Day 03

P-hacking
Publication Bias
Multiple testing

- Hypothesis testing
- P-value
- P-hacking, Publication Bias
- Multiple hypothesis testing

Statistical hypothesis testing

Abstract

Formula display: MathJax

Background

Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation plays during normal cellular differentiation and tissue specification. We also need thorough, genome level studies to determine the meaning of methylation of individual CpG dinucleotides in terms of gene expression.

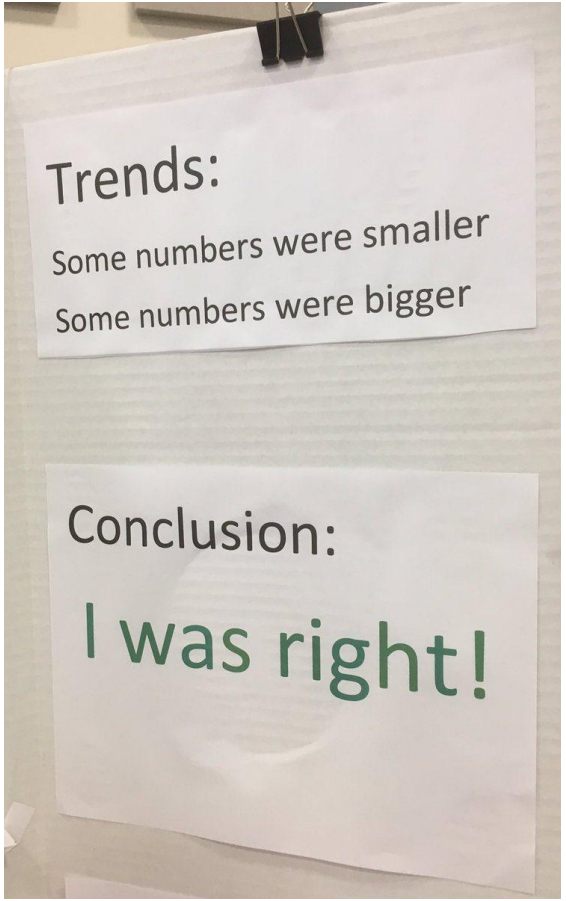
Results

In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arraysand compared those to gene expression by RNA sequencing. We have identified unique signatures of global DNA methylation for human heart, kidney and liver, and showed that DNA methylation data can be used to correctly classify various tissues. It indicates that DNA methylation reflects tissue specificity and may play an important role in tissue differentiation. The integrative analysis of methylation and RNA-Seq data showed that gene methylation and its transcriptional levels were comprehensively correlated. The location of methylation markers in terms of distance to transcription start site and CpG island showed no effects on the regulation of gene expression by DNA methylation in normal tissues.

Conclusions

This study showed that an integrative analysis of methylation array and RNA-Seq data can be utilized to discover the global regulation of gene expression by DNA methylation and suggests that DNA methylation plays an important role in normal tissue differentiation via modulation of gene expression.

https://nsaunders.files.wordpress.com/2012/07/bmcsysbiol.png



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

Statistical hypothesis testing

1. **Decide on the effect** that you are interested in, design a suitable experiment or study, pick a data summary function and test statistic.
2. **Set up a null hypothesis**, which is a simple, computationally tractable model of reality that lets you compute the null distribution, i.e., the possible outcomes of the test statistic and their probabilities under the assumption that the null hypothesis is true.
3. **Decide on the rejection region**, i.e., a subset of possible outcomes whose total probability is small.
4. **Do the experiment** and collect the data, compute the test statistic.
5. **Make a decision**: reject the null hypothesis – i.e. conclude that it is unlikely to be true – if the test statistic is in the rejection region.

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.

