POLS201 Spring 2019

Basic Hypothesis Testing

March 13

Agenda

- We're going to consider hypothesis testing and try it with R
- Log into your RStudio Cloud account and move into the Lewis & Clark workspace
- You can see these slides from the POLS201_Spring_2019_190313.pdf
- You will see a live example of data loaded raw from the Pew survey.

We can use our mastery of Normal Distributions to Test Hypotheses

- Suppose we compare two different values even though they were produced by the same data generating process.
- Or, they came from the same distribution
- Or, they are really different and weren't generated from the same distribution.
  - Which means: they represent two things that are truly different.
- How do we know?

Hypothesis Testing

- Null Hypothesis: A statement about what we would expect to observe if our values came from the same distribution.
- Alternative Hypothesis: A statement about what we would expect to observe if our values did not come from the same distribution.

Null Hypotheses:

- There is no difference in population means
- $H_o : \mu = \mu$ or $H_o : \mu - \mu = 0$
- H_o: There is no relationship between our independent and dependent variable.
- H_o: corr(x, y) $= 0$
- We may see some correlation in our samples, or not, but we conclude the populations have none.

Alternative Hypotheses:

- There is meaningful difference in population means
- $H_a : \mu \neq mu$ or $H_o : \mu - \mu \neq 0$
- H_a: There is a relationship between our independent and dependent variable.
- H_a: corr(x, y) $\neq 0$

Conventions

- By convention, when we engage in hypothesis testing, we seek to "reject the null hypothesis"
- We never accept the alternative hypothesis until demonstrated
- The burden of proof lies with the alternative hypothesis
- Note that we are generally interested in two-tailed tests, which as a practical matter, are harder to pass.

Different data require different tests

Table 8.1. Variable types and appropriate bivariate hypothesis tests

		Independent variable type	
		Categorical	Continuous
Dependent variable type	Categorical	*Tabular analysis*	Probit/logit
	Continuous	*Difference of means*	*Correlation coefficient*; bivariate regression model

Notes: Tests in italics are discussed in this chapter.

Different data require different tests

- All tests have a common logic: You are comparing the identified relationship of X and Y to what we would expect to see if NO relationship existed.
- Express Null and Alternative Hypotheses
- Central Limit Theorem gives us the parameters of what we would expect to see “by chance”

Another Convention

- An old fashioned convention says: Would we expect to see the findings we got less than 5% of the time?
- A P-value tell us this result.
- But beware. . . .there is no magic threshold that "proves" we have a relationship.

P-Values

- Tells us the probability we would expect to see a given relationship because of random chance
- Limitations:
- Does not tell us if the relationship is causal!
- Does not tell us the STRENGTH of the relationship if it exists.
- It merely says: "we have some basis for rejecting the null hypothesis"

How could we use this idea in a simple research project?

- Draw a sample of Facebook users
- Randomly assign users to control and treatment group
- Give treatment group a political ad. No ad for control group.
- Measure voter turnout of control and treatment groups.
- Could the mean turnout of the treatment group have been pulled from the same distribution, or is it different?

How could we use this idea in a simple research project?

- REMEMBER: The central limit theorem is the backbone for all hypothesis testing!
- We know how much random error will affect estimates. It's simply a function of standard deviation and sample size!
- We ask: "If they were pulled from the same distribution, what size of differences could be explained by mere randomness?"

Difference of Means Testing: Two Varieties

- Often, we can analyze sample means for an answer to this question. We might ask:
- Is a sample mean different than a population mean?
- Are two sample means different from each other?

Sample VS Population Means

- Say you are told told that $\mu = 12$

Sample VS Population Means

- Say you are told told that $\mu = 12$
- You are given a sample of 100 with a sample mean of 10 and a sample standard deviation of 4.

Sample VS Population Means

- Say you are told told that $\mu = 12$
- You are given a sample of 100 with a sample mean of 10 and a sample standard deviation of 4.
- You could build a confidence interval with this information OR calculate Z!

