What you need to do before the next class

- Join slack and look out for messages on all channels: bmb961-statgaps-nov18.slack.com
- Read the course website: bit.ly/bmb961-nov18
- Fill out the incoming survey: bit.ly/bmb961-nov18-incoming

Topic 1: Statistical hypothesis testing

Lectures 2 & 3

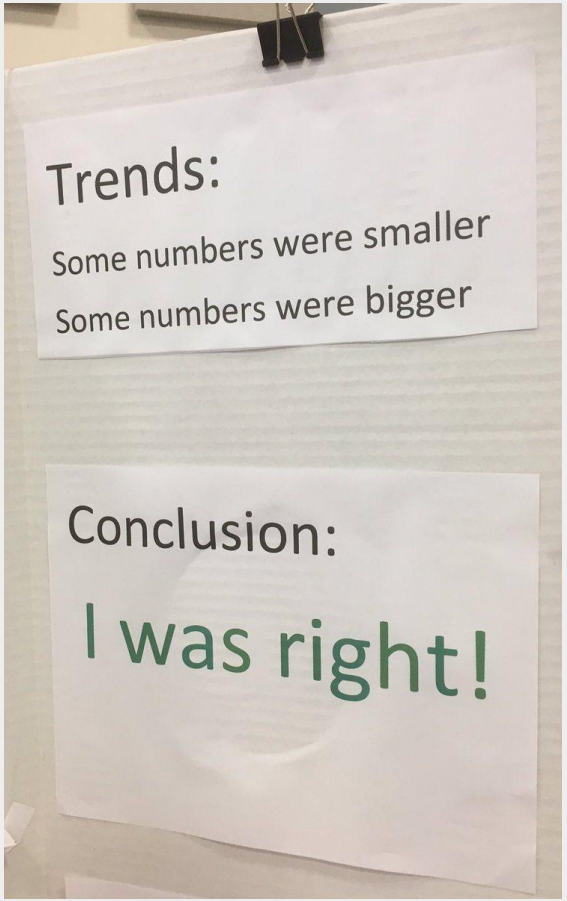
- P-value & P-hacking
- Multiple hypothesis correction
- Estimation of error & uncertainty

Topic 1: Statistical hypothesis testing

Lectures 2 & 3

The first principle is that you must not fool yourself, and you are the easiest person to fool.

– Richard Feynman



Statistical hypothesis testing

- Many scientific studies are interested in quantifying the difference in a particular parameter between two groups.
  - There's always some difference → Is it statistically significant difference?

- Say you're testing the efficacy of a cold medicine:
  - Two groups given placebo/medication
  - Followed-up: how long the cold lasted in each person in both groups
  - Null: Ineffective; Alternative: Effective

P-value

- The next step is to perform a statistical hypothesis test and get a **p-value**.
- The p-value is:
  - The amount of evidence that there is an effect?
  - The probability that the observed outcome is important?
  - The probability that the medication is ineffective?

The p-value is the probability that the experiment would have produced the observed outcome (or something more extreme) even if the medication were completely ineffective.

Write code to simulate two distributions and calculate p-values both using a t-test and a permutation test.

P-value

- P-values are dependent on:
  a. Size of the effect (effect size)
  b. Sample size
  c. Variance within each group
  d. The underlying experimental design & the null hypothesis (need always be random chance).
     - Conversely, two completely different experiments can give same data but end up very different p-values.
       - 3 out of 9: Binomial p-value = 0.073; Neg. Binomial p-value = 0.033.

P-value - History

- Fisher (1920s):
  - Informal method to help interpret the data along with prior experience, domain knowledge, size of the effect, etc.
- Neyman & Pearson:
  - Control false positive rate at α, set by the experimenter based on what can be tolerated.
  - Formulate null and alternative hypothesis.
  - Reject null when $p < \alpha$.
    - The threshold $\alpha = 0.05$ is merely a convention.