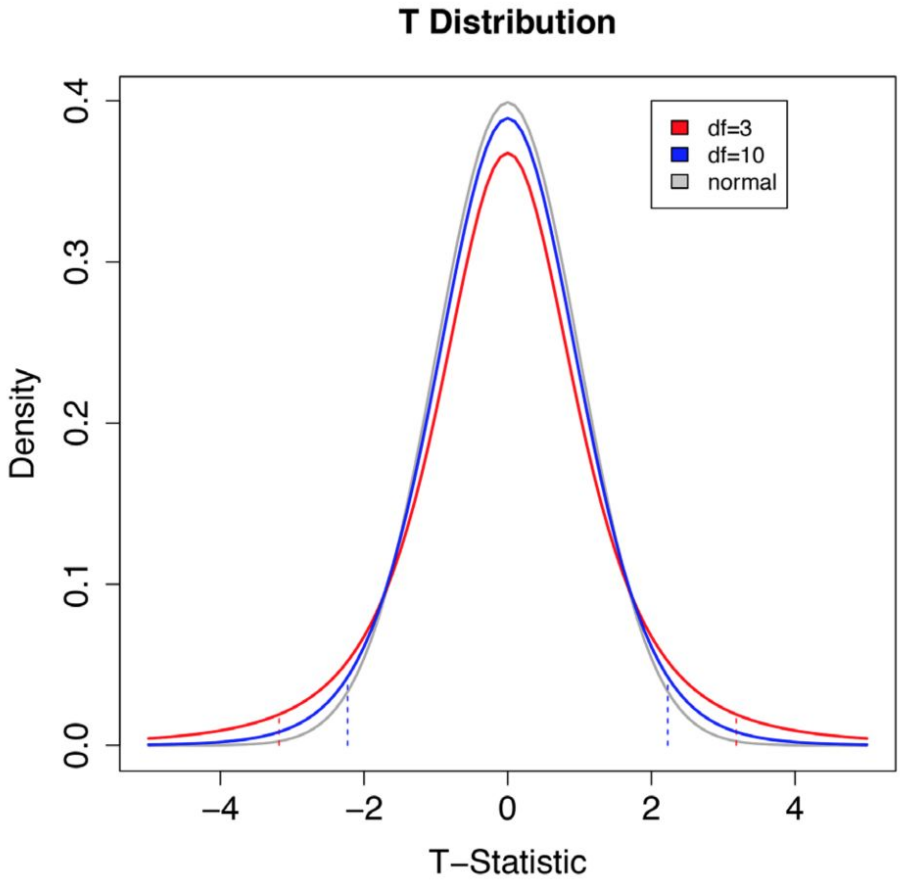
P-value

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.

1. Simulate data from two normal distributions: (mean, sd) → (0, 0.5) and (0.2, 0.5)
2. Effect = Difference in the mean of two distributions.
3. Repeat the following 10,000 times to set up the null hypothesis for this test statistic
   - Randomly assign data points to groups
   - Record the test statistic of the permuted groups
4. Calculate the p-value of the real test statistic.

```r
sample_size <- 20
effect_size <- 0.25
stddev <- 0.5

group1 <- rnorm(sample_size, mean = 0, sd = stddev)
group2 <- rnorm(sample_size, mean = effect_size, sd = stddev)

hist(group1, col = rgb(0, 0, 1, 0.25), breaks = seq(-2, 2, 0.25), xlim = c(-2, 2))
hist(group2, col = rgb(1, 0, 0, 0.25), breaks = seq(-2, 2, 0.25), add = T)

test_statistic <- mean(group2) - mean(group1)

#########################################################

permuted_test_statistics <- {}

for(i in 1:10000) {
  permuted_data <- sample(c(group1, group2))

  rand_group1 <- permuted_data[1:20]
  rand_group2 <- permuted_data[21:40]

  permuted_test_statistics <- c(permuted_test_statistics,
                                mean(rand_group2) - mean(rand_group1))
}

hist(permuted_test_statistics)
abline(v = test_statistic, col = "red", lwd = 2, lty = 2)
```

P-value



One-sample test

$$t = \frac{\bar{x} - \mu_0}{\text{SEM}}$$

Two-sample test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{N_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{N_2}(x_j - \bar{x}_2)^2}{N_1 + N_2 - 2}$$

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.

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 (α)
- Type II error: False-negative rate (β)
- 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