Let's ask:

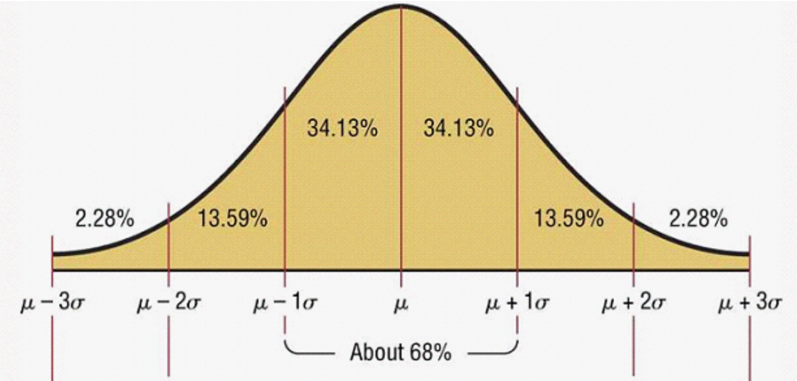
- "How often would our sample distribution emerge from a population of with a mean of 12?"
- Our sample has a mean ($\bar{x}$) of 10 and an s of 4.
- Is it plausible that a distribution with $\mu = 12$ produced our sample?
- What are the ranges of plausible population means? How often would we see the sample mean of 10 produced from a population with mean of 10? Or 5? Or 20?

We compute the *confidence interval*

- The confidence interval defines _____ % of the possible population means that could have produced this result.
  - Often, we define this range as 95%.
- What is the range of means that defines these boundaries?

Observe the standard deviation of 2

- Notice that slightly more than 95% of the curve covers the area +/- 2 standard deviations.
- With 1.96 standard deviations, we have covered 95% exactly.



How many standard deviations away is our population mean?

- A Z score of +/- 1.96 will define those boundaries.
- The formula is this:

$$Z = \frac{(x-\mu)}{\sigma \div \sqrt{n}}$$

- For our example, the 95% interval is 11.216 to 12.784.
- Does 10 fall inside or outside that interval?
- Notice that we're figuring out a standard error, so we include $\sqrt{n}$ in the denominator.

Remember: We apply the standard deviation of the sampling distribution

- Bigger samples will narrow the range of a confidence interval. Do you see why?

$$Z = \frac{(x-\mu)}{\sigma \div \sqrt{n}}$$

- In our example:

$$Z = (10 - 12)/(4/10) = -5$$

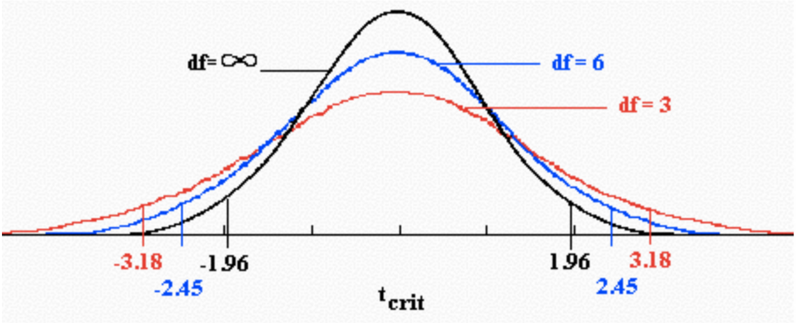
- A Z score of -5 is completely out of bounds. Which means: it's extremely unlikely that a population with a mean of 12 would have produced our sample!

Usually we compare two samples, instead of a sample v. population

- So we make a slight adjustment from our normal distribution
- We use the *Student-t Distribution* and compute a t-score instead of a Z-score.
- As the sample size increases, the t-distribution approaches the shape of a normal distribution.

The look of a t-distribution

- But when the sample size is small, the t-distribution is a bit flatter than normal, with fatter tails.



