Statistical Tests

Jordan Boyd-Graber

University of Maryland

Fall 2020

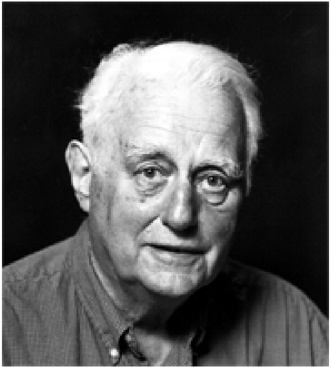
What's Necessary for (Data/Information/Computer) Science: Scepticism

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

What's Necessary for (Data/Information/Computer) Science: Scepticism

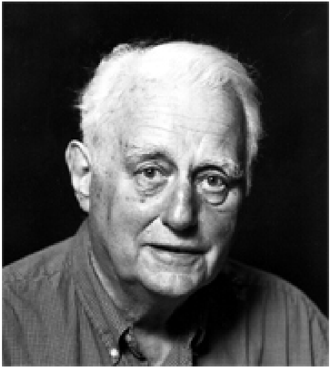
- We've assumed
  - Our models are right
  - Our parameter estimates are good
- Not always true
  - Learning the mindset
  - Not trusting your data
  - Communicating uncertainty
  - How do we know if distributions / parameters are any good?

Lincoln Moses



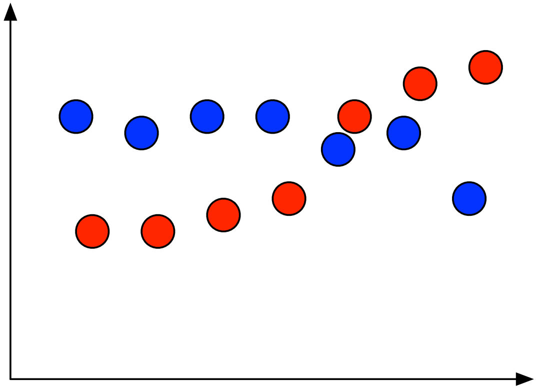
- Stanford Statistician
- Learn one thing: Use Error Bars

Lincoln Moses

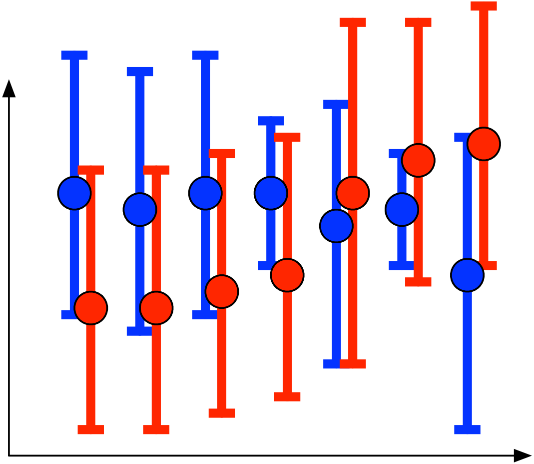


- Stanford Statistician
- Learn one thing: Use Error Bars
- After visiting US government: Use data

Point Estimates Lie



Point Estimates Lie



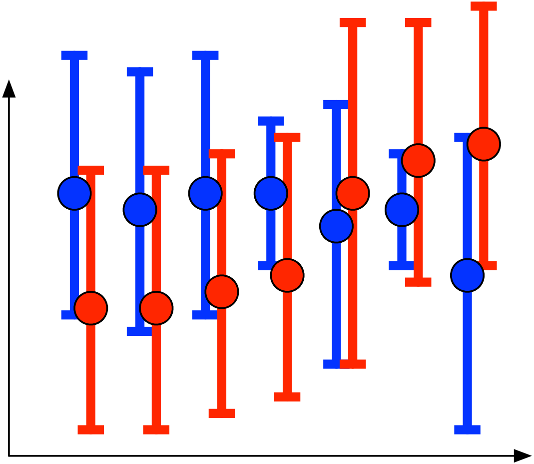
So how can you make a decision?

- 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

Lingo

- Confidence interval
- Null hypothesis
- test statistic
- *p*-value
- *p*-hacking

Confidence Intervals



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?

