Testing for Significance

What does it mean?

Null hypothesis testing (NHT)

- Recall from beginning of the semester when we wanted to rule out that chance (randomness) is responsible for the observed correlation

- To do so, we undertake significance testing

- Significance testing allows us to judge the likelihood of observing a sample as extreme without the hypothesized relationship being true

Hypotheses

- Null hypothesis: generally a hypothesis about a null finding (i.e. no) relationship — i.e. due to chance

Hypotheses

- Null hypothesis: generally a hypothesis about a null finding (i.e. no) relationship — i.e. due to chance

- In research we generally formulate a(n) (alternative) hypothesis and then test is against the null

Hypotheses

- Null hypothesis: generally a hypothesis about a null finding (i.e. no) relationship — i.e. due to chance

- In research we often formulate a(n) (alternative) hypothesis and then test is against the null

- Example:

    - Hypothesis: Americans under 40 years old are more likely to be registered Democrats, than Americans over 40.

    - Null hypothesis: Americans under 40 years old are NOT more likely to be registered Democrats, than Americans over 40.

- Note that our hypotheses are about the population (box) while the statistics we observe are about samples (draws)

Hypotheses

- Null hypothesis: generally a hypothesis about a null finding (i.e. no) relationship — i.e. due to chance

- In research we often formulate a(n) (alternative) hypothesis and then test is against the null

- Any hypothesis testing is done in an “inverted” manner, i.e. the significance test indicates how likely it is that we observe data this extreme if the NULL hypothesis is correct

Hypotheses

- So, even though we are interested in finding evidence for our argument, we do so by testing **against** a **null hypothesis**

- The p-value (test result) is the chance of getting the sample statistics as extreme as the one observed given that the null hypothesis is correct.

- If the p-value is very small (generally smaller than 0.05), then we reject the **null**

Hypotheses

- If the p-value is very small (generally smaller than 0.05), then we reject the **null**

- A rejection of the null hypothesis is generally seen as support for the alternative hypothesis

- **BUT**, we **never** prove our hypothesis or interpret p-value as probability that hypothesis is correct

Z Test statistic

- (Observed value - hypothesis value)/SE = z-stat

- Z-stat is value of how many SEs the observed value is away from the expect value

- Any idea where this is coming from? Similarity?

Z Test statistic

- (Observed value - hypothesis value)/SE = z-stat
- Z-stat is value of how many SEs the observed value is away from the expect value
- We take this to the z-table -> area under the curve outside of z divided by 2 (one direction)

Test statistic

The observed significance level is the chance of getting a test statistic as extreme as, or more extreme than, the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller this chance, the stronger the evidence against the null.

Example

Can people guess the number a random number generator generates? 15 subjects do 500 guesses, I.e. total of 7500 guesses. What is the null?

Example

Can people guess the number a random number generator generates? 15 subjects do 500 guesses, I.e. total of 7500 guesses. What is the null? Expected value if guesses are just as random… i.e. should be correct 1/4 of time

Expected value under null: 7500/4 = 1,875

Test statistic = (observed - 1875)/SE

Example

Can people guess the number a random number generator generates? 15 subjects do 500 guesses, I.e. total of 7500 guesses. What is the null? Expected value if guesses are just as random… i.e. should be correct 1/4 of time

Expected value under null: 7500/4 = 1,875

We observed 2,006 correct guesses

Test statistic = (2,006 - 1875)/SE

SE: as if 7500 draws from box 0,0,0,1 ->
sqrt(7500)*sqrt(0.25*0.75) = 37

Example

Can people guess the number a random number generator generates? 15 subjects do 500 guesses, I.e. total of 7500 guesses. What is the null? Expected value if guesses are just as random… i.e. should be correct 1/4 of time

Expected value under null: 7500/4 = 1,875

We observed 2,006 correct guesses

Test statistic = (2,006 - 1875)/37 = 3.5

P-value: 2/10,000 -> evidence against the null!

Significance Testing

(Observed - expected)/SE

Very similar to normal approximation and confidence intervals

Make sure SE is appropriate for statistics you are interested in, i.e. percentage, average, etc

Remember to divide area under curve for z-value by 2

Who thinks Sumlin should have been fired?

