DSO 699: Modern Statistical Inference
Special Topics in Data Sciences an Operations

Week 1

Adel Javanmard

USC Marshall
Department of Data Sciences and Operations

Administration and logistics

Learning objectives

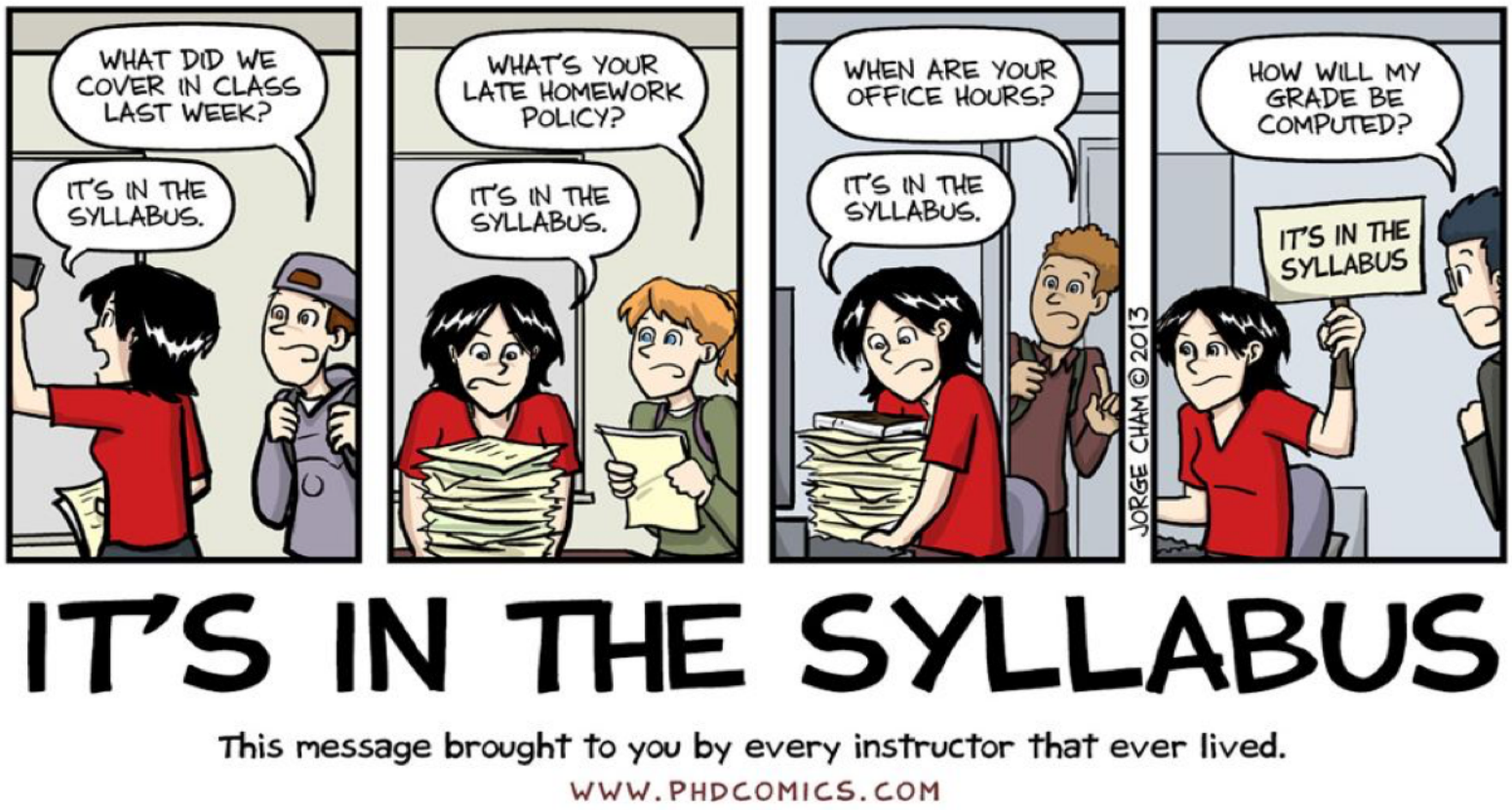
(Excerpted from syllabus)

- Motivated by high-dimensional data, we consider multiple hypothesis testing problems, and focus on the reliability of statistical methods and reproducibility of findings.
- Learn several common statistical tests and applications
- Learn about the challenges of multiple testing problems, common measures of error, and procedures to control these measures
- Conformal prediction and Inferential fairness
- Conditional randomization test
- Gaussian comparisons inequality and applications
- The course is theory oriented.
  - ✓ We use “Proof language”
  - ✓ We will have project presentation in the second half on some of the most recent work on statistical inference. The projects hopefully lead to research papers.

Syllabus

Available on Blackboard.

Covers all the administrative details (and due dates).



Blackboard

- Covers all the administrative details
  ✓ Assignments and due dates
- I will make announcements via Blackboard
- **Office hours:** After our class or send me an email and we schedule a zoom meeting
  ✓ My email: ajavanma@usc.edu

Course Readings

There is no required text books but as a PhD level course you are highly encouraged to consults outside resources.
Some useful references for background reading:

- An introduction to Statistical Learning by G. James, D. Witten, T. Hastie, R. Tibshirani
- Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction by B. Efron
- Testing Statistical Hypotheses by E. L. Lehmann, Joseph P. Romano
- High-dimensional probability: An introduction with applications in data science by Roman Vershynin

Grading

Participation/ Discussion	5%
Homework (3 sets)	30%
Midterm Exam	25%
Final Project/ Presentation/ Report	40%

- No final exam!
- Homework should be written in Latex with a copy of code included (if applicable).
- You are free in choosing your software: R, Python, Matlab, ...

Overview of Statistical Inference

three major task of statistics: estimation, prediction, inference

Statistical estimation

$$y = f_{\theta_0}(x; \text{noise})$$

$\theta \in \mathbb{R}^p$	$\rightarrow$	Unknown parameters
$y \in \mathbb{R}$	$\rightarrow$	(Observed) response variable
$x \in \mathbb{R}^d$	$\rightarrow$	(Observed) covariate vector (a.k.a features/ attributes)
$f_{\theta_0}(\cdot, ; \text{noise})$	$\rightarrow$	Parametric model

Estimation Problem: Estimate θ_0 from observations $\{x_i, y_i\}_{i=1}^n$

estimator is a random variable

Performance measure: $||\hat{\theta} - \theta_0||$ (probabilistic bounds), $\mathbb{E}[||\hat{\theta} - \theta_0||]$

Statistical prediction

in many cases, we do regression function estimation through in-sample prediciton; prediction calculates error on y, while estimation calculates error on f or theta.