P-value

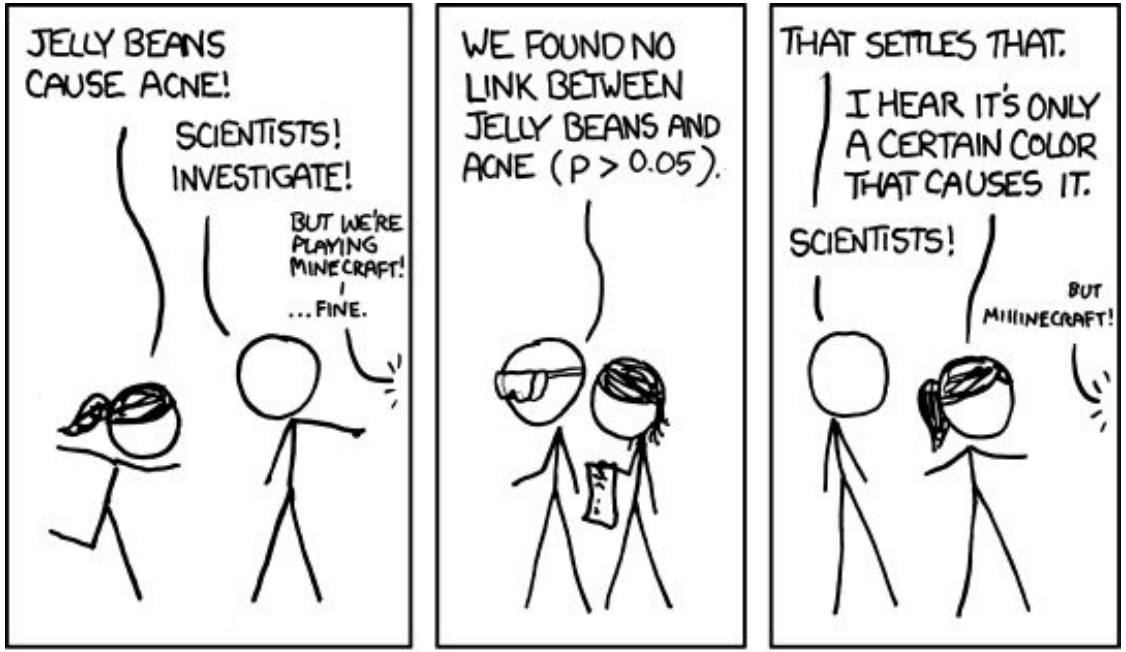
Significant or not!

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

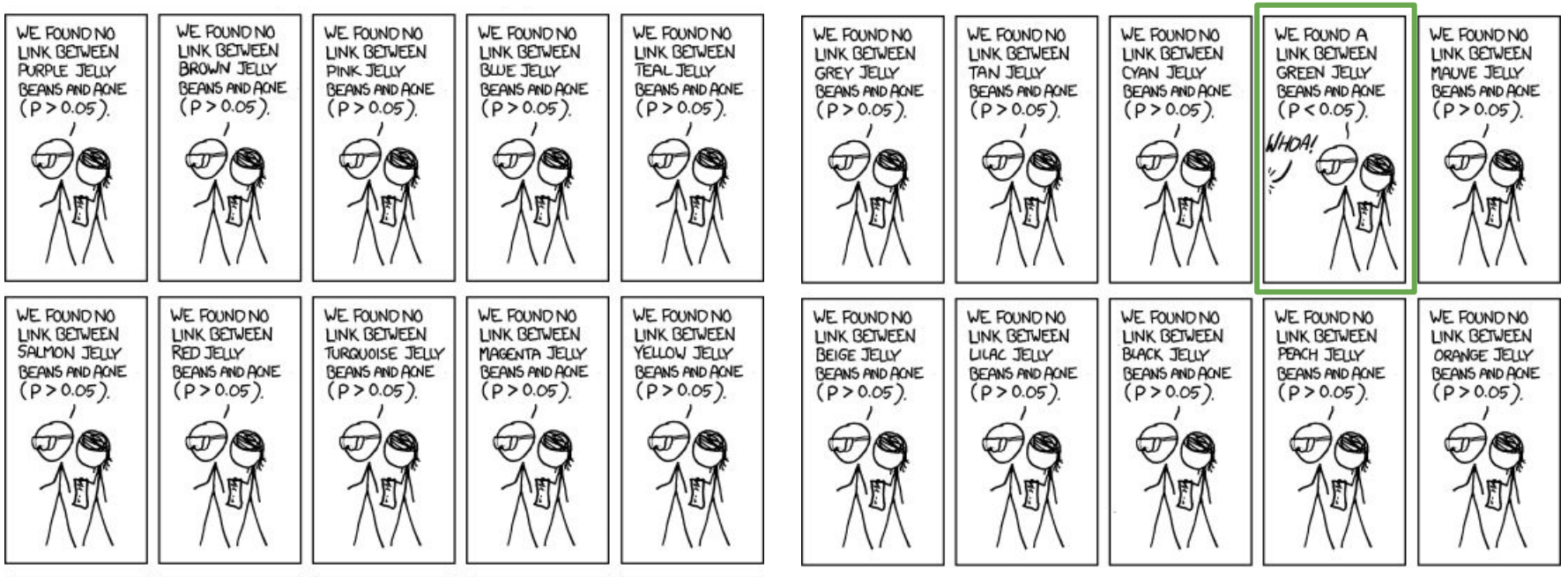
The following list is culled from peer-reviewed journal articles in which:
(a) the authors set themselves the threshold of 0.05 for significance,
(b) failed to achieve that threshold value for p and
(c) described it in such a way as to make it seem more interesting.