Interpretation of t

- We use t when we compare two *samples*
- The interpretation of t is identical to the interpretation of Z.
- High (or low) t values represent a difference that cannot be explained by chance alone.
- The "critical" t-value will vary on sample size, but with very large sample sizes it is precisely $+/-$ 1.96
- Considering sample size complicates the formula, however

Difference of Means and the T-Test

- With two samples, we ask a different question: what is the likelihood both samples came from the same population
- We can analyze this likelihood by comparing means, and using standard errors and sample sizes in the formula
- That formula is:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^{\,2}}{N_1} + \frac{S_2^{\,2}}{N_2}}}$$

But Use the Computer to Calculate a Difference of Means Test

- We want to suppose a distribution of means differences.
- In other words, a distribution of $\bar{x_1} - \bar{x_2}$
- If the samples came from the same source, what is the confidence interval of this difference?
- Or: is the true difference of means $= 0$?
- At some point, the differennce of means is too great, the same source couldn't have produce both samples.

Example

- A random sample of 300 men has an income of \$32,800 (sd of 4k)
- A random sample of 200 women has anincome of \$32,000 (sd of 3k)
- Is this mean difference real or an artifact of random error?

R's t.test function can handle it

- Generate our samples

Suppose our means are identical. This really supports the null hypothesis


```
rnorm2 <- function(n,mean,sd) {mean+sd*scale(rnorm(n)

men <- rnorm2(300, 32000, 4000)
women <- rnorm2(200, 32000, 3000)
(d_mean <- t.test(men, women))
```

```
##
##  Welch Two Sample t-test
##
## data:  men and women
## t = 0, df = 491.11, p-value = 1
## alternative hypothesis: true difference in means i
## 95 percent confidence interval:
##  -616.1268  616.1268
## sample estimates:
## mean of x mean of y
```

Now let's say mean for men is 32,800 and 32,000 for women


```
rnorm2 <- function(n,mean,sd) {mean+sd*scale(rnorm(n)

men <- rnorm2(300, 32800, 4000)
women <- rnorm2(200, 32000, 3000)
(d_mean <- t.test(men, women))
```

```
##
##  Welch Two Sample t-test
##
## data:  men and women
## t = 2.5512, df = 491.11, p-value = 0.01104
## alternative hypothesis: true difference in means i
## 95 percent confidence interval:
##   183.8732 1416.1268
## sample estimates:
## mean of x mean of y
```

Let's bump the confidence level from .95 to .99


```
men <- rnorm2(300, 32800, 4000)
women <- rnorm2(200, 32000, 3000)
(d_mean <- t.test(men, women, conf.level = 0.99))
```

```
##
##  Welch Two Sample t-test
##
## data:  men and women
## t = 2.5512, df = 491.11, p-value = 0.01104
## alternative hypothesis: true difference in means i
## 99 percent confidence interval:
##   -10.88315 1610.88315
## sample estimates:
## mean of x mean of y
##     32800     32000
```

How to interpret?

- We see that the null hypothesis (difference = 0) is extremely unlikely.
- We would expect the differences to range between -10.88 and 1610.88 95% of the time.
- So (we argue) the difference is not just random error. Something else is producing the difference.

Do Catholics like Pope Francis better than non-Catholics?

- After some data wrangling from the Pew Survey, we see the following (1: Very favourable . . . 4: Very unfavourable)
  NB: Excuse the presentation

```
## # A tibble: 4 x 3
##                          q84 Catholic `Not Catholic`
##                    <dbl+lbl>    <int>          <int>
## 1 1 [Very favorable]               96            137
## 2 2 [Mostly favorable]            141            549
## 3 3 [Mostly unfavorable]           49            227
## 4 4 [Very unfavorable]             20            153
```

```
## # A tibble: 2 x 2
##   CATHOLIC      Mean
##   <chr>        <dbl>
## 1 Catholic      1.98
## 2 Not Catholic  2.37
```

```
t.test(q84 ~ CATHOLIC, sep_18)
```

```
##
##  Welch Two Sample t-test
##
## data:  q84 by CATHOLIC
## t = -7, df = 500, p-value = 6e-12
## alternative hypothesis: true difference in means i
## 95 percent confidence interval:
##  -0.5044 -0.2843
## sample estimates:
##     mean in group Catholic mean in group Not Catho
##                      1.977                       2.
```

This crude boxplot suggests a similar story

POLS201
Spring 2019