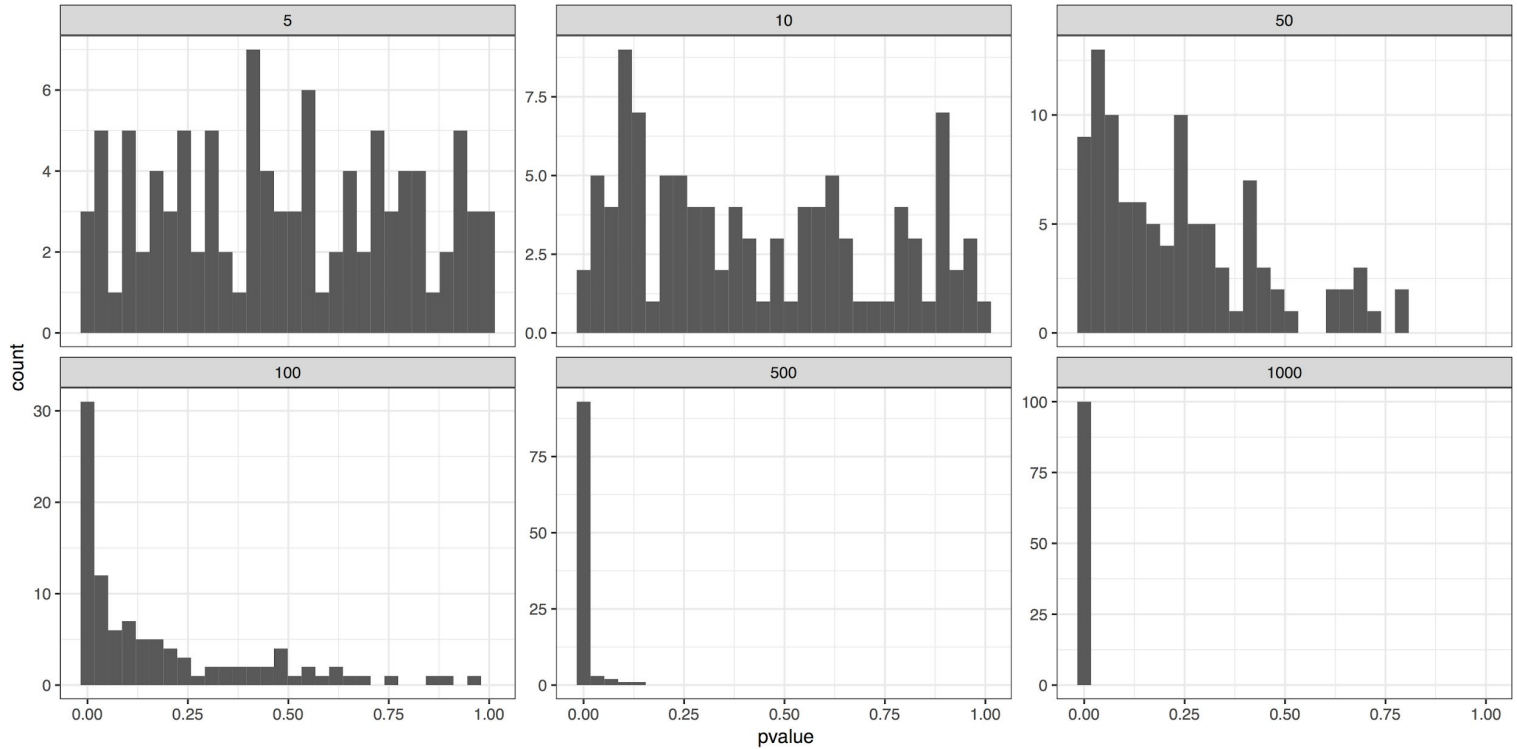
Giving mistakes numbers instead of names was a real Type IV error

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 not 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

- P-values are dependent on: sample_size (effect_size = 0.25, std_deviation = 1)



P-value

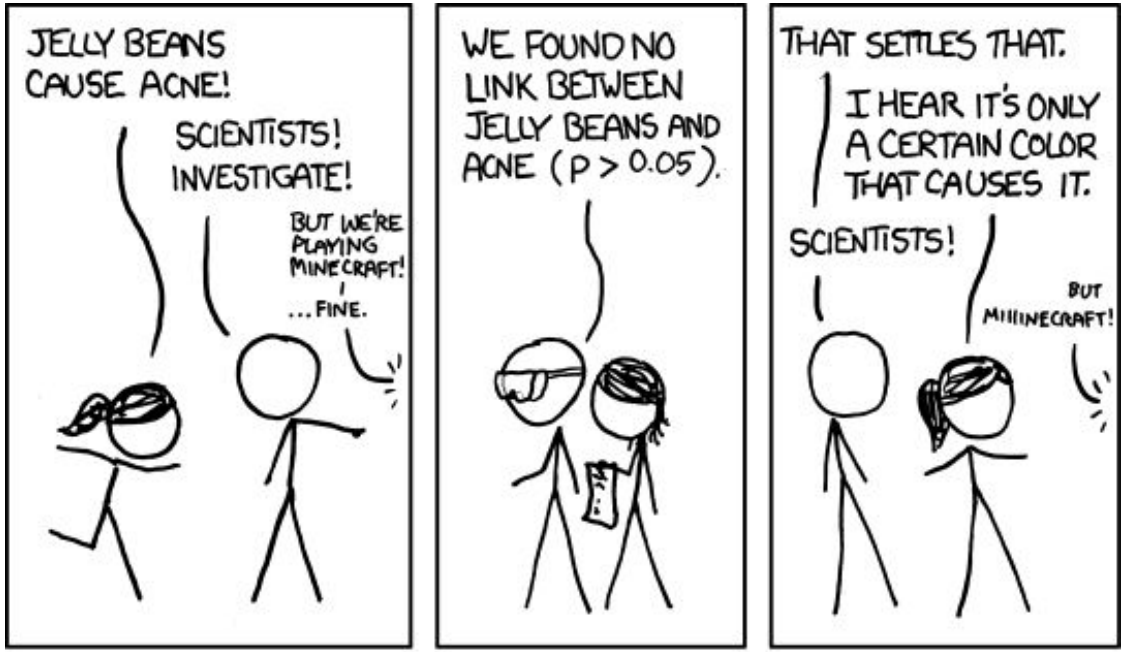
Significant or not!