- Null hypothesis: 70% of TAMU students approve of firing
- Write 1 = firing, 0 = should’ve kept him

Remember

- Generally we reject the null hypothesis if the p-value is less than 5%
- The area under the curve is already in percent
- So a result of 0.001 is actually 0.001 of 1%

Exercises

A) other things being equal, which of the following p-values is best for the null hypothesis?

0.1 of 1%. 3%. 17% 32%

B) Repeat, for the alternative hypothesis.

Exercises

One hundred draws are made at random with replacement from a box. The average of the draws is 22.7 and the SD is 10. Someone claims that the average of the box equals 20. Is this plausible? Do a significance test.

Exercises

According to a investigator's model, the data are like 400 draws made at random from a large box. The null hypothesis says that the average is 50. The alternative says that the average of the box is more than 50. In fact, the data averaged out to 52.7 and the SD was 25. What should we conclude?

Z-stat

- Based on the normal approximation, i.e. still needs large number of draws and/or approximately close to normal histogram
- Need null hypothesis in terms of box model for the data
- Difference between data and what is expected under null
- SE for stat of interest
- Compute significance level

Exercise

As part of a statistics project in the early 1970s, Mr. Frank Albert approached the first 100 students he saw one day on Sproul Plaza at the University of California, Berkeley and fund out the school or college in which they were enrolled. There were 53% men in this sample. From Registrar's data, 25,000 students were registered at Berkeley that term, and 67% were male. Was his sampling procedure like taking a simple random sampling?

Fill in the blanks. This will lead us to fill the box model for the null hypothesis.

A) There is one ticket in the box for each _____ (student registered at Berkeley or person in sample)

B) The ticket is marked ____ for men and ___ for women.

C) The number of tickets in the box is ____ and the number of draws is ____

D) The null hypothesis says the sample is like _____ _draws made at random from the box.

E) The percentage of 1's in the box is _____ .

F) Observed number of men is ____

G) Expected number of men is ____

T-statistics

- When we don’t have enough data, the normal approximation for the z-statistic is not appropriate

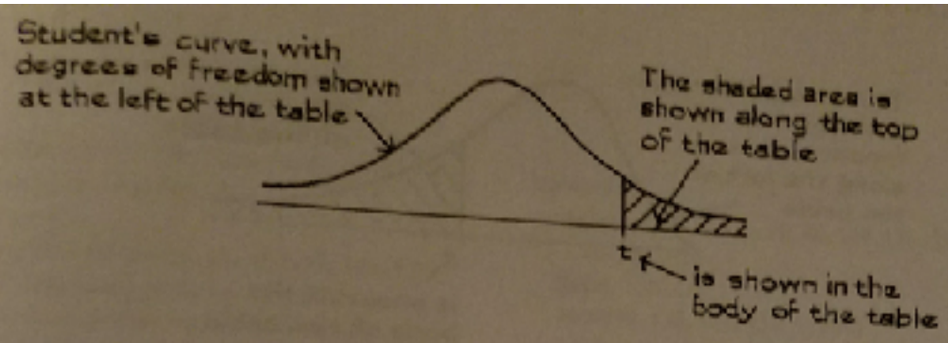
- So we rely on a different distribution for our inference: the student’s t distribution

T-statistics

- The test statistic is still (observed - expect)/SE
- But, we can’t reliably calculate a SD with so few observations
- So we approximate the SD with SD+ ->>>>
    - sqrt(# observations/(# observations - 1)) * SD
    - Similar to earlier correction factor for without replacement

T-statistics

- The test statistic is still (observed - expect)/SE+
- Now instead of the z-table in the back, we need to use the student's curve (table for student T distribution)
- This table requires another part, however, degrees of freedom (DF)
- Degrees of freedom are the number of data points (observations) - 1



Degrees of freedom	*25%*	*10%*	*5%*	*2.5%*	*1%*	*0.5%*
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.41	1.89	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.05
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.35	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95
16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.85
21	0.69	1.32	1.72	2.08	2.52	2.83
22	0.69	1.32	1.72	2.07	2.51	2.82
23	0.69	1.32	1.71	2.07	2.50	2.80
24	0.68	1.32	1.71	2.06	2.49	2.80
25	0.68	1.32	1.71	2.06	2.49	2.79

T-statistics

- When number of observations is small (less than say 25) and SD of population is unknown, use SD+ to approximate SD
    - SD+ = SD * sqrt(# observations /(# observations - 1))
- Calculate DF = # of observations -1
- Use t-table to find p-value

- What is the area under the student's t-curve with 5 degrees of freedom

  - To the right of 2.02?
  - Between - 2.02 and 2.02?
  - To the left of 2.02?