Estimation and prediction are cousins!

- **In-sample prediction:**

$$\frac{1}{n}\ell(y_i, f_{\widehat{\theta}}(x_i)) \,,$$

where $\ell(\cdot, \dot{)}$ is a loss function (e.g., quadratic loss).

- **Out-of-sample prediction:**

$$\mathbb{E}[\ell(y, f_{\widehat{\theta}}(x))] \,,$$

where (x, y) are generated from the same distribution as training data.

these two quantities are similiar (proof uses uniform convergence) when p<n
if p~~n than they can differ much (overfitting)

Statistical Inference

beyond point estimates, want to talk about how certain I am
Not only the answer, but want to say how confident I am with the answer
confident: about whether the answer will change if the dataset is changed

Inferential thinking: going from data to the underlying phenomena behind the data

Hypothesis testing (HT): A rigorous statistical framework to answer Yes/No questions about underlying phenomena

Statistical Inference

Inferential thinking: going from data to the underlying phenomena behind the data

Hypothesis testing (HT): A rigorous statistical framework to answer Yes/No questions

- Is there a difference in the mean blood pressure of mice in the control group and mice in the treatment group?
- Is there a difference in the mean Glucose level of COVID patients and people in the control group (healthy)?
- Does drinking coffee increase your chance of cancer?

Null hypothesis

We divide the world into two possibilities:

null hypothesis and *alternative hypothesis*

Null hypothesis: denoted by H_0, is the default state of belief about the world.

H_0: There is no difference between the mean glucose level of people with and without COVID

Null hypothesis

We divide the world into two possibilities:

null hypothesis and *alternative hypothesis*

Null hypothesis: denoted by H_0, is the default state of belief about the world.

H_0: There is no difference between the mean glucose level of people with and without COVID

- Null hypothesis is boring!
- We may hope that our data tell us otherwise
- Alternative hypothesis, denoted by H_a represents something different than null and interesting! (A discovery)
- Typically, H_a is the opposite of H_0 (H_a: H_0 does not hold)
- We either reject H_0 (data provides evidence in favor of H_a) or we fail to reject H_0

Null hypothesis

We divide the world into two possibilities:

null hypothesis and *alternative hypothesis*

Null hypothesis: denoted by H_0, is the default state of belief about the world.

H_0: There is no difference between the mean glucose level of people with and without COVID

- Null hypothesis is boring!
- We may hope that our data tell us otherwise
- Alternative hypothesis, denoted by H_a represents something different than null and interesting! (A discovery)
- Typically, H_a is the opposite of H_0 (H_a: H_0 does not hold)
- We either reject H_0 (data provides evidence in favor of H_a) or we fail to reject H_0

usually complement, but not always
e.g. difference=0 vs difference >5

fail to reject $H_0 \neq$ accept H_0

Testing a hypothesis

1. Define the Null and Alternative Hypotheses
2. Construct the Test Statistic
3. Compute the p-Value
4. Decide Whether to Reject the Null Hypothesis

Construct the Test Statistic

We want to use our data to find evidence for or against the null hypothesis

Test statistic: A quantity that summarized the extent to which the data are consistent with H_0

Construct the Test Statistic

We want to use our data to find evidence for or against the null hypothesis

Test statistic: A quantity that summarized the extent to which the data are consistent with H_0 captures degree of evidence for H0

- There is no unique test statistic!
- The choice of test statistics depends on both H_0 and H_a. also depend on alternative!!
- Commonly-used test statistics follow a well known statistical distribution under H_0
  (e.g., normal, t-distribution, χ^2-distribution, F-distribution, etc)
  but sometimes we don't know the test statistic distribution under H0

Test Statistic: an example

Let $x_1^p, \ldots, x_{n_p}^p$ denote the glucose level for (COVID) patients group
Let $x_1^h, \ldots, x_{n_h}^h$ denote the glucose level for control (healthy) group
Let $\mu_p = \mathbb{E}(X^p)$ and $\mu_h = \mathbb{E}(X^h)$. We want to test

$$H_0 : \mu_p = \mu_h \quad \text{versus} \quad \mu_p \neq \mu_h$$

Test Statistic: an example

Let $x_1^p, \ldots, x_{n_p}^p$ denote the glucose level for (COVID) patients group
Let $x_1^h, \ldots, x_{n_h}^h$ denote the glucose level for control (healthy) group
Let $\mu_p = \mathbb{E}(X^p)$ and $\mu_h = \mathbb{E}(X^h)$. We want to test

$$H_0 : \mu_p = \mu_h \quad \text{versus} \quad \mu_p \neq \mu_h$$

Two-sample t-statistic:

$$T = \frac{\hat{\mu}_p - \hat{\mu}_h}{s\sqrt{\frac{1}{n_p} + \frac{1}{n_h}}}$$

where

$$\hat{\mu}_p = \frac{1}{n_p}\sum_{i=1}^{n_p} x_i^p, \ \hat{\mu}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_i^h, \quad s = \sqrt{\frac{(n_p-1)s_p^2 + (n_h-1)s_h^2}{n_p + n_h - 2}}$$

if we have all population data, we only have to look at mean diff; but we only have sample so we have to normalize. Even though the mean diff is large, if the sd is large we can't say it's significant.

Test Statistic: an example

Let $x_1^p, \ldots, x_{n_p}^p$ denote the glucose level for (COVID) patients group
Let $x_1^h, \ldots, x_{n_h}^h$ denote the glucose level for control (healthy) group
Let $\mu_p = \mathbb{E}(X^p)$ and $\mu_h = \mathbb{E}(X^h)$. We want to test

$$H_0 : \mu_p = \mu_h \quad \text{versus} \quad \mu_p \neq \mu_h$$

Two-sample t-statistic:

$$T = \frac{\hat{\mu}_p - \hat{\mu}_h}{s\sqrt{\frac{1}{n_p} + \frac{1}{n_h}}}$$

where

$$\hat{\mu}_p = \frac{1}{n_p}\sum_{i=1}^{n_p} x_i^p, \ \hat{\mu}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_i^h, \quad s = \sqrt{\frac{(n_p-1)s_p^2 + (n_h-1)s_h^2}{n_p + n_h - 2}}$$

$|T| \uparrow$ evidence against H_0 (in favor of H_a)

Construct p-values

- Why should we go from test statistic to p-value?
  How much evidence against H_0 is provided by a given value of $|T|$?
  (How large is large?!)