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

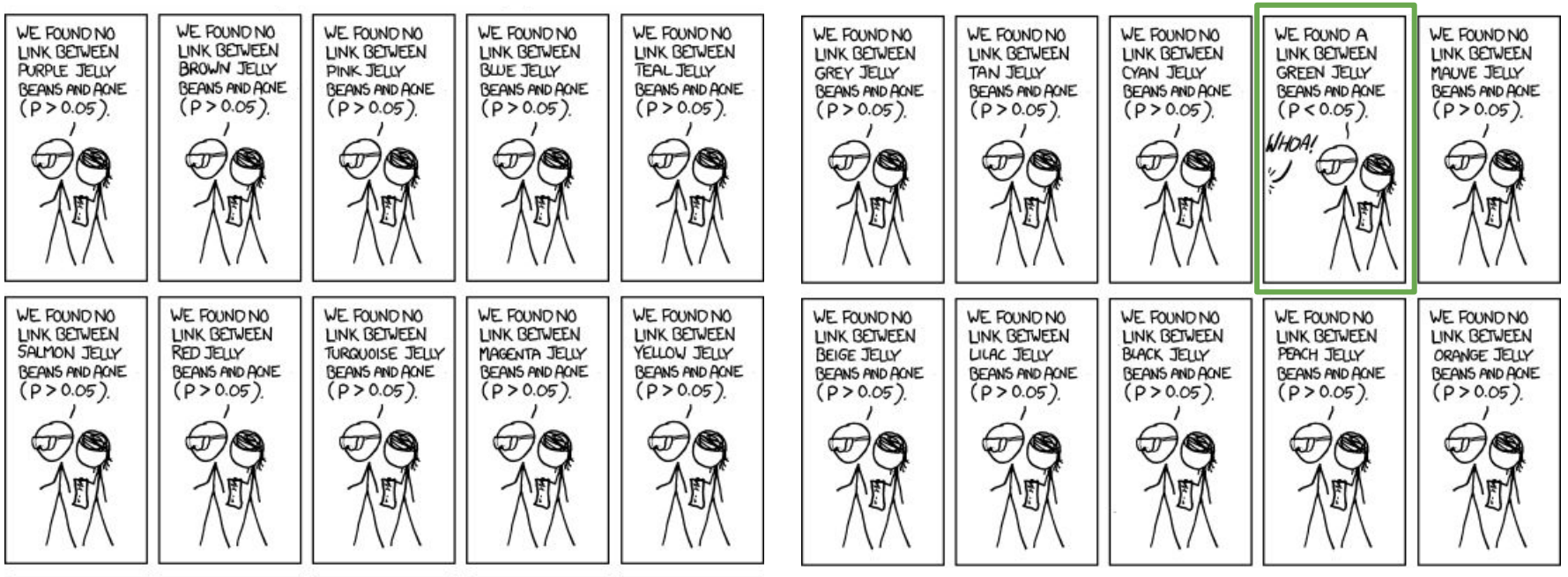
The following list is culled from peer-reviewed journal articles in which:
(a) the authors set themselves the threshold of 0.05 for significance,
(b) failed to achieve that threshold value for p and
(c) described it in such a way as to make it seem more interesting.