(barely) not statistically significant (p=0.052)
a barely detectable statistically significant difference (p=0.073)
a borderline significant trend (p=0.09)
a certain trend toward significance (p=0.08)
a clear tendency to significance (p=0.052)
a clear trend (p<0.09)
a clear, strong trend (p=0.09)
a considerable trend toward significance (p=0.069)
a decreasing trend (p=0.09)
a definite trend (p=0.08)
a distinct trend toward significance (p=0.07)
a favorable trend (p=0.09)

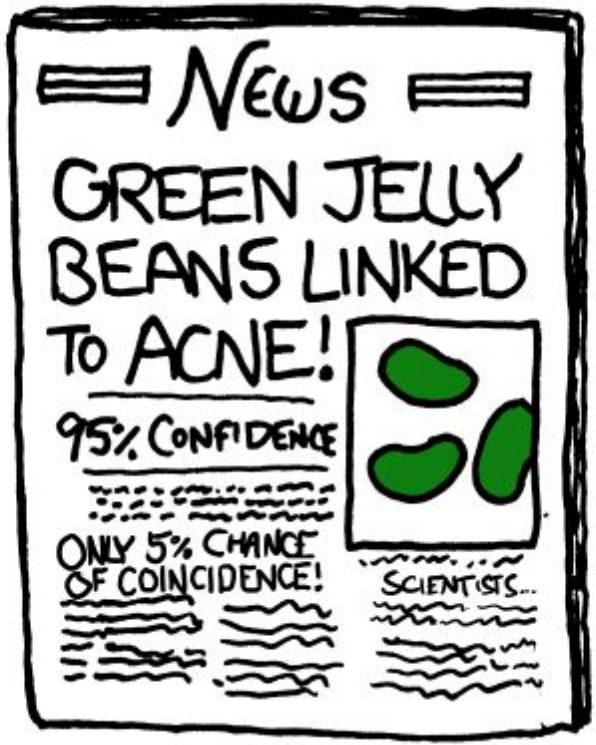
Multiple hypothesis testing



Multiple hypothesis testing



Multiple hypothesis testing



- The more inferences are made, the more likely erroneous inferences are to occur.
- Several statistical techniques have been developed to prevent this from happening.
- These techniques generally require a stricter significance threshold for individual comparisons, so as to compensate for the number of inferences being made.

What is the probability of obtaining at least 1 false positive? Family-wise error rate (FWER)

Type I & type II errors

P-value captures if there is “sufficient” inconsistency with the null hypothesis.

Choosing $p < \alpha$ controls type I error at α.

- Type I error: False-positive rate
- Type II error: False-negative rate
- Remember the story of the boy that cried wolf!