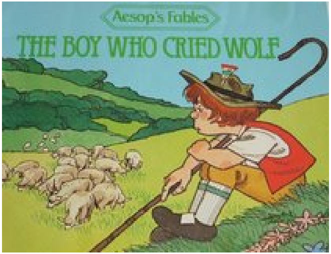
Test Statistic

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

What can happen

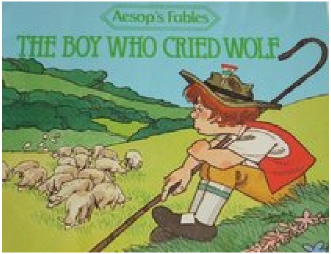
Measured / Perceived	Reality: True	Reality: False
True	Correct ☺	Type I False Positive
False	Type II False Negative	Correct ☺

Boy who cried wolf



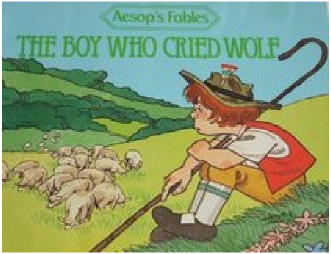
- Null hypothesis (status quo): no wolf

Boy who cried wolf



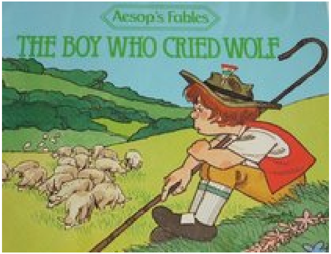
- Null hypothesis (status quo): no wolf
- First error, Type I: villagers believed there was wolf (but there wasn't), **False Positive**

Boy who cried wolf



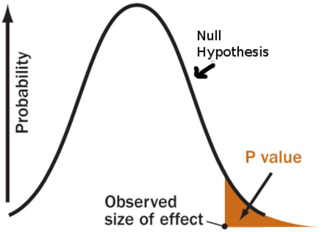
- Null hypothesis (status quo): no wolf
- First error, Type I: villagers believed there was wolf (but there wasn't), **False Positive**
- Second error, Type II: villagers believed there was no wolf (when there was), **False Negative**

Boy who cried wolf



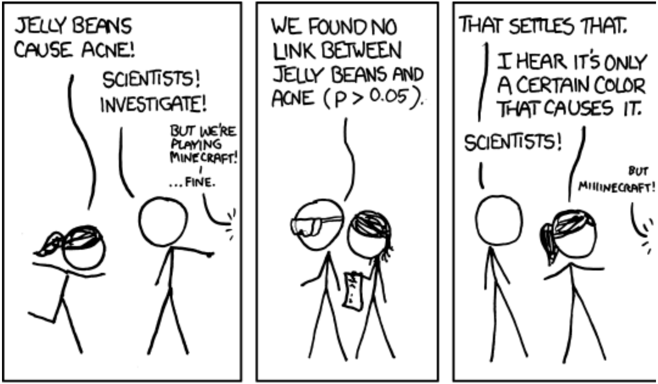
- Null hypothesis (status quo): no wolf
- First error, Type I: villagers believed there was wolf (but there wasn't), **False Positive**
- Second error, Type II: villagers believed there was no wolf (when there was), **False Negative**
- The villagers had Type I and Type II in that order

p-value



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

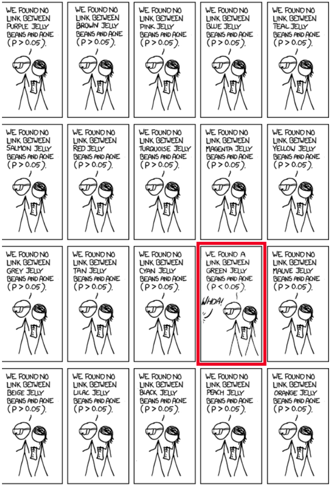
p-hacking


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

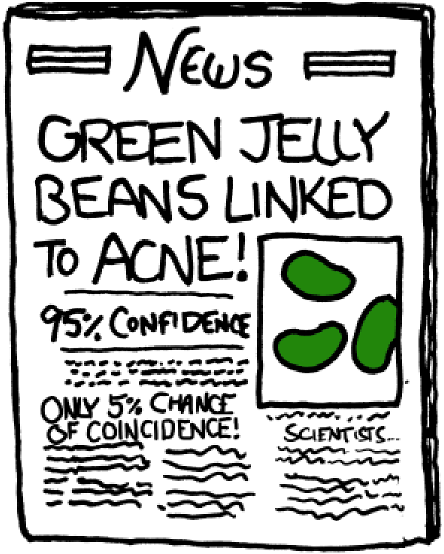
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND 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).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE

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 α 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

What does this have to do with deep learning / natural language processing / what I care about?

