College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

# Hypothesis Testing I: Why, Learning Lingo, $\chi^2$

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

APRIL 24, 2017

## Does the model fit?

- We've assumed
  - Our models are right
  - Our parameter estimates are good

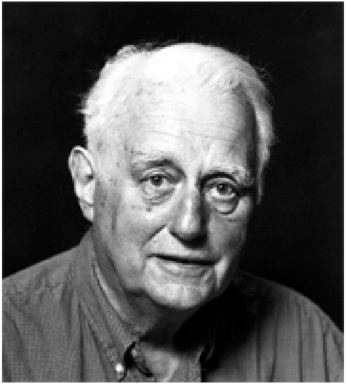
## Does the model fit?

- We've assumed
  - Our models are right
  - Our parameter estimates are good
- Not always true
- How do we know if distributions / parameters are any good?

## Importance for Data Science

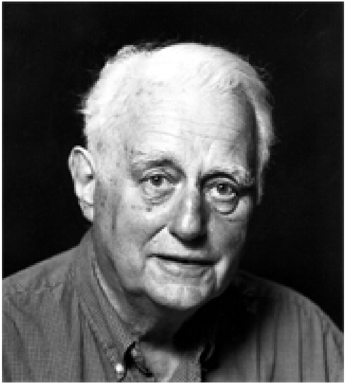
- Learning the mindset
- Not trusting your data
- Communicating uncertainty
- Testing hypotheses

## Lincoln Moses



- Stanford Statistician
- Learn one thing: Use Error Bars

## Lincoln Moses



- Stanford Statistician
- Learn one thing: Use Error Bars
- After visiting US government: Use data

College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

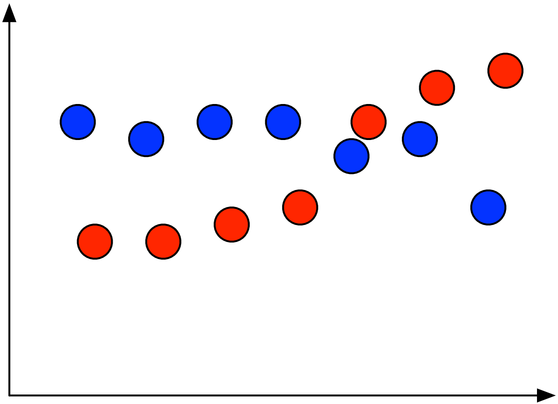
# Hypothesis Testing I: Making Decisions

INFO-2301: Quantitative Reasoning 2

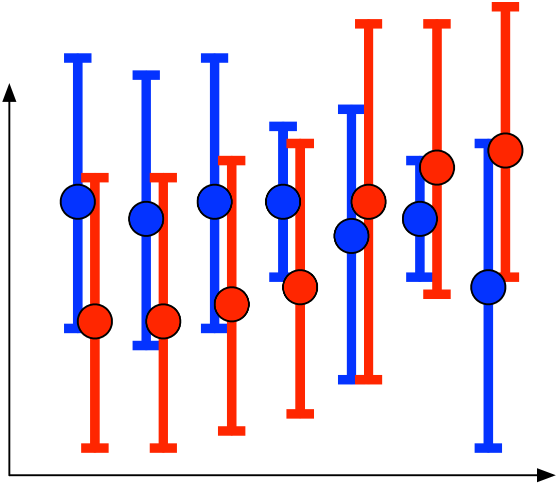
Michael Paul and Jordan Boyd-Graber

APRIL 24, 2017

## Point Estimates Lie



## Point Estimates Lie



## So how can you make a decision?

- Error bars help, but not systematic
- Make the point that decisions need to not just look at single estimates but *distributions*
- Statistical Test: Deciding whether a hypothesis is true or not

## Statistical Test Lingo

- Null hypothesis
- test statistic
- p-value
- p-hacking

## Null hypothesis

**Null Hypothesis**

A statement that can be validated through a statistic derived from observations.

- Often status quo
- Goal prove false: “reject the null”
- Phrased in terms of distributions

**Examples**

- Average body temperature 98.6?
- Voting republican and education independent?

## Body temperature

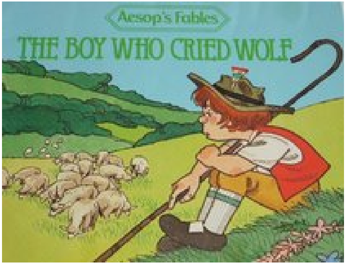
$n = 130$, $\bar{x} = 98.249$, standard deviation $s = 0.7332$.