Construct p-values

large statistic value indicates much evidence agains h0

- Why should we go from test statistic to p-value?
  How much evidence against H_0 is provided by a given value of $|T|$?
  (How large is large?!)
  to answer this question, we have to know the typical value of T (or equivalently,
- What is p-value? its distribution) under H0
  - ✓ it is the probability of observing a comparable or more extreme value of the test statistics under the null hypothesis.

$$p = \mathbb{P}_{H_0}\left(|T| \geq |T^{\text{data}}|\right)$$

small p-value $\Rightarrow$ evidence *against* H_0

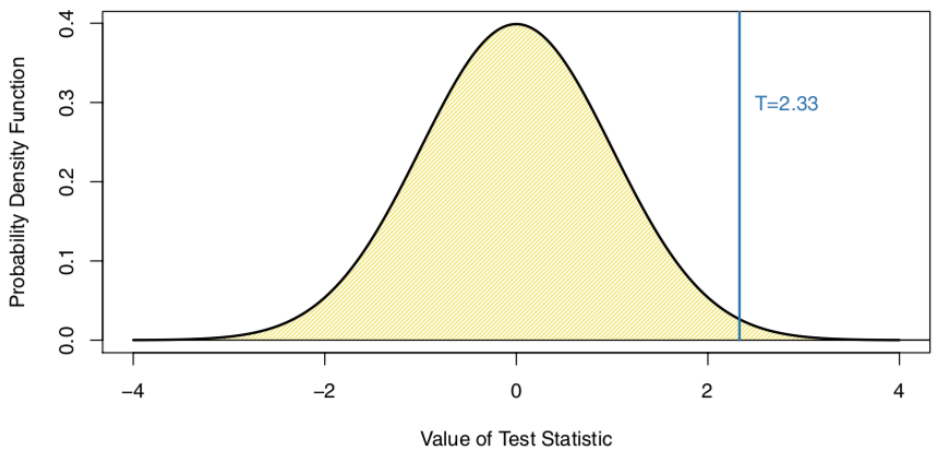
Construct p-values

- Why should we go from test statistic to p-value?
  How much evidence against H_0 is provided by a given value of $|T|$?
  (How large is large?!)

- What is p-value?
  - ✓ it is the probability of observing a comparable or more extreme value of the test statistics under the null hypothesis.

$$p = \mathbb{P}_{H_0}\left(|T| \geq |T^{\text{data}}|\right)$$

small p-value $\Rightarrow$ evidence *against* H_0



p-value = 0.02

Construct p-values

- Why should we go from test statistic to p-value?
  How much evidence against H_0 is provided by a given value of $|T|$?
  (How large is large?!)

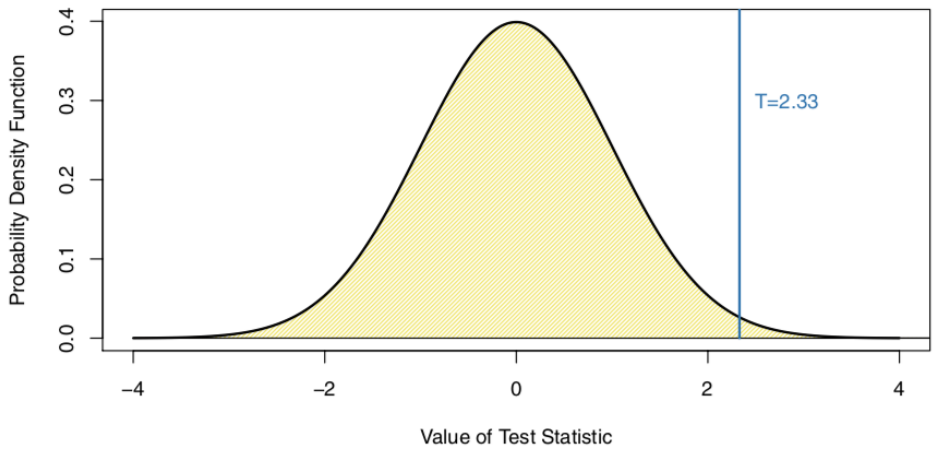
if the variance of the statisitc is large,
p-value is large for the same T value

- What is p-value?
  - ✓ it is the probability of observing a comparable or more extreme value of the test statistics under the null hypothesis.

range 0-1 so easy to interpret
no arbitrarily large number

$$p = \mathbb{P}_{H_0}\left(|T| \geq |T^{\text{data}}|\right)$$

small p-value $\Rightarrow$ evidence *against* H_0



p-value = 0.02

Exercise: Show that under null, p-value is uniform.

for ANY test statistic
besides t-statistic

Test statistic's null distribution

Test statistic is a random variable
p-value is just a deterministic transformation of the test statistic

To construct p-value we need to know the test statistic's null distribution.

What if we don't?

$$|T^{data}| \sim F$$
$$pval = g(|T|) = 1 - F(T^{data})$$

the CDF of $g(|T^{data}|)$ is:

$$\begin{aligned}
\mathbb{P}[g(|T| \leq \alpha)] &= \mathbb{P}[1 - F(T^{data} \leq \alpha)] \\
&= \mathbb{P}[(1-\alpha) \leq F(|T^{data}|)] \\
&= \mathbb{P}[F^{-1}(1-\alpha) \leq |T^{data}|] \\
&= 1 - \mathbb{P}[|T^{data} \leq F^{-1}(1-\alpha)] \\
&= 1 - F(F^{-1}(1-\alpha)) \\
&= 1 - (1-\alpha) = \alpha
\end{aligned}$$

Test statistic's null distribution

To construct p-value we need to know the test statistic's null distribution.

What if we don't?

- Data-driven approach (resampling?)
- helps to avoid unrealistic assumptions about data!
- More in HW1 ....

  but the computational cost is high

Decide Whether to Reject the Null Hypothesis

Recall that the smaller the p-value, the stronger the evidence against H_0.

Decide Whether to Reject the Null Hypothesis

Recall that the smaller the p-value, the stronger the evidence against H_0.

Decision rule:

$$R = \begin{cases} 1 & \text{if } p \leq \alpha\,, \\ 0 & \text{otherwise} \end{cases}$$

What is the right choice of α?

Decide Whether to Reject the Null Hypothesis

Recall that the smaller the p-value, the stronger the evidence against H_0.