- You collect a lot of data
- Run a bunch of experience
- There's some natural varience
  - How do you know if what you did is better?
  - How do you know if two populations are different?
  - Modern methods often have hundreds or thousands of experiments

Courses, Lectures, Exercises and More

http://boydgraber.org

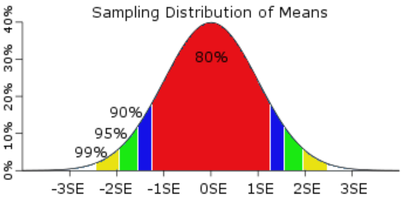
Statistical Tests

Jordan Boyd-Graber

University of Maryland

Fall 2020

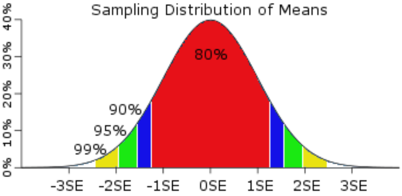
Normal Distribution Confidence Interval



- You observe $\{x_1 \dots x_N\}$
- Obtain mean $\bar{x}$
- Sample standard deviation (standard deviation is square root of variance σ^2)

$$S = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}} \tag{1}$$

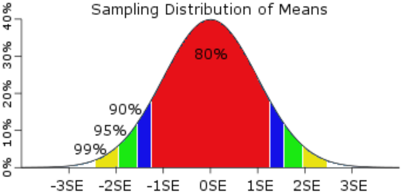
Normal Distribution Confidence Interval



- You observe $\{x_1 \dots x_N\}$
- Obtain mean $\bar{x}$
- Sample standard deviation (standard deviation is square root of variance σ^2)

$$S = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}} \tag{1}$$

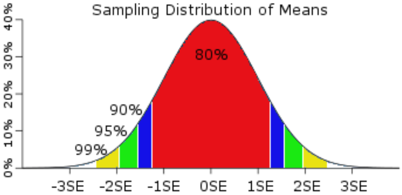
Normal Distribution Confidence Interval



- You observe $\{x_1 \ldots x_N\}$
- Obtain mean $\bar{x}$
- Sample standard deviation (standard deviation is square root of variance σ^2)

$$S = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}} \tag{1}$$

Normal Distribution Confidence Interval



- You observe $\{x_1 \ldots x_N\}$
- Obtain mean $\bar{x}$
- Sample standard deviation (standard deviation is square root of variance σ^2)

$$S = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}} \tag{1}$$

Example Data

```
   Name,Birth Date,Inaug,End,Age
1  George Washington,"Feb 22, 1732","Apr 30, 1789","Mar 4, 1797",57
2  John Adams,"Oct 30, 1735","Mar 4, 1797","Mar 4, 1801",61
3  Thomas Jefferson,"Apr 13, 1743","Mar 4, 1801","Mar 4, 1809",57
4  James Madison,"Mar 16, 1751","Mar 4, 1809","Mar 4, 1817",57
5  James Monroe,"Apr 28, 1758","Mar 4, 1817","Mar 4, 1825",58
```

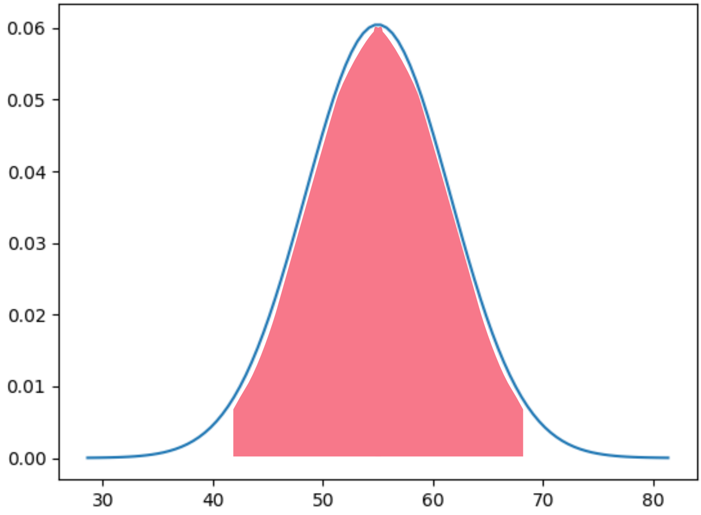
President CI

Pandas: For reading data from CSV file

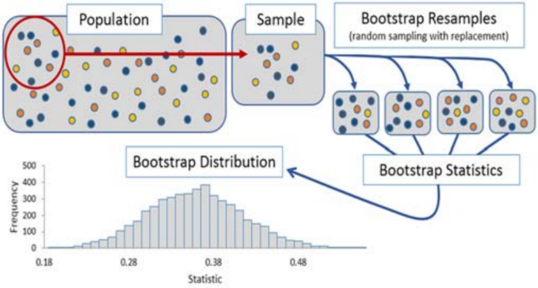
```
import pandas
import numpy
from scipy import stats
import matplotlib.pyplot as plt
```