(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)

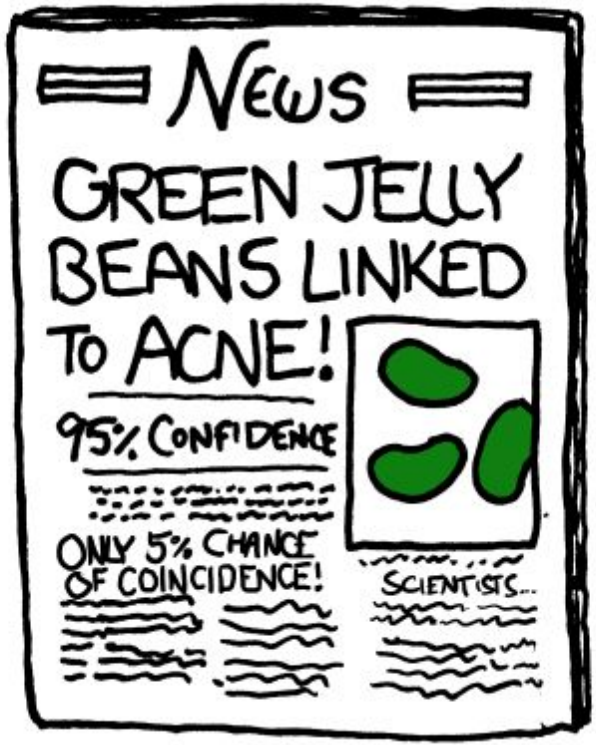
Multiple hypothesis testing



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)