Exercise

True or False, to make a t-test with 4 measurements, use the Student's t curve with 4 degrees of freedom

Exercise

Five readings of the pressure of a football:

10.5, 11.1, 11, 11.5, 10.9, pounds per square inch

By rules NFL footballs have to be between 12.5 and 13.5 pounds per square inch. Are these low measurements due to chance? Can we reject the null hypothesis that the measured ball had a pressure of 12.5 pounds per square inch?

Comparing Averages

- Often times we are interested in figuring out if two groups are really different from one another

- For example, let's say we compare the height of women and men in the class

- Are these significantly different from each other?

Comparing Averages

- More so, for experiments we might want to compare control and treatment group on some outcome

- For example, one groups gets the placebo pill and one group gets a weight loss pill. Are the groups significantly different from each other after the treatment period?

Comparing Averages

- To compare averages of two groups, we first calculate their difference
    - E.g.
        - Avg. height of men 1.8m
        - Avg. height of women 1.65 m
        - DIfference 0.15 m
- But to figure out whether this is significantly different from zero (our null hypothesis), we need what?

Comparing Averages

- To compare averages of two groups, we first calculate their difference
    - E.g.
        - Avg. height of men 1.8m
        - Avg. height of women 1.65 m
        - Difference 0.15 m
- But to figure out whether this is significantly different from zero (our null hypothesis), we need what? A standard error for the difference!

Comparing Averages

- To compare averages of two groups, we first calculate their difference
  - E.g.
    - Avg. height of men 1.8m
    - Avg. height of women 1.65 m
    - Difference 0.15 m
- The standard error for the difference:
    - Sqrt(a^2 + b^2) where a and b are the SEs of the respective groups
    - Sqrt(2^2 + 1.5^2) = sqrt(6.25) = 2.5

Comparing Averages

- When can we compare averages like this?

- The groups have to be independent from each other!

Comparing Averages - Example

- Box A has an average of 100 and SD of 10, Box B has an average of 50 and a SD of 18. Now 25 draws are made with replacement from box A and independently 36 draws are made at random with replacement from box B. Find the expected value and SE for the difference between the averages.

Comparing Averages - Example

- A coin is tossed 500 times. Find the expected value and SE for the difference between the percentage of heads in the first 400 and last 100 tosses.

Comparing Averages - significance

- Once we have the difference in averages and the SE for the difference we can use the z-test to test for significance against an expected value

- Usually the expected value for the difference is zero - hence null hypothesis

Comparing Averages - Experiments

- In experiments we usually have one subject pool from which we randomly sample a number of participants to be in the treatment group, the rest of the pool is in the control group

- Is this a problem for comparing the averages and testing for significance? Why?

Comparing Averages - Experiments

1. Draws are made without replacement!

2. Groups are not independent!

- Yet, the SE calculated as before is approximately correct, since mistake 1 inflates the SE and mistake 2 decreases the SE

- Essentially the two mistakes cancel each other out

Exercise

A study is done on elementary school students. 499 students agree to participate. After the midterm 250 are randomly selected into the treatment group and 249 into the control group. The treatment group is fed wheaties for breakfast, the control group gets sugar pops.

A) final scores averaged 66 for the treatment group (SD = 21). For the control group the average was 59 (Sd = 20). What do you conclude?

B) What aspect of the study could have been done “blind”?

Exercise

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

By MARIANNE BERTRAND AND SENDHIL MULLAINATHAN*

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (*JEL* J71, J64).

Exercise

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

By MARIANNE BERTRAND AND SENDHIL MULLAINATHAN*

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (*JEL* J71, J64).

- Let's look at the data!

Resume i	Black-sounding name sounding name T_i	Callback $Y_i(1)$	Callback $Y_i(0)$	Age	Education
1	1	1	?	20	College
2	0	?	0	55	High school
3	0	?	1	40	Graduate school
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	0	?	62	College

Exercise

- Let’s look at the data!
- 2435 resumes with black sounding names
- 2435 resumes with white sounding names
- Average percent of call backs:
    - white: 9.6%, SD = 0.245
    - black: 6.4%, SD = 0.295