President CI

```
if __name__ == "__main__":
    p = pandas.read_csv("../data/presidents.csv")

    # Compute sample standard deviation
    mu = numpy.mean(p["Age"])
    s = numpy.std(p["Age"], ddof=1)

    print(stats.norm.interval(0.95, loc=mu, scale = s))

    # Plot distribution
    x = numpy.linspace(mu - 4*s, mu + 4*s, 100)
    plt.plot(x, stats.norm.pdf(x, mu, s))
    plt.show()
```


0.06
0.05
0.04
0.03
0.02
0.01
0.00
30
40
50
60
70
80

Bootstrap Sample



(From Banjanovic and Osborne)

- Compute CI of more complicated distributions
- Example: Effect of Tweets on DL system
  - You have 10k tweets
  - Sample 10k tweets with replacement
  - Train complicated system
  - Repeat
  - Compute CI using the result

Courses, Lectures, Exercises and More

http://boydgraber.org

Statistical Tests

Jordan Boyd-Graber

University of Maryland

Fall 2020

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

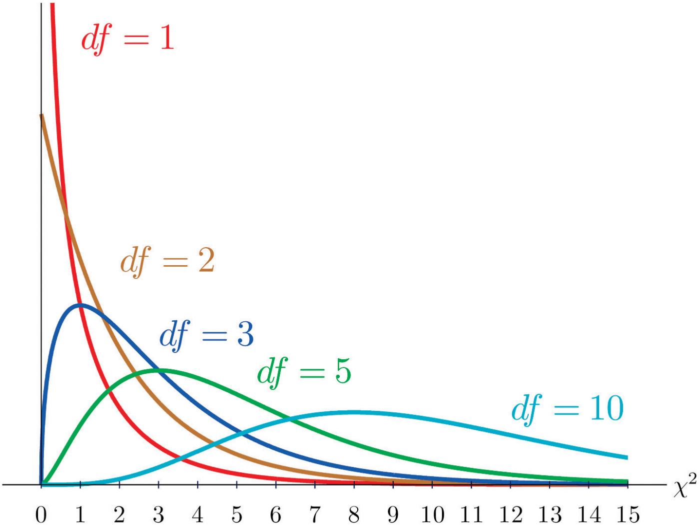
- H_0: fair die
- How far does it deviate from uniform distribution?
- χ^2 distribution

Chi-Square Definition

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

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

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

$$\sum \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

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

$$\sum \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

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

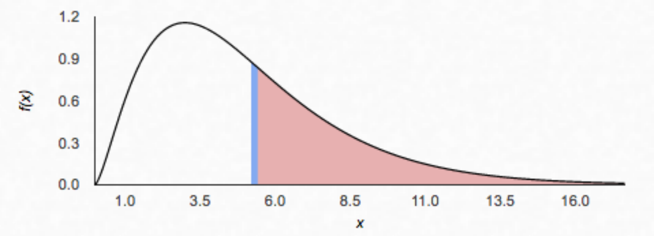
Degrees of Freedom

- We condition on the number of observations (36) into each of the cells (one for each type of observation)
- So after filling in the cells for five observations, one is known
- So total of $k-1=5$ degrees of freedom

Degrees of Freedom

- We condition on the number of observations (36) into each of the cells (one for each type of observation)
- So after filling in the cells for five observations, one is known
- So total of $k-1=5$ degrees of freedom
- Important because it specifies which χ^2 distribution to use

Test Statistic and p-value



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

Independence

Random variables X and Y are *independent* if and only if
$P(X = x, Y = y) = P(X = x)P(Y = y)$.
Mathematical examples:

- If I flip a coin twice, is the second outcome independent from the first outcome?

Independence

Random variables X and Y are *independent* if and only if $P(X=x, Y=y) = P(X=x)P(Y=y)$.

Mathematical examples:

- If I flip a coin twice, is the second outcome independent from the first outcome?