David Robinson
@drob

Follow

Remember, mixing up Type I and Type II errors is called a Type III error

David Robinson
@drob

Follow

Giving mistakes numbers instead of names was a real Type IV error

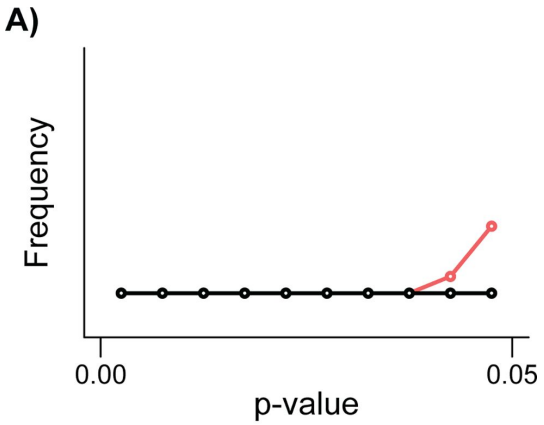
Multiple hypothesis testing

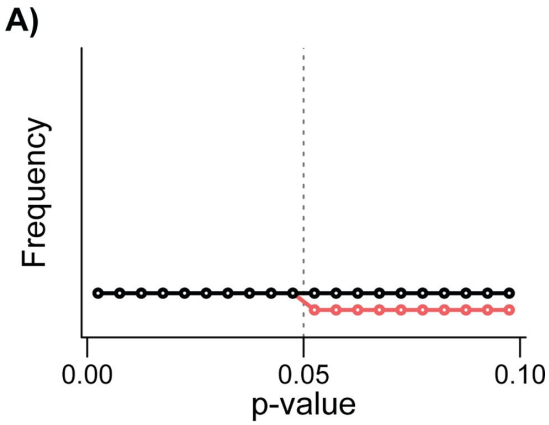
- FWER = Pr(#FP ≥ 1)
- False discovery rate (FDR) = E[#FP / #Discoveries]
- Suppose 550 out of 10,000 genes are found to have different expression levels between disease and control samples at $p < 0.05$.
    - If p-value is chosen to control FWER, what is the #FP?
    - If p-value is chosen to control FDR, what is the #FP?

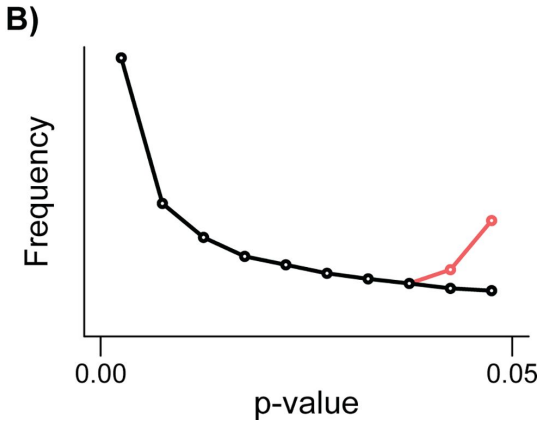
P-hacking and Publication bias

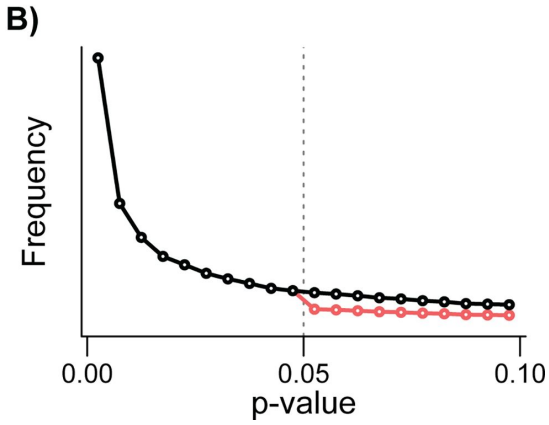
- P-hacking: Collect or select data or statistical analyses until nonsignificant results become significant.
    - Conducting analyses midway through experiments to decide whether to continue collecting data.
    - Recording many response variables and deciding which to report postanalysis
    - Deciding whether to include or drop outliers postanalyses
    - Excluding, combining, or splitting treatment groups postanalysis
    - Including or excluding covariates postanalysis, and
    - Stopping data exploration if an analysis yields a significant p-value.
- Publication bias: studies with nonsignificant results have lower publication rates.