Decision rule:

$$R = \begin{cases} 1 & \text{if } p \leq \alpha\,, \\ 0 & \text{otherwise} \end{cases}$$

p-value measures how much our result came from the fundamental phenomenon rather than from the randomness of the data

What is the right choice of α? It is a matter of convention:

- some areas of science are happy with $\alpha = 0.05$
- some areas of physics set $\alpha = 10^{-9}$

Errors in hypothesis testing

		Truth	
		H_0	H_a
Decision	Reject H_0	Type I error (false positive)	Correct (true positive)
	Fail to Reject H_0	Correct (true negative)	Type II error (false negative)

something interesting happened, but failed to detect it
because of randomness
or my test statistic is not good enough
(does not extract enough information from the data)
to detect the signal

reject H0 = positive = scientific discovery
false positive means we got positive result falsely

Errors in hypothesis testing

		Truth	
		H_0	H_a
Decision	Reject H_0	Type I error (false positive)	Correct (true positive)
	Fail to Reject H_0	Correct (true negative)	Type II error (false negative)

- **Size of a test** (a.k.a, type I error rate, false positive rate):

$$\mathbb{P}(\text{test rejects } H_0 | H_0)$$

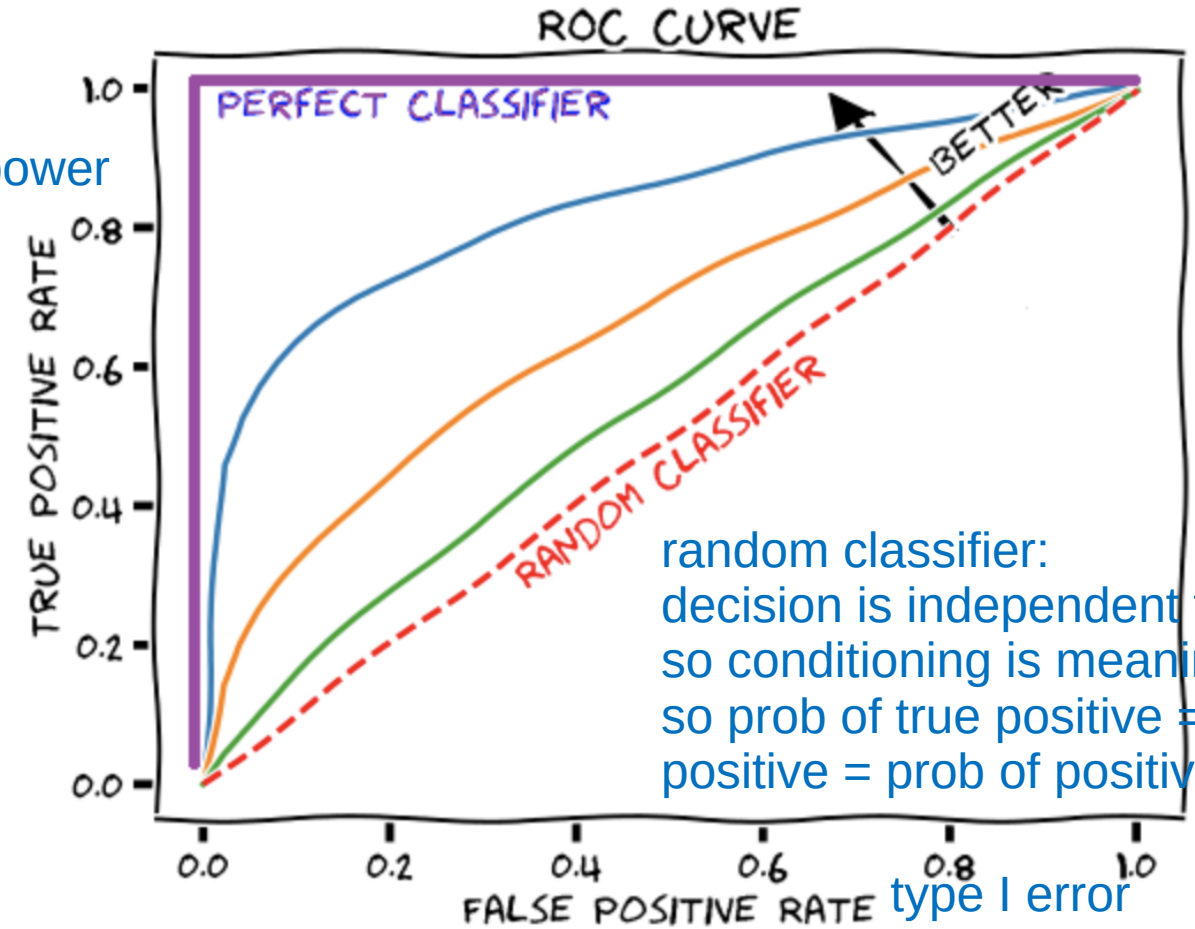
- **Power of a test** (not making type II, aka, true positive rate):

$$\mathbb{P}(\text{rejects } H_0 | H_a)$$

control size first, than try to maximize power
or sometimes sum them and show ->0

Receiver operating characteristic (ROC) curve

conditioning is different, so do not sum to 1



power

random classifier:
decision is independent from data
so conditioning is meaningless
so prob of true positive = prob of false positive = prob of positive

type I error

how to draw it? change alpha (maybe for fixed alternative)
by simulation or math

Performance measure: Area under curve (AUC), a.k.a Concordance Statistic (C-)Statistic

False positive versus False negative

- Recall that the treatment of the null hypothesis and the alternative hypothesis is asymmetric.
- False positive rate (size of the test) is more important to be controlled.
- False negative control (i.e. power of the test) is a luxury of the test!
- **Example:** Cheating in an exam ...

Is p-value the right approach?

- The concept of p-values in recent years have been the topic of extensive commentary
- some social science journals have gone so far as to ban the use of p-values altogether!

Is p-value the right approach?

- The concept of p-values in recent years have been the topic of extensive commentary
- some social science journals have gone so far as to ban the use of p-values altogether!

- p- value is one of the most used and abused notion in all of statistics!

- For example, it is sometimes said that the p-value is the probability that H_0 holds. Is this correct?

Is p-value the right approach?

- The concept of p-values in recent years have been the topic of extensive commentary
- some social science journals have gone so far as to ban the use of p-values altogether!

- p- value is one of the most used and abused notion in all of statistics!

- For example, it is sometimes said that the p-value is the probability that H_0 holds. Is this correct?
- No, p-value is the fraction of the time that we would expect to see such an extreme value of the test statistic.

Pitfalls of Modern Statistical Inference

Multiple testing

- Much of classical statistics has been focused on single hypothesis testing.
- Multiple testing is indispensable in the era of data deluge.