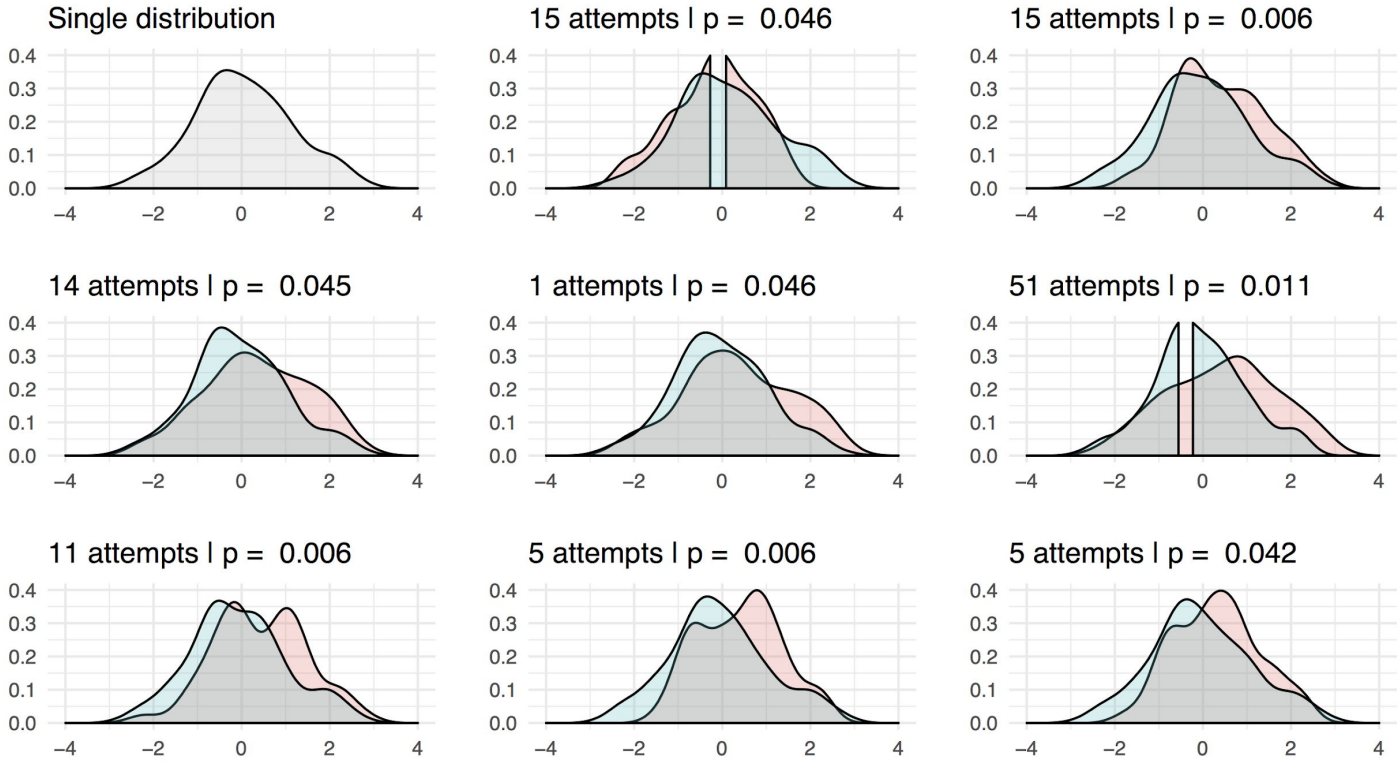
Multiple hypothesis testing

- 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

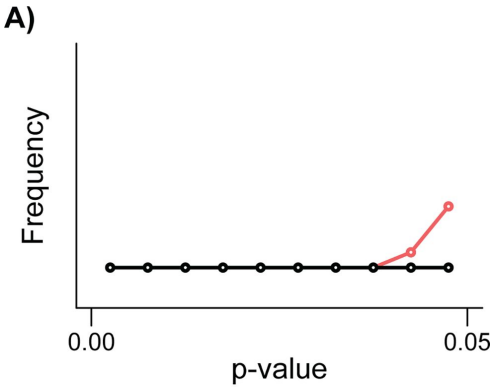
- P-hacking: Collect or select data or statistical analyses until nonsignificant results become significant.
    - Conducting analyses midway through experiments to decide whether to continue collecting data.
    - Recording many response variables and deciding which to report post-analysis
    - Deciding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

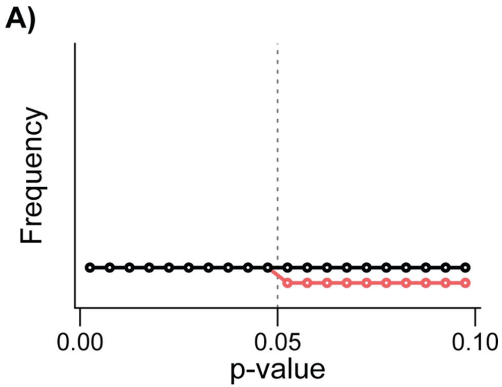
P-hacking and Publication bias

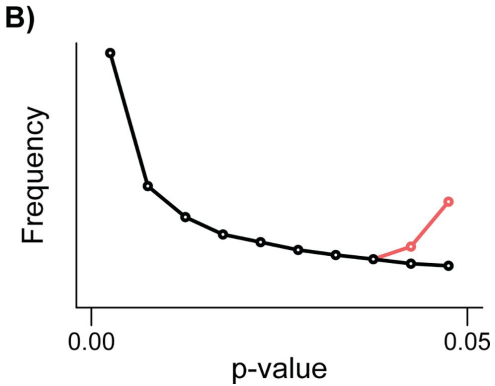


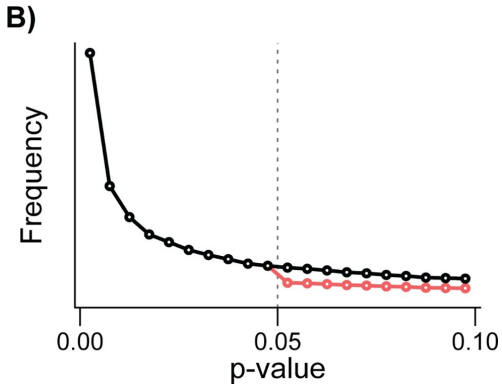
"When a measure become a target, it ceases to be a good measure" – Goodhart's Law