P-hacking and Publication bias

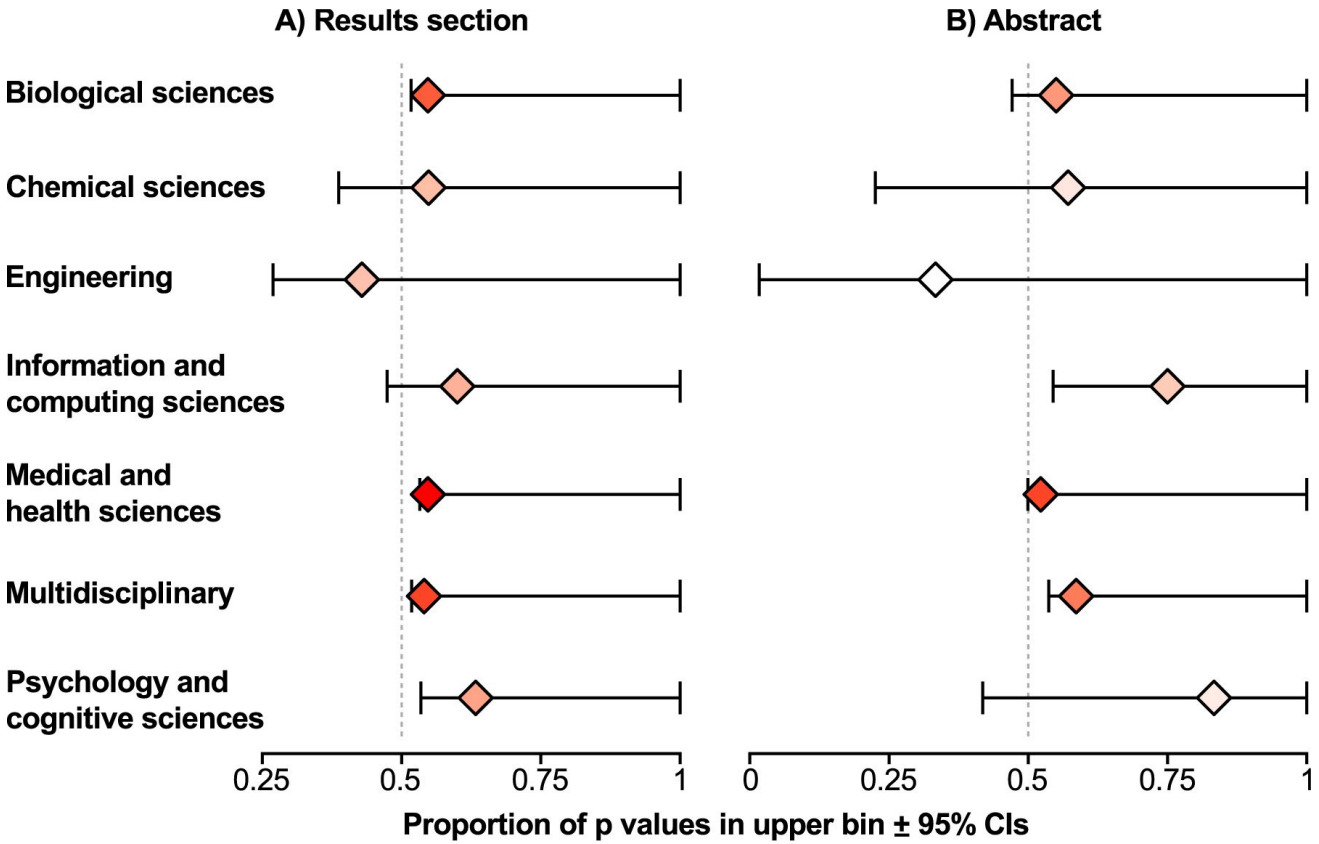




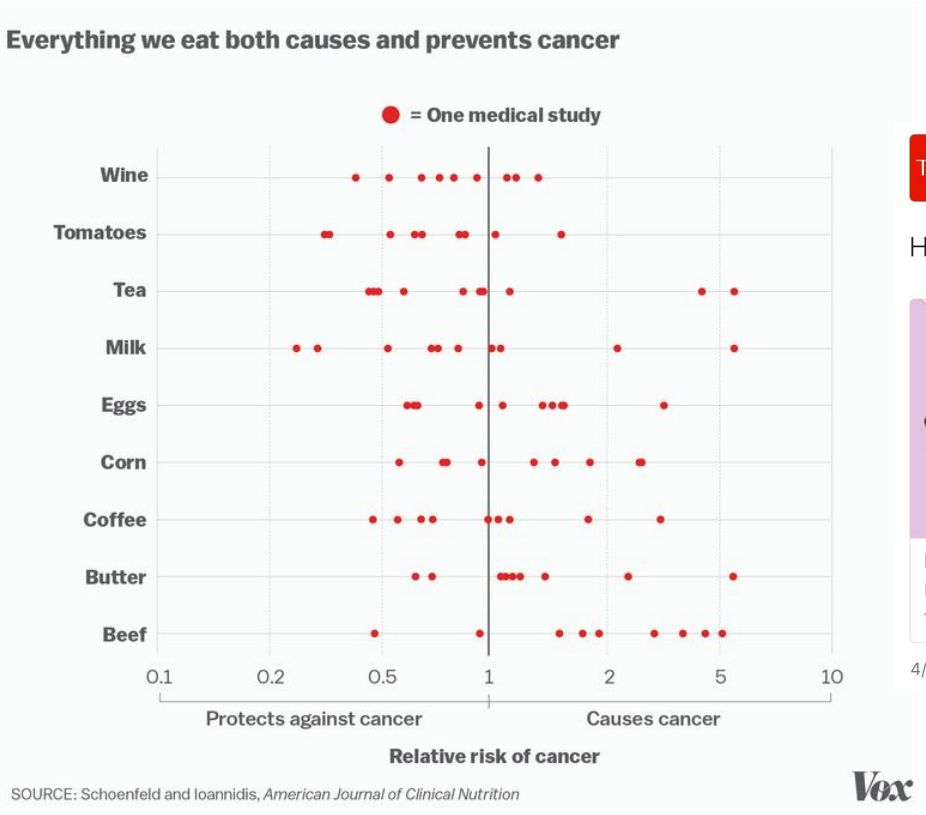




Publication bias and P-hacking



Statistical hypothesis testing



TIME
@TIME

How coffee can help you live longer

How Coffee Can Help You Live Longer
New findings add to growing evidence that co...
time.com