- If I draw two socks from my (multicolored) laundry, is the color of the first sock independent from the color of the second sock?

Independence

Intuitive Examples:

- Independent:
    - you use a Mac / the Green Line is on schedule
    - snowfall in the Himalayas / your favorite color is blue

Independence

Intuitive Examples:

- Independent:
  - you use a Mac / the Green Line is on schedule
  - snowfall in the Himalayas / your favorite color is blue
- Not independent:
  - you vote for Larry Hogan / you are a Republican
  - there is a traffic jam Baltimore / there's a home game

Independence

Sometimes we make convenient assumptions.

- the values of two dice (ignoring gravity!)
- whether it is raining and the number of taxi licenses
- whether it is raining and the amount of time it takes me to hail a cab
- the first two words in a sentence

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 χ^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 to find Collocations

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

	$w_1 = \text{new}$	$w_1 \neq \text{new}$
$w_2 = \text{companies}$	8 (new companies)	4667 (e.g., old companies)
$w_2 \neq \text{companies}$	15820 (e.g., new machines)	14287181 (e.g., old machines)

Joint distribution

- Typically, we consider collections of random variables.
- The *joint distribution* is a distribution over the configuration of all the random variables in the ensemble.
- For example, imagine flipping 4 coins. The joint distribution is over the space of all possible outcomes of the four coins.

$$
\begin{aligned}
P(HHHH) &= 0.0625 \\
P(HHHT) &= 0.0625 \\
P(HHTH) &= 0.0625 \\
&\dots
\end{aligned}
$$

- You can think of it as a single random variable with 16 values.

Marginalization

If we know a joint distribution of multiple variables, what if we want to know the distribution of only one of the variables?

We can compute the distribution of $P(X)$ from $P(X, Y, Z)$ through *marginalization*:

$$\sum_y \sum_z P(X = x, Y = y, Z = z) = P(X)$$

Marginalization

If we know a joint distribution of multiple variables, what if we want to know the distribution of only one of the variables?

We can compute the distribution of $P(X)$ from $P(X, Y, Z)$ through *marginalization*:

$$\sum_y \sum_z P(X = x, Y = y, Z = z) = P(X)$$

We'll explain this notation more next week for now the formula is the most important part.

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X, Y) = \sum_z P(X, Y, Z = z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather
- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold

- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold

- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15		

- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X, Y) = \sum_z P(X, Y, Z = z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

W=Sunny
W=Cloudy

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

W=Sunny
W=Cloudy

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

W=Sunny	.40
W=Cloudy	

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

W=Sunny	.40
W=Cloudy	.60

Contingency tables

	$w_1 = \text{new}$	$w_1 \neq \text{new}$
$w_2 = \text{companies}$	8 (new companies)	4667 (e.g., old companies)
$w_2 \neq \text{companies}$	15820 (e.g., new machines)	14287181 (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 first quiz)
- 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

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{4}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{5}$$

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

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{4}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{5}$$

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

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{4}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{5}$$

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

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{4}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{5}$$

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

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{4}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} \tag{5}$$

Observed

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

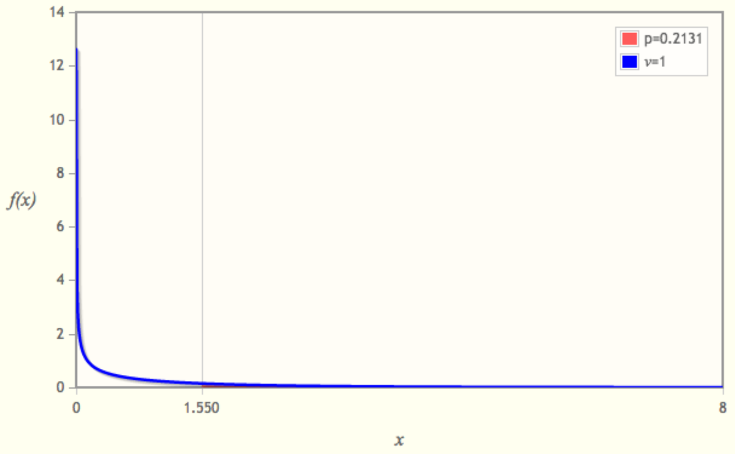
Expected

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

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-1669.83)^2}{4667} + \frac{(15820-15822.83)^2}{15820} \tag{4}$$

$$+ \frac{(14287181-14287178.17)^2}{14287181} = 1.55 \tag{5}$$

Can we reject the null?



