (shaded red area) is the probability of an observed (or more extreme) result arising by chance

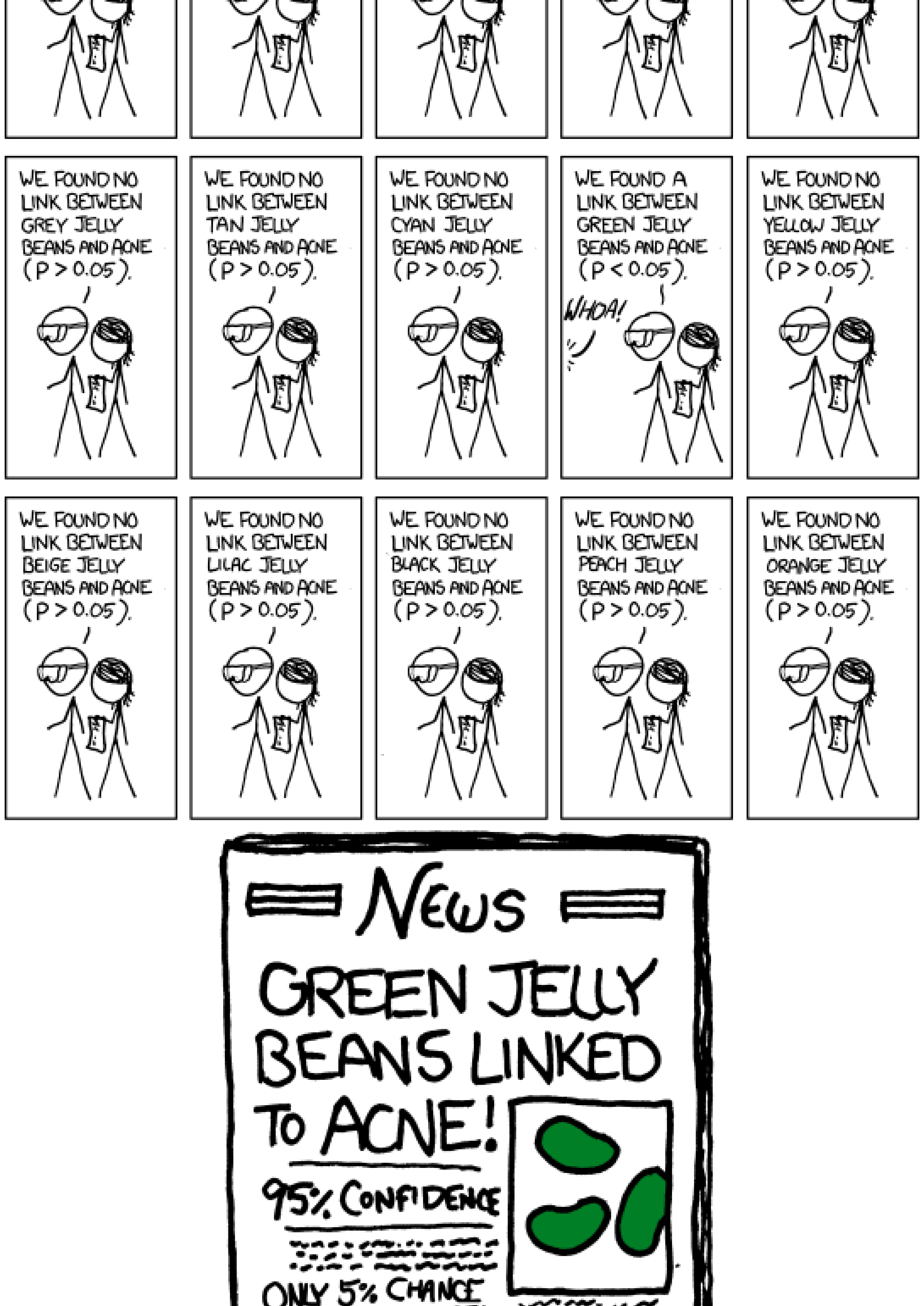
XKCD ON P-VALUES

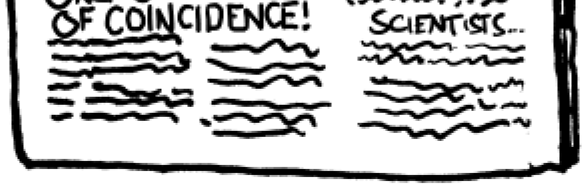


P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

XKCD ON P-VALUES


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE


OF COINCIDENCE!
SCIENTISTS...

CONFIDENCE INTERVALS

Conf intervals

FINAL PROJECT PART 1 LIGHTNING TALK

FINAL PROJECT

Final Project Part 1 Lightning Talk

YOUR TURN

YOUR TURN

https://github.com/alexperrier/ds-curriculum/blob/master/lessons/lesson-04/code/starter-code/lab-starter-code-4.ipynb

LESSON REVIEW

BEFORE NEXT CLASS

5 QUESTIONS ABOUT TODAY

EXIT TICKET

EXIT TICKET

- Really good. read this if you read one thing An Introduction to Causal Inference
- Datarobot Notebook on linear regression
- Correlation / Association is not causalation
- Assumptions of Linear Regression
- Regression diagnostics: testing the assumptions of linear regression
- Do your data violate linear regression assumptions?
- Confounding
- http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html
- http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/ols.html
- https://github.com/statsmodels/statsmodels/blob/master/examples/notebooks/ols.ipynb
- http://statisticalhorizons.com/prediction-vs-causation-in-regression-analysis