P-hacking and Publication bias



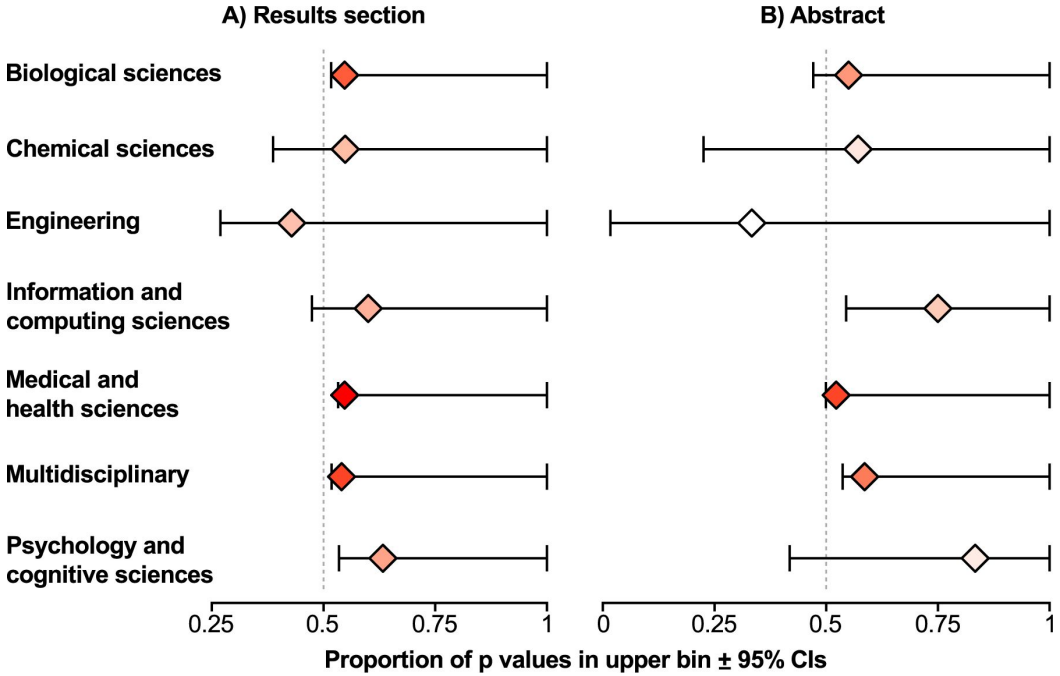






The Extent and Consequences of P-Hacking in Science https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106

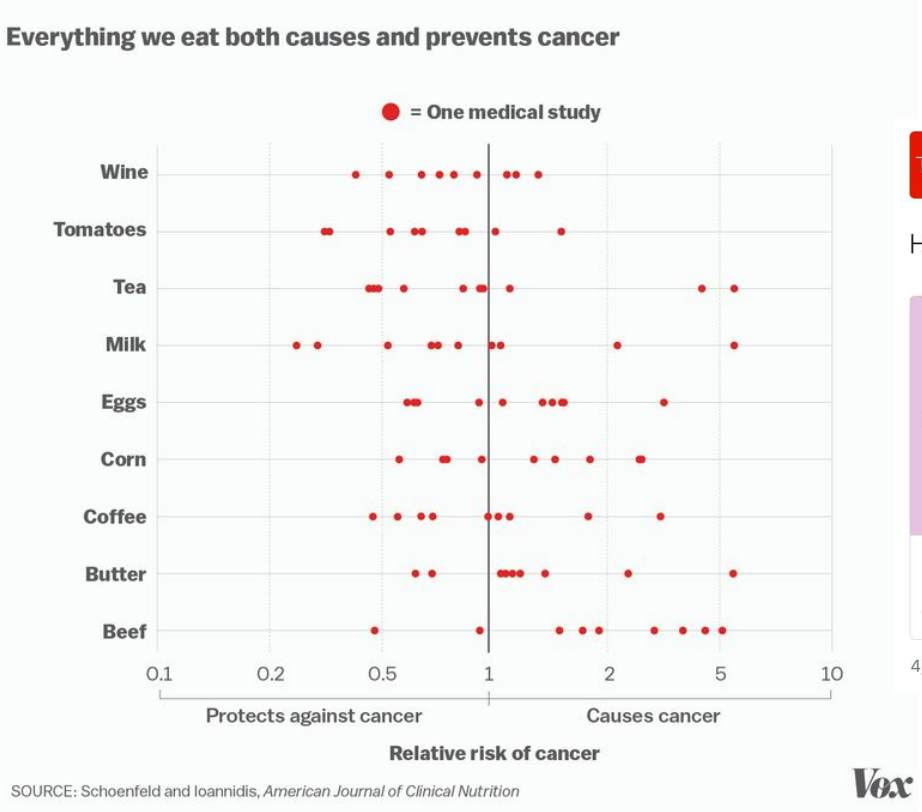
Publication bias and P-hacking



Hack Your Way To Scientific Glory: https://projects.fivethirtyeight.com/p-hacking/

The Extent and Consequences of P-Hacking in Science https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106