Multiple testing

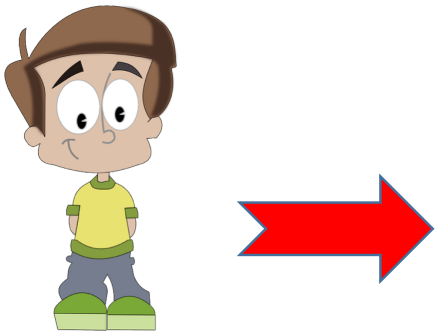
- Much of classical statistics has been focused on single hypothesis testing.
- Multiple testing is indispensable in the era of data deluge.
  Suppose we wish to test m null hypotheses simultaneously:

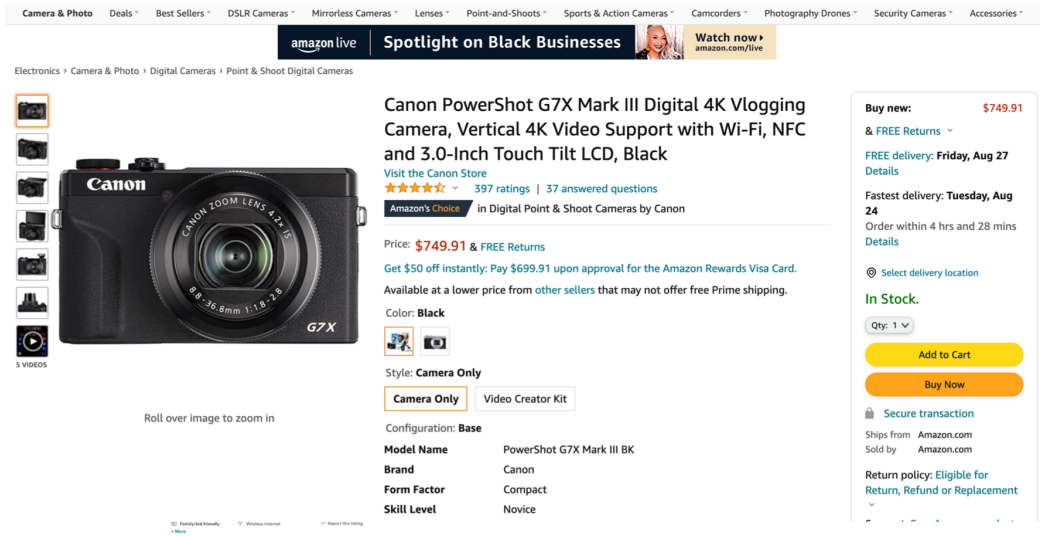
  $H_{0,m}$: the mean value of the m^{th} biomarker among mice in the control group equals that among mice in the treatment group.

Multiple testing

- Much of classical statistics has been focused on single hypothesis testing.

- Multiple testing is indispensable in the era of data deluge.





User (with features $x_t \in \mathbb{R}^d$, cookies, gender, age, location, search history, ...)

$H_{0,m}$: the mean value of the m^{th} features is the same among users who purchased the product versus those who did not.

The challenge of multiple testing

- Suppose we are testing 10^6 hypotheses. If we reject a null hypothesis whose p-value is less than 0.05, then how many type I errors do we have?

The challenge of multiple testing

- Suppose we are testing 10^6 hypotheses. If we reject a null hypothesis whose p-value is less than 0.05, then how many type I errors do we have?

▷ stockbroker example

▷ Fair coin example

The challenge of multiple testing

- Suppose we are testing 10^6 hypotheses. If we reject a null hypothesis whose p-value is less than 0.05, then how many type I errors do we have?

▷ stockbroker example

▷ Fair coin example

- We need to account for multiplicity!
- This is one (out of many) reasons for the reproducibility crisis.

Reproducibility

$$(y^{(1)}, X^{(1)}) \longrightarrow \widehat{\theta}^{(1)}, \qquad \widehat{\theta}_i^{(1)} \neq 0$$

$$(y^{(2)}, X^{(2)}) \longrightarrow \widehat{\theta}^{(2)}, \qquad \widehat{\theta}_i^{(2)} \overset{?}{\neq} 0$$

Reproducibility

Can we reproduce our findings in other independent studies, not exactly but up to statistical errors?

The reproducibility crisis

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

The reproducibility crisis

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

- Conducted replications of 100 experimental and correlational studies published in three psychology journals
- 97% of original studies had significant results (P < .05)
- Only 36% of original findings could be reproduced.

Science Contents News Careers Journals

Read our COVID-19 research and news.

SHARE

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration*,†

*All authors with their affiliations appear at the end of this paper.

↵†Corresponding author. E-mail: nosek@virginia.edu

– Hide authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

The reproducibility crisis

- Chose 53 studies that are considered landmarks in basic cancer science
- Only 6 out of 53 ($\sim 11\%$) could be replicated by Amgen

MENU nature

Published: 28 March 2012

Drug development

Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis

Nature **483**, 531–533(2012) | Cite this article

44k Accesses | **1449** Citations | **2034** Altmetric | Metrics

A Clarification to this article was published on 02 May 2012

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

The reproducibility crisis

- Chose 53 studies that are considered landmarks in basic cancer science
- Only 6 out of 53 ($\sim 11\%$) could be replicated by Amgen

MENU nature

Published: 28 March 2012

Drug development

Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis

Nature **483**, 531–533(2012) | Cite this article

44k Accesses | **1449** Citations | **2034** Altmetric | Metrics

A Clarification to this article was published on 02 May 2012

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

- Bad Science can kill!

January 20, 2017 06:54 AM EST | Pharma

The FDA's cautionary Hall of Shame: 22 'breakthrough' drugs that suddenly crashed in PhIII

John Carroll
Editor & Founder

Reproducibility crisis in media

THE MAGAZINE: SEPTEMBER 20, 2011

THE NEW YORKER

SUBSCRIBE | IN THE MAGAZINE | BLOGS | AUDIO & VIDEO | LISTINGS

POLITICS | PROFILES | THE TALK OF THE TOWN | COMMENT | TABLE OF CONTENTS | THE FINANCIAL PAGE

ANNALS OF SCIENCE

THE TRUTH WEARS OFF

Is there something wrong with the scientific method?