- Not exactly equal (but wouldn't expect that)
- Is the difference meaningful?
- Null hypothesis, $H_0 : \mu = 98.6$
- Alternative hypothesis, $H_a : \mu \neq 98.6$

## What can happen

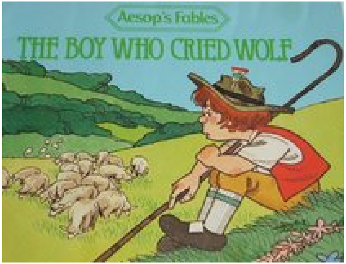
| Measured / Perceived | Reality: True | Reality: False |
|---|---|---|
| True | Correct ☺ | Type I<br>False Positive |
| False | Type II<br>False Negative | Correct ☺ |

## Boy who cried wolf



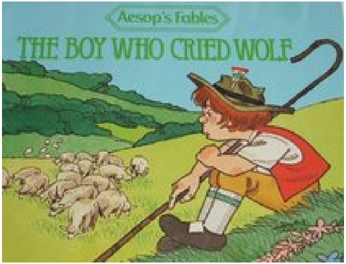

- Null hypothesis (status quo): no wolf

## Boy who cried wolf



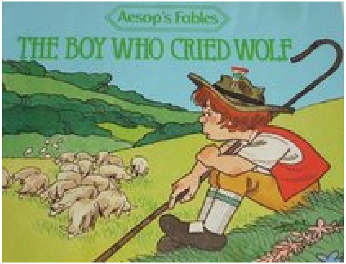

- Null hypothesis (status quo): no wolf
- First error, Type I: villagers believed there was wolf (but there wasn't)

## Boy who cried wolf




- Null hypothesis (status quo): no wolf
- First error, Type I: villagers believed there was wolf (but there wasn't)
- Second error, Type II: villagers believed there was no wolf (when there was)

## Boy who cried wolf

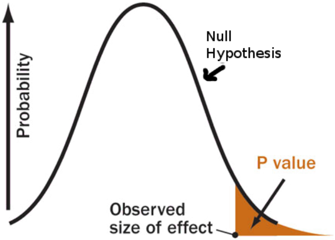



- Null hypothesis (status quo): no wolf
- First error, Type I: villagers believed there was wolf (but there wasn't)
- Second error, Type II: villagers believed there was no wolf (when there was)
- Type I and Type II in that order

## Test Statistic

- Measurement of how far observations deviate from null hypothesis (e.g., $\bar{x}$ far from $\mu$)
- Test statistic is paired with a distribution that measures deviation
- Lower probability test statistics let you reject the null

## $p$-value




- Probability of null hypothesis being true
- Lower is better
- Common critical values $\alpha$: 0.05, 0.01
- We'll see examples in a bit

# Hypothesis Testing I: $\chi^2$ distribution

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

APRIL 24, 2017

## Goodness of Fit

Suppose we see a die rolled 36 times with the following totals.

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 8 | 5 | 9 | 2 | 7 | 5 |

- $H_0$: fair die
- How far does it deviate from uniform distribution?

## Goodness of Fit

Suppose we see a die rolled 36 times with the following totals.

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 8 | 5 | 9 | 2 | 7 | 5 |

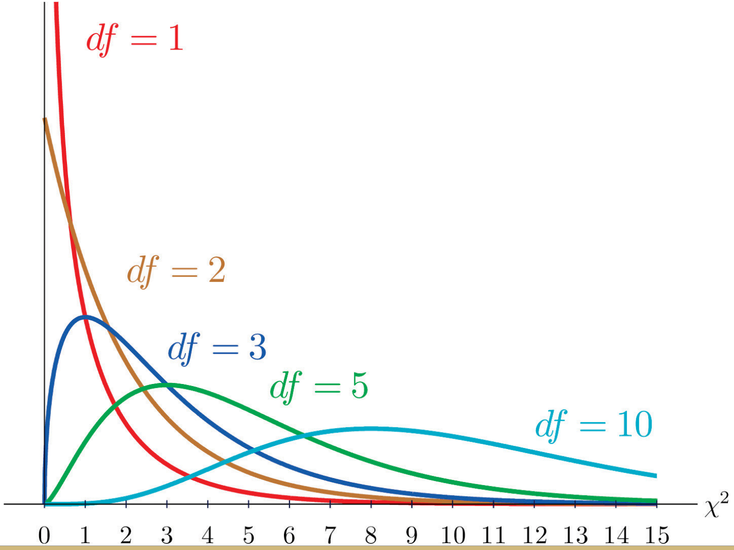
- $H_0$: fair die
- How far does it deviate from uniform distribution?
- $\chi^2$ distribution