4/9/17, 6:45 AM

TIME
@TIME

The problem with your coffee

Hot Drinks a Probable Cancer Cause, Says WHO
time.com

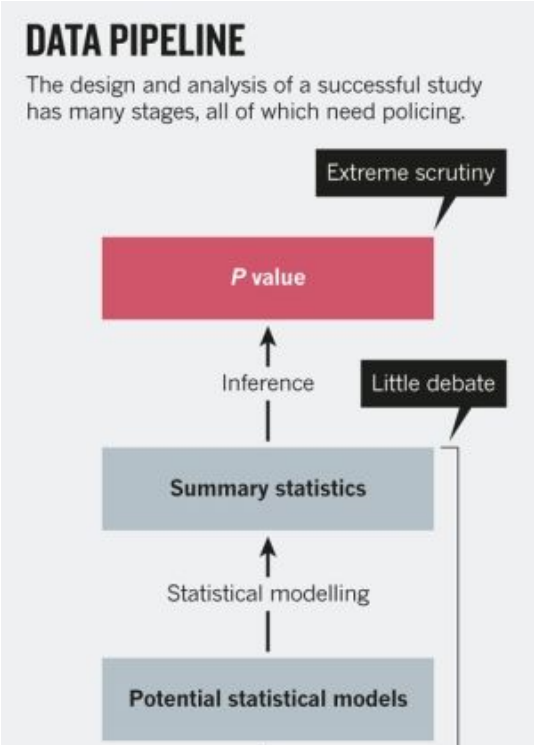
4/9/17, 6:15 AM

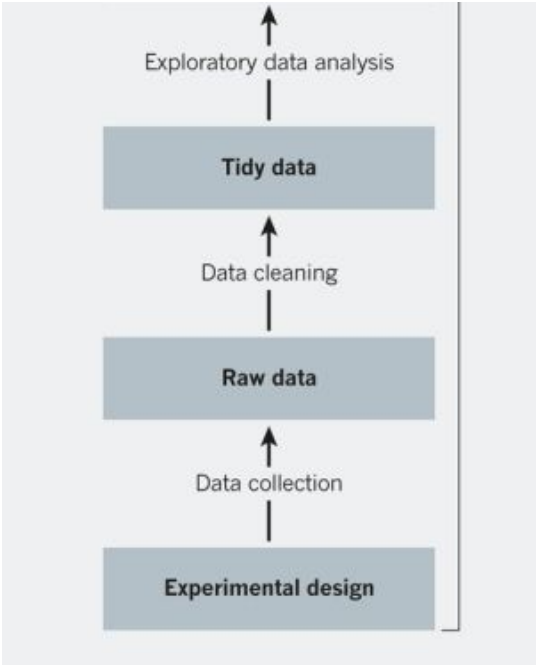
Questionable research practices

- Exclusively using p-values to determine the relevance and sanity of the results of a statistical test.
- Analyzing the data until the desired results are found.
- Collecting more data to reach smaller p-values.
- Trying many hypothesis until one of them gives a low p-value, and reporting just that final result.

P-value

P values are just the tip of the iceberg!

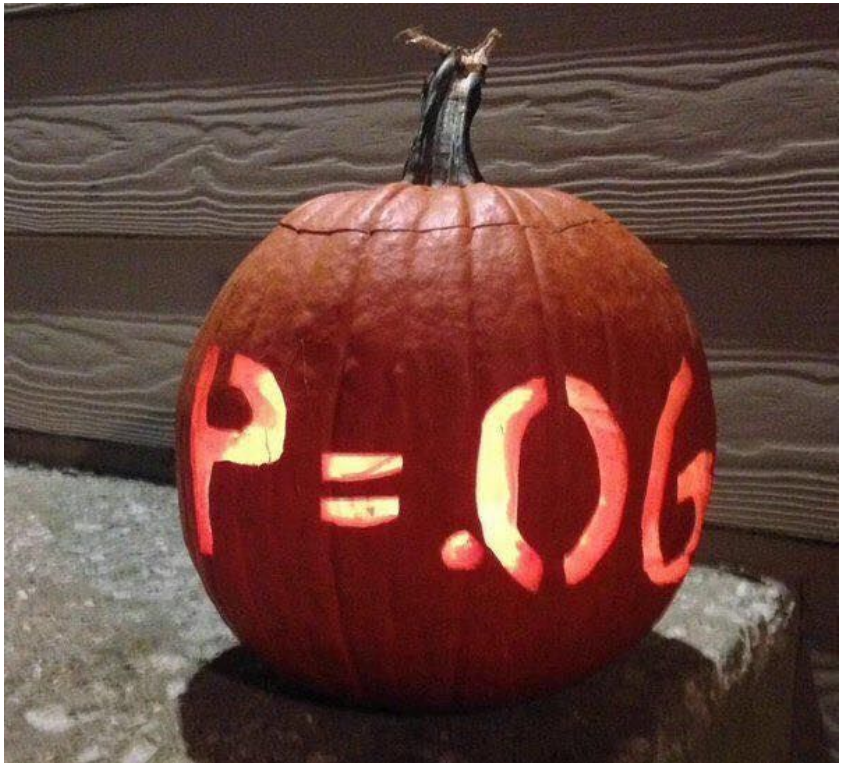




JT Leek, RD Peng http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412

P-value

The 'p' in p-value actually stands for p-otentially interesting!



ALTBIER - 4.9% ABV

The original amber ale as created by the Germans. Slightly drier than the American version, this beer drinks easy and satisfies the palette with notes of toffee and caramel without being thick or too dark which makes it a good idea.

P-VALUE

DRY-HOPPED AMERICAN PALE ALE - 5.4% ABV

This Pale Ale is light and hoppy with just the right amount of malt depth. This beer challenges the notion that hops and grain can't be balanced. Reject the null hypothesis.

SENSORY OVERLOAD

NEW ENGLAND IPA - 6.1% ABV

Sensory Overload doesn't let bitterness get in the way as your senses go into overdrive trying to keep up with the juicy citrus a