BY JONAH LEHRER

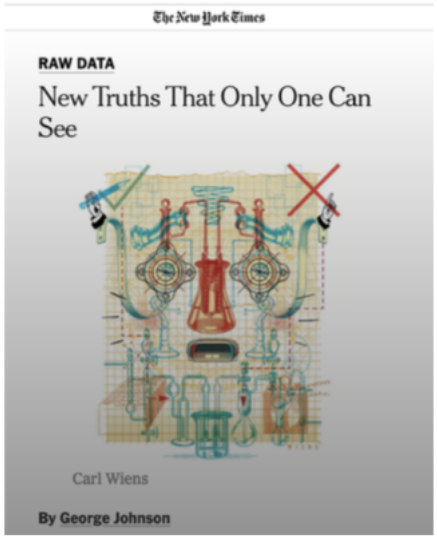
DECEMBER 13, 2010

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects' psychiatric symptoms. As a result, second-generation antipsychotics had become one of the fastest-growing and most profitable pharmaceutical classes. By 2001, Eli Lilly's Zyprexa was generating more revenue than Prozac. It remains the company's top-selling drug.

Many results that are rigorously proved and accepted start shrinking in later studies.

KEYWORDS

Scientific Experiments; Decline Effect; Replicability; Scientists; Statistics; Jonathan Schooler; Scientific Theories



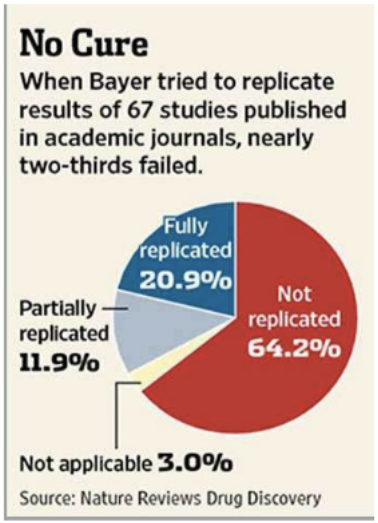
The New York Times

RAW DATA

New Truths That Only One Can See

Carl Wiens

By George Johnson



No Cure

When Bayer tried to replicate results of 67 studies published in academic journals, nearly two-thirds failed.

Fully replicated 20.9%

Partially replicated 11.9%

Not replicated 64.2%

Not applicable 3.0%

Source: Nature Reviews Drug Discovery

The Economist

Washington's lawyer surplus

How to do a nuclear deal with Iran

Investment tips from Nobel economists

Junk bonds are back

The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG.

Why are we facing the reproducibility crisis?

1. **Age of public datasets! (accounting for multiplicity)**

Why are we facing the reproducibility crisis?

1. **Age of public datasets! (accounting for multiplicity)**
   - Popular public datasets are used over and over by independent research teams to evaluate theories

   - Without proper accountability for multiple tests this leads to many spurious discoveries.

Why are we facing the reproducibility crisis?

A new scientific framework:

Hypothesis-driven approach to science

- Formulate hypothesis
- Collect data to test prediction
- Reject or accept the theory

Data-driven approach to science

- Collect data
- Formulate hypothesis
- Reject or accept the theory

Why are we facing the reproducibility crisis?

A new scientific framework:

Hypothesis-driven approach to science

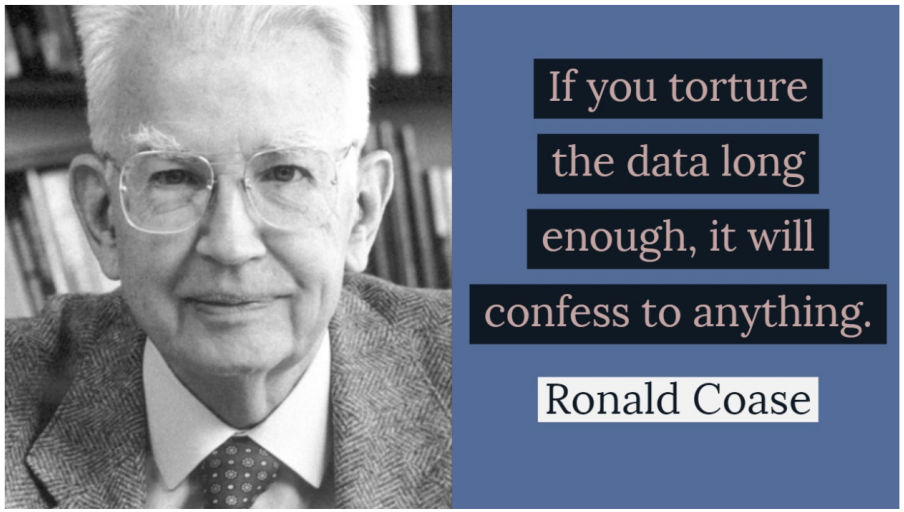
- Formulate hypothesis
- Collect data to test prediction
- Reject or accept the theory

Data-driven approach to science

- Collect data
- Formulate hypothesis
- Reject or accept the theory

Data Snooping (data dredging, data fishing, p-hacking)

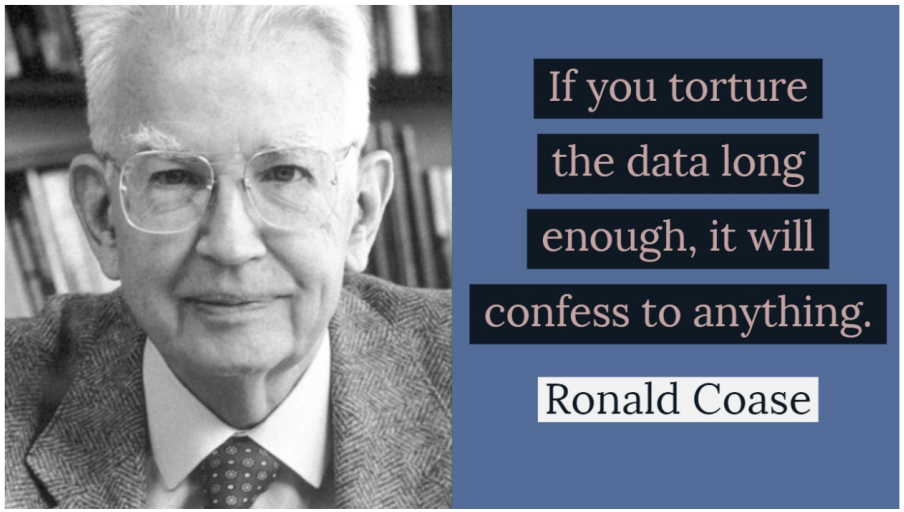
Data snooping is a form of statistical bias manipulating data or analysis to artificially get statistically significant results.



Data Snooping (data dredging, data fishing, p-hacking)

Data snooping is a form of statistical bias manipulating data or analysis to artificially get statistically significant results.

You observe statistically significant results not because of some underlying phenomena, but because of probabilistic nature of all statistical tests.