## Chi-Square Definition

Let $Z_1, \ldots Z_n$ be independent random variables distributed $N(0,1)$. The $\chi^2$ distribution with $n$ degrees of freedom can be defined by

$$\chi_n^2 \equiv Z_1^2 + Z_2^2 + \cdots + Z_n^2 \tag{1}$$

## Chi-Square Definition



## Chi-Square Distributions

### PDF

$$\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}\exp\{-x/2\}$$

### CDF

$$\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\gamma\left(\frac{n}{2},\frac{x}{2}\right)$$

- $\gamma(s,x) \equiv \int_0^x t^{s-1}\exp\{-t\}\,dt$
- $\Gamma(x) \equiv \int_0^\infty t^{x-1}\exp\{-t\}\,dt$, $\Gamma(n) = (n-1)!$

## Goodness of Fit

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Observed | 8 | 5 | 9 | 2 | 7 | 5 |
| Expected | 6 | 6 | 6 | 6 | 6 | 6 |

- If this were a fair die, all observed counts would be close to expected
- We can summarize this with a test statistic

$$\sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

## Goodness of Fit

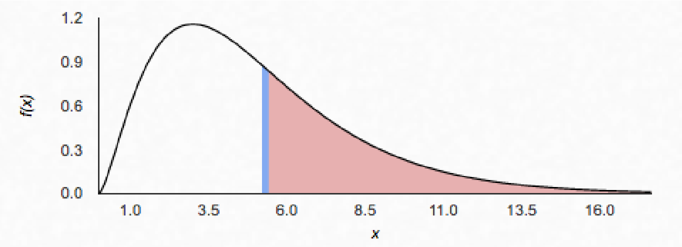
| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Observed | 8 | 5 | 9 | 2 | 7 | 5 |
| Expected | 6 | 6 | 6 | 6 | 6 | 6 |

- If this were a fair die, all observed counts would be close to expected
- We can summarize this with a test statistic

$$\sum \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

- In our example, 5.33
- Approximately distributed as $\chi^2$ with $k-1$ degrees of freedom

## Test Statistic and $p$-value




- Expected value of $\chi^2$ with df=5 is 5
- 5.33 is not that far away
- 0.38 probability of rejecting the null

## Degrees of Freedom

- We condition on the number of observations (36)
- So after filling in the cells for five observations, one is known
- So total of $k-1$ degrees of freedom

## Degrees of Freedom

- We condition on the number of observations (36)
- So after filling in the cells for five observations, one is known
- So total of $k-1$ degrees of freedom
- Important because it specifies which $\chi^2$ distribution to use

# Hypothesis Testing I: $\chi^2$ for collocations

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

APRIL 24, 2017

## Distributional Independence

- If $x$ and $y$ are independent, $P(x,y) = P(x)P(y)$.
- Can we test of two distributions are independent?
- This also is a $\chi^2$ test

## Example: Collocations

- Selectional preferences: “strong tea”, not “powerful tea”
- Phrases: “intents and purposes”, “helter skelter”
- Some words just go together more than others
- I.e., they’re not independent

## Can't use frequency

Most frequent bigrams are just the most frequent words. (Independent distribution.)

| | |
|---|---|
| 80871 | of the |
| 58841 | in the |
| 26430 | to the |
| 21842 | on the |
| 21839 | for the |
| 18568 | and the |
| 16121 | that the |
| 15630 | at the |
| 15494 | to be |
| 13899 | in a |
| 13689 | of a |
| 13361 | by the |

## Contingency tables

| | $w_1 =$ new | $w_1 \neq$ new |
|---|---|---|
| $w_2 =$ companies | 8<br>(new companies) | 4667<br>(e.g., old companies) |
| $w_2 \neq$ companies | 15820<br>(e.g., new machines) | 14287181<br>(e.g., old machines) |