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

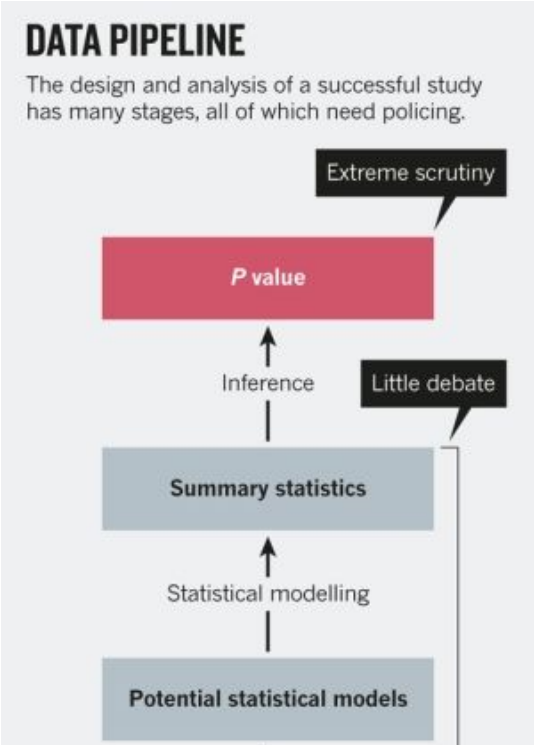
4/9/17, 6:15 AM

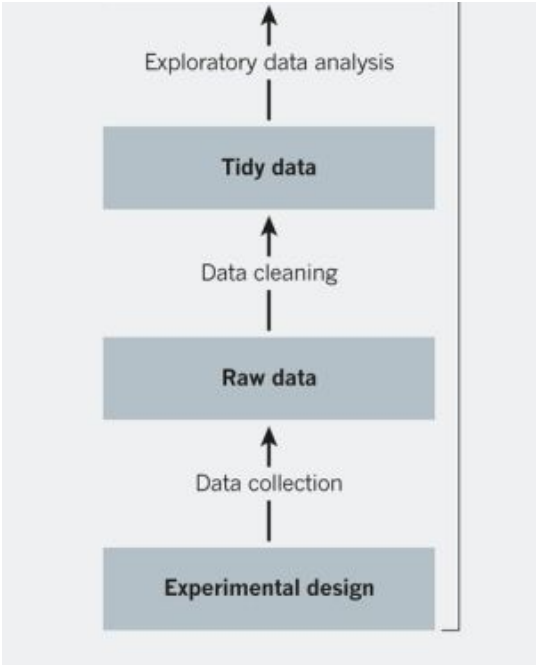
Questionable research practices

- 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!

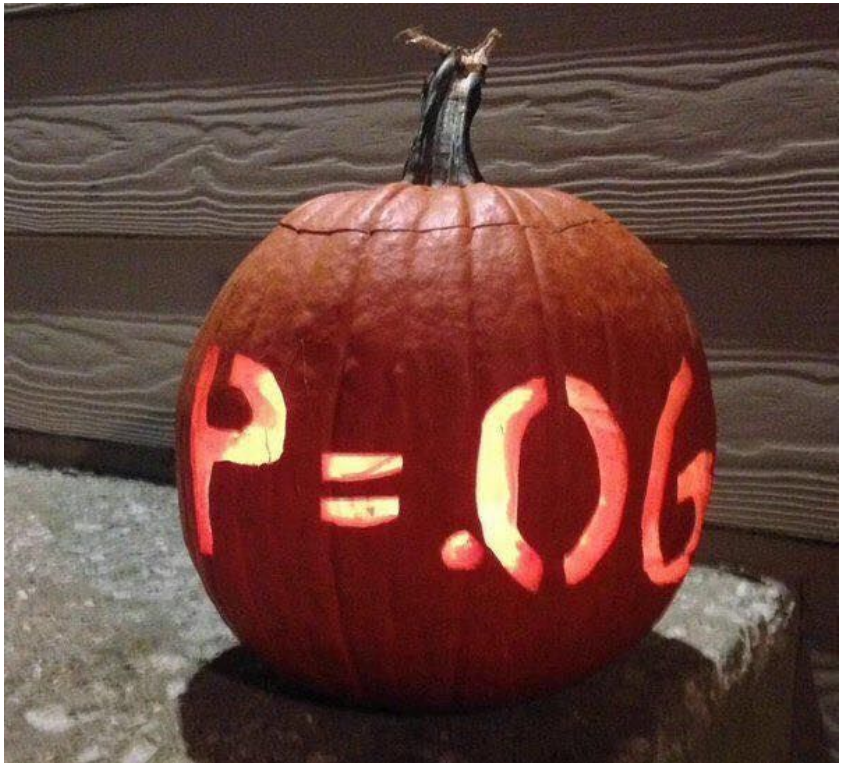




JT Leek, RD Peng http://www.nature.com/news/statistics-p-values-are-just-the-tip-of-the-iceberg-1.17412

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

What you need to do before the next class

- Complete the assignment
  - Implementing the permutation test
  - Exploring the dependence of p-value on effect size, sample size, & variance

- Concepts
  - Brush-up: Statistical power
  - Brush-up: Replication

What you need to do before the next class

- Answer the poll in the #syllabus-schedule channel on slack to figure out a good “final exam" time.
  - Not a typical exam but a really useful exercise.
  - Mandatory for all students taking this for credit.
  - Others, I strongly recommend doing it.