Jelly beans experiment

Part 1: Does eating jelly beans cause acne?
You conduct a survey among 500 volunteers and collect the following information:

- if a participant eats jelly beans on a regular basis;
- participant's acne condition measured as a value between 0 and 1.

Jelly beans experiment

Part 1: Does eating jelly beans cause acne?
You conduct a survey among 500 volunteers and collect the following information:

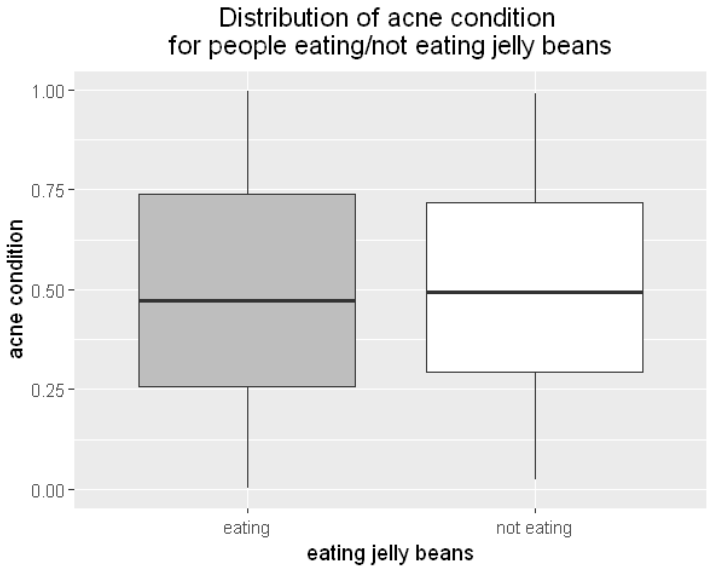
- if a participant eats jelly beans on a regular basis;
- participant's acne condition measured as a value between 0 and 1.

Data simulation: We generate two independent random variables:

- "acne condition" indicating the acne condition of a participant drawn from uniform distribution $U(0, 1)$;
- eating indicating jelly bean consumption will drawn from Bernoulli distribution $Bern(0.9)$;

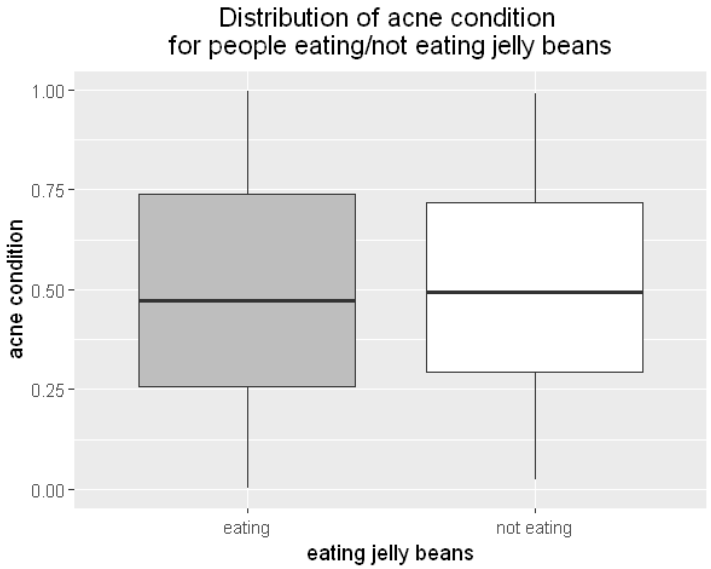
Jelly beans experiment

Part 1: Data visualization



Jelly beans experiment

Part 1: Data visualization



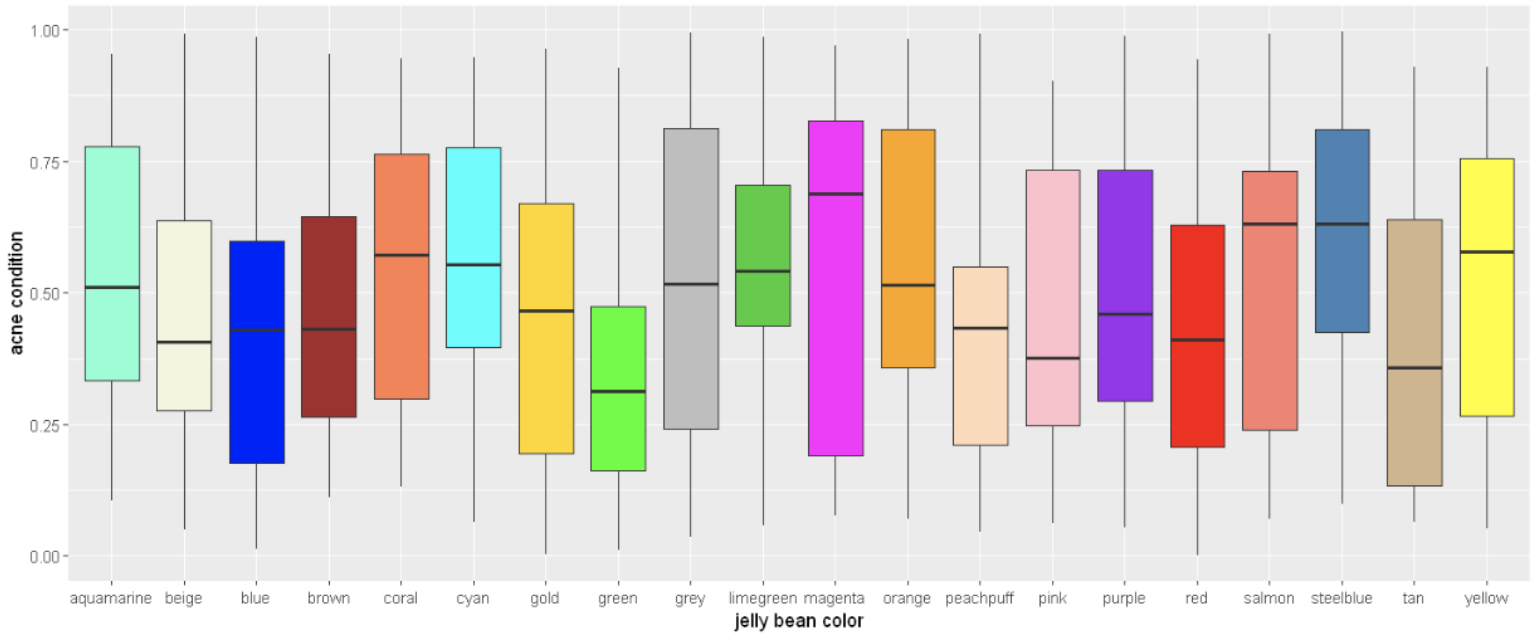
Statistical significance: You run t-test and get p-value = 0.60

Jelly beans experiment

Part 2: Which jelly bean color causes acne?
Your colleague suggests you to further investigate the data and check different colors of jelly beans.