## Contingency tables: degrees of freedom

- Given row and column totals, one cell can fill in the rest (as you did in earlier practice problems)
- In general, for a contingency table with $r$ rows and $c$ columns, $(r-1)(c-1)$ degrees of freedom

## Observed

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 8 | 4667 |
| $w_2 \neq$ **companies** | 15820 | 14287181 |

## Observed

| | $w_1 =$ **new** | $w_1 \neq$ **new** | |
|---|---|---|---|
| $w_2 =$ **companies** | 8 | 4667 | *4675* |
| $w_2 \neq$ **companies** | 15820 | 14287181 | *14303001* |
| | *15828* | *14291848* | *14307676* |

## Expected

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | $\frac{15828}{14307676} \frac{4675}{14307676} \cdot 14307676 = 5.17$ | 1669.83 |
| $w_2 \neq$ **companies** | 15822.83 | 14287178.17 |

**Observed**

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 8 | 4667 |
| $w_2 \neq$ **companies** | 15820 | 14287181 |

**Expected**

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 5.17 | 1669.83 |
| $w_2 \neq$ **companies** | 15822.83 | 14287178.17 |

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \quad (1)$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \quad (2)$$

## Observed

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 8 | 4667 |
| $w_2 \neq$ **companies** | 15820 | 14287181 |

## Expected

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 5.17 | 1669.83 |
| $w_2 \neq$ **companies** | 15822.83 | 14287178.17 |

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{1}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{2}$$

## Observed

| | $w_1 = \textbf{new}$ | $w_1 \neq \textbf{new}$ |
|---|---|---|
| $w_2 = \textbf{companies}$ | 8 | 4667 |
| $w_2 \neq \textbf{companies}$ | 15820 | 14287181 |

## Expected

| | $w_1 = \textbf{new}$ | $w_1 \neq \textbf{new}$ |
|---|---|---|
| $w_2 = \textbf{companies}$ | 5.17 | 1669.83 |
| $w_2 \neq \textbf{companies}$ | 15822.83 | 14287178.17 |

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \qquad (1)$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \qquad (2)$$

## Observed

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 8 | 4667 |
| $w_2 \neq$ **companies** | 15820 | 14287181 |

## Expected

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 5.17 | 1669.83 |
| $w_2 \neq$ **companies** | 15822.83 | 14287178.17 |

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{1}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{2}$$

## Observed

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 8 | 4667 |
| $w_2 \neq$ **companies** | 15820 | 14287181 |

## Expected

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 5.17 | 1669.83 |
| $w_2 \neq$ **companies** | 15822.83 | 14287178.17 |

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{1}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{2}$$

**Observed**

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 8 | 4667 |
| $w_2 \neq$ **companies** | 15820 | 14287181 |

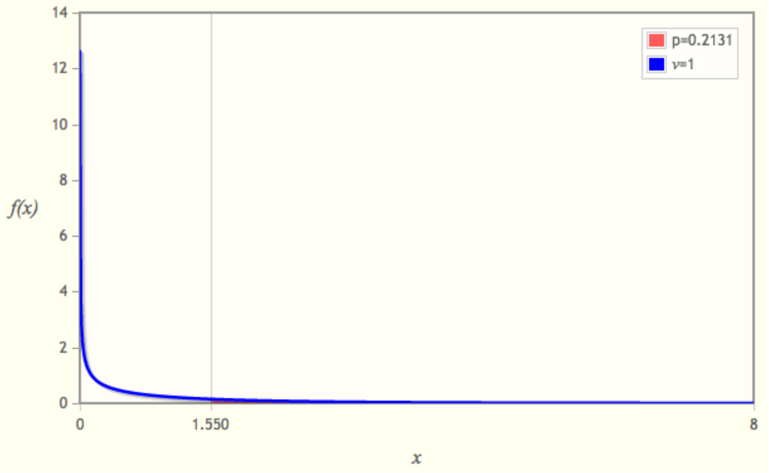
**Expected**

| | $w_1 =$ **new** | $w_1 \neq$ **new** |
|---|---|---|
| $w_2 =$ **companies** | 5.17 | 1669.83 |
| $w_2 \neq$ **companies** | 15822.83 | 14287178.17 |

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \quad (1)$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} = 1.55 \quad (2)$$