Courses, Lectures, Exercises and More

http://boydgraber.org

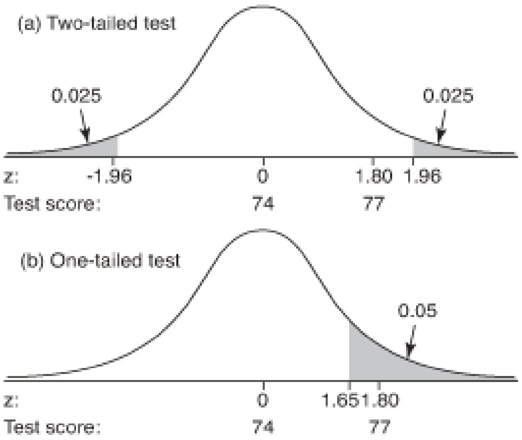
Statistical Tests

Jordan Boyd-Graber

University of Maryland

Fall 2020

Two-tailed vs. one-tailed tests



- Two tail: Alternative $\mu \neq \mu_0$
- One tail: Alternative $\mu > \mu_0$

What if you don't know variance?

- *t*-test allows you to test hypothesis if you don't know variance
- Sometimes called "small sample test": same as *z* test with enough observations
- William Gossett: check that yeast content matched Guiness's standard (but couldn't publish)
- I.e., checking whether yeast content equal to μ_0

t-test statistic

- Need to estimate variance

$$s^2 = \sum_i \frac{(x_i - \bar{x})^2}{N-1} \tag{6}$$

- $n-1$ removes bias (expected value is less than truth)
- Test statistic looks similar

$$T \equiv \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{N}}} \tag{7}$$

Degrees of Freedom

- Like χ^2, *t*-distribution parameterized by degrees of freedom
- $\nu = N - 1$ degress of freedom

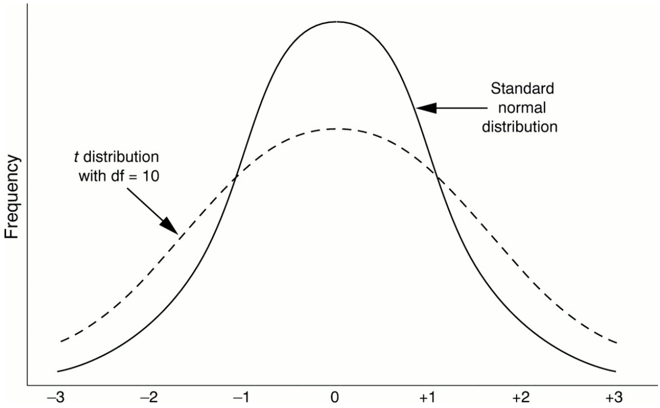
PDF

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{8}$$

CDF

$$\frac{1}{2}+x\Gamma\left(\frac{\nu+1}{2}\right)\frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \tag{9}$$

Shape of *t*-distribution



Example

- Suppose observe $\{0, 1, 2, 3, 4, 5\}$
- Test whether $\mu \neq 1$

Example

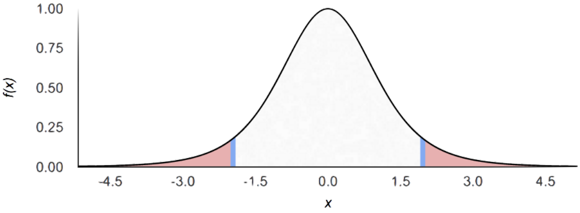
- Suppose observe $\{0,1,2,3,4,5\}$
- Test whether $\mu \neq 1$
- $\bar{x} = 2.5$, $s^2 = 3.5$

Example

- Suppose observe $\{0, 1, 2, 3, 4, 5\}$
- Test whether $\mu \neq 1$
- $\bar{x} = 2.5$, $s^2 = 3.5$
- $T = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} = \frac{2.5 - 1.0}{\sqrt{\frac{3.5}{6}}} = 1.9640$

Example

- Suppose observe $\{0, 1, 2, 3, 4, 5\}$
- Test whether $\mu \neq 1$
- $\bar{x} = 2.5$, $s^2 = 3.5$
- $T = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} = \frac{2.5 - 1.0}{\sqrt{\frac{3.5}{6}}} = 1.9640$
- Double area under the at two tailed CDF



$\mu = E(X) = 0 \quad \sigma = SD(X) = 1.291 \quad \sigma^2 = Var(X) = 1.667$