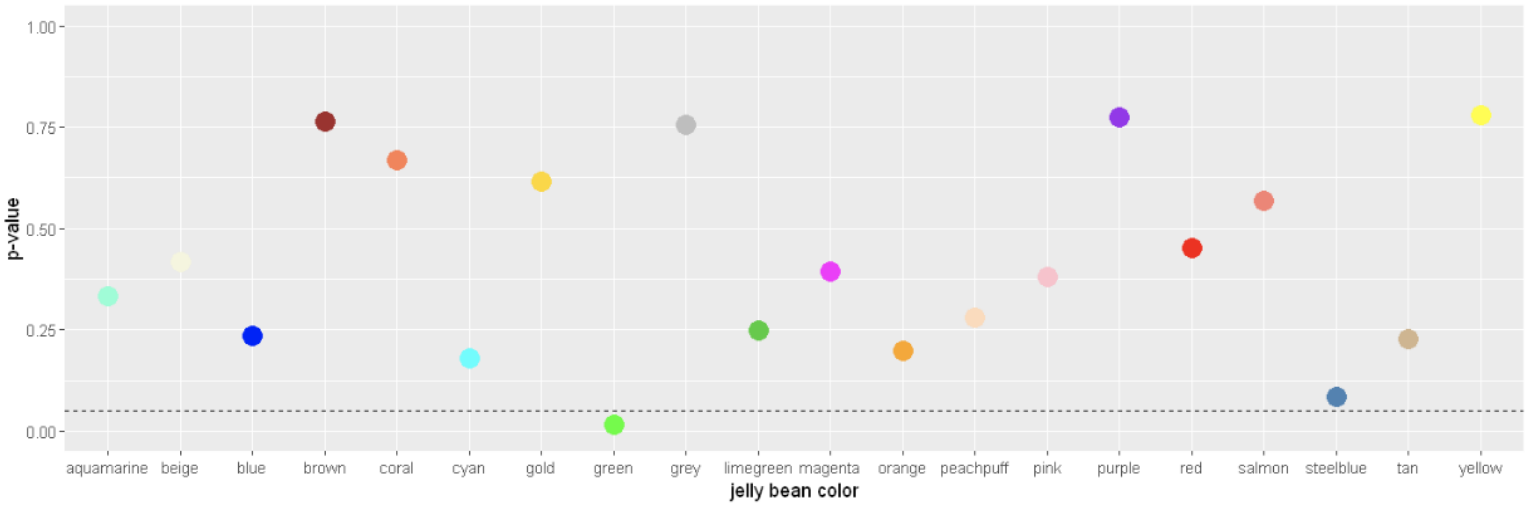
Jelly beans experiment

Part 2: Which jelly bean color causes acne?
Your colleague suggests you to further investigate the data and check different colors of jelly beans.
Suppose there exist 20 different jelly bean colors and that people eat jelly beans of each color with equal probability.



Jelly beans experiment

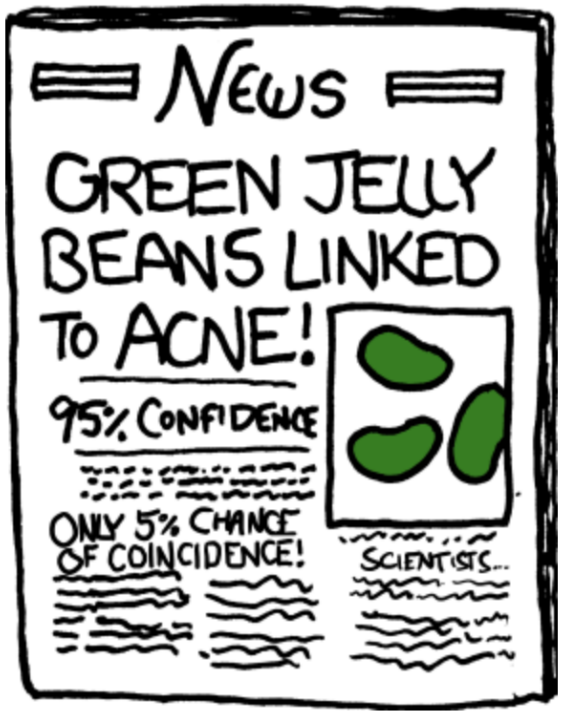
Part 2: Let's look at p-values (based on t-test)



p value for green color = 0.01605542

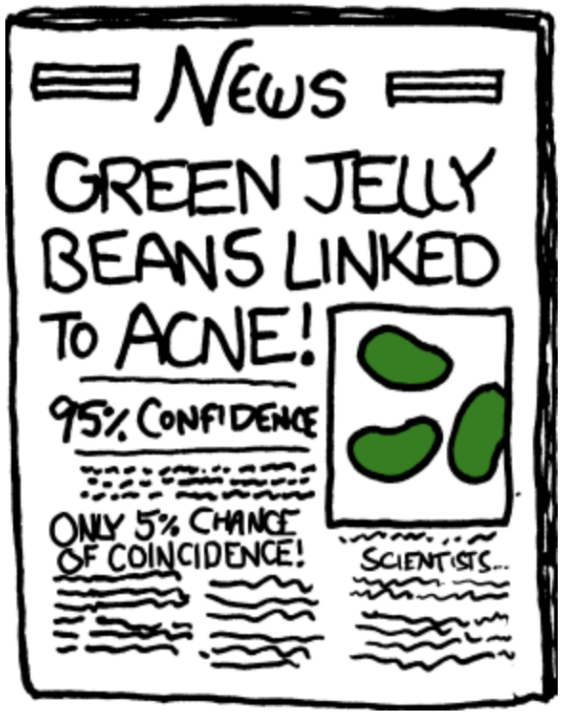
Jelly beans experiment

Yay, You finally publish your breaking news!



Jelly beans experiment

Yay, You finally publish your breaking news!



- What happened?

There is 5% chance to observe a significant result under the null, so testing 20 hypotheses will result in $1 - 0.95^{20} \approx 64\%$ chance to observe at least one significant test among these 20 experiments.

Recap

We talked about

- hypothesis testing
- notions of p-value, test statistics
- Pitfalls of modern statistical inference
  - ✓ Multiple hypotheses testing
  - ✓ Data snooping