## Can we reject the null?



# Hypothesis Testing I: Limitations

## INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

APRIL 24, 2017

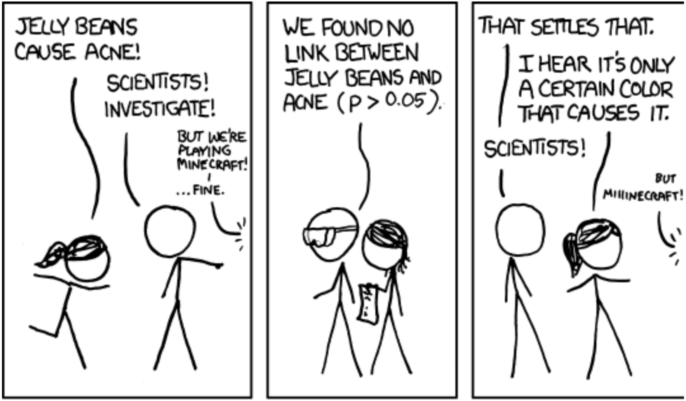
## $\chi^2$ is not exact

- $\chi^2$ is not exact
- Should not use if any cells are $< 5$
- Fischer's exact test (hypergeometric distribution)

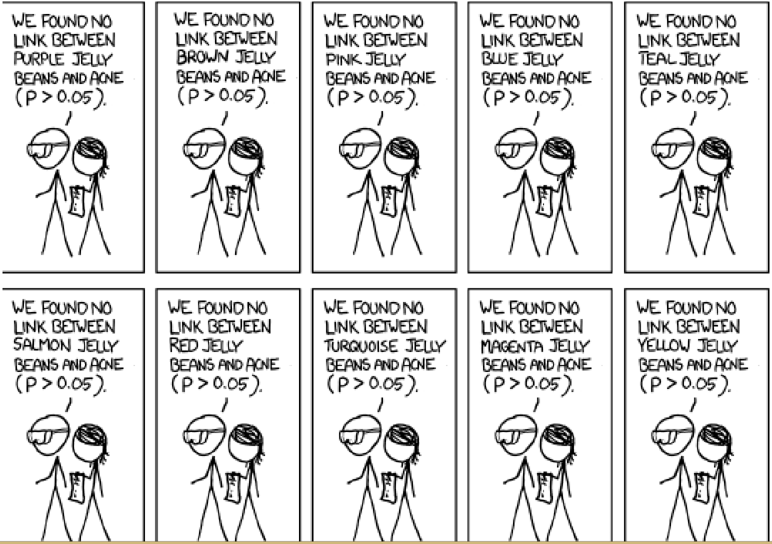
| | |
|---|---|
| a | b |
| c | d |

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;b!\;c!\;d!\;n!}$$

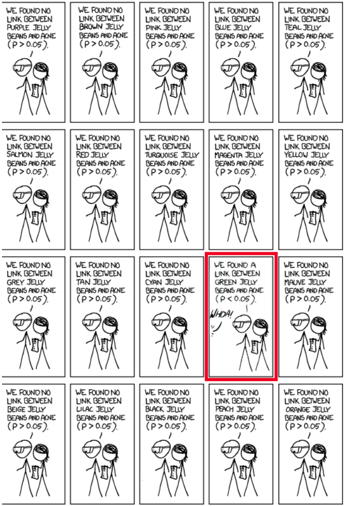
## $p$-hacking


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 MIIINECRAFT!

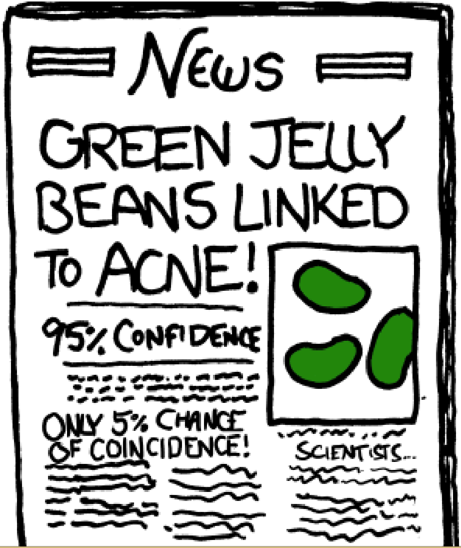
## *p*-hacking


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).

# $p$-hacking


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 MAUVE 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).

## *p*-hacking


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

## Bonferroni Correction

- If you conduct multiple statistical tests, you must divide $\alpha$ by number of tests
- If you have $m$ tests and reject null at 0.05 for any of them, chance of Type I error is multiplied by $m$