Intro to Social Science Data Analysis

Lecture 9: Statistical Inference with Large Samples

Christopher Gandrud

October 15, 2012

Assignment 3

Intro to Statistical Inference: Quick Quiz (1)

Give an example of a population parameter and its corresponding point estimate.

Intro to Statistical Inference: Quick Quiz (2)

What is the sampling distribution of the sampling mean?

In general, what is the sampling distribution of the sampling mean centered on?

Intro to Statistical Inference: Quick Quiz (3)

What do we use to find the standard error of a point estimate?

What do we use the standard error for?

Intro to Statistical Inference: Quick Quiz (4)

What is a confidence interval?

Why is it more useful to show the confidence interval than just the standard error?

Last class we largely looked at how to draw inferences about a population **mean** from a sample **mean**.

Today we will expand our inferential tools by learning about:

- Hypothesis testing & p-values,
- Comparing 2 population means,
- Making inferences with population proportions,
- Inferential statistics with categorical variables.

Last class we largely looked at how to draw inferences about a population **mean** from a sample **mean**.

Today we will expand our inferential tools by learning about:

- Hypothesis testing & p-values,
- Comparing 2 population means,
- Making inferences with population proportions,
- Inferential statistics with categorical variables.

Today

Last class we largely looked at how to draw inferences about a population **mean** from a sample **mean**.

Today we will expand our inferential tools by learning about:

- Hypothesis testing & p-values,
- Comparing 2 population means,
- Making inferences with population proportions,
- Inferential statistics with categorical variables.

Last class we largely looked at how to draw inferences about a population **mean** from a sample **mean**.

Today we will expand our inferential tools by learning about:

- Hypothesis testing & p-values,
- Comparing 2 population means,
- Making inferences with population proportions,
- Inferential statistics with categorical variables.

Hypothesis Testing Setup

Imagine that we have a sample of 200 Cherrry Blossom Run Finishing times from the 2009 race.

(This example is largely from Diaz et al. 2011. See last week's lecture for more details)

The mean finishing time in the sample is 95.5 minutes with a standard deviation of 16.1

Question

The mean finishing time in 2006 was 93.29.

Is there strong evidence that on average the 2009 runners are faster/slower than the 2006 runners?

The language of hypothesis testing.

We can think that there are two **competing** possibilities:

- H_0: There is *no difference* in the average finishing times between the 2006 and 2009 runners (**the null hypothesis**).
- H_a: The average finishing time in 2006 *is different* from the average finishing time in 2009 (**the alternative hypothesis**).

The language of hypothesis testing.

We can think that there are two **competing** possibilities:

- H_0: There is *no difference* in the average finishing times between the 2006 and 2009 runners (**the null hypothesis**).
- H_a: The average finishing time in 2006 *is different* from the average finishing time in 2009 (**the alternative hypothesis**).

In other words, if the population mean for the 2009 is called μ_{09}:

- $H_0 : \mu_{09} = 93.29$
- $H_A : \mu_{09} \neq 93.29$

93.29 is called the **null value**, as it is the value of the parameter **if** the null hypothesis is true.

In other words, if the population mean for the 2009 is called μ_{09}:

- $H_0 : \mu_{09} = 93.29$
- $H_A : \mu_{09} \neq 93.29$

93.29 is called the **null value**, as it is the value of the parameter **if** the null hypothesis is true.

The language of hypothesis testing

The null hypothesis is the **skeptical possibility**.

If we do not find evidence against the null hypothesis we say that we: *fail to reject the null hypothesis.*

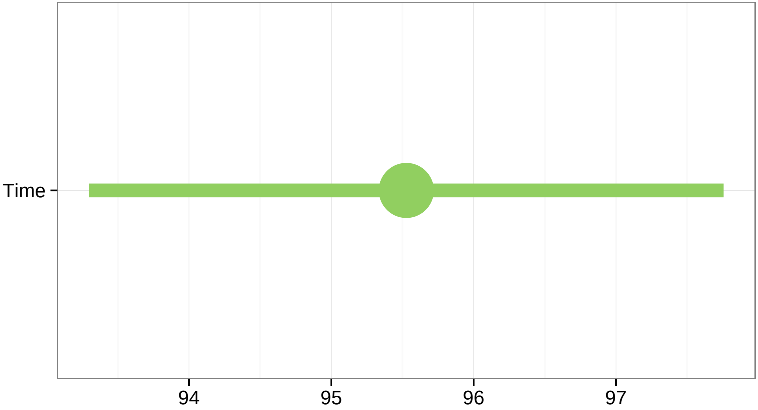
If we do find evidence against the null hypothesis we say that we: *found evidence for the alternative hypothesis.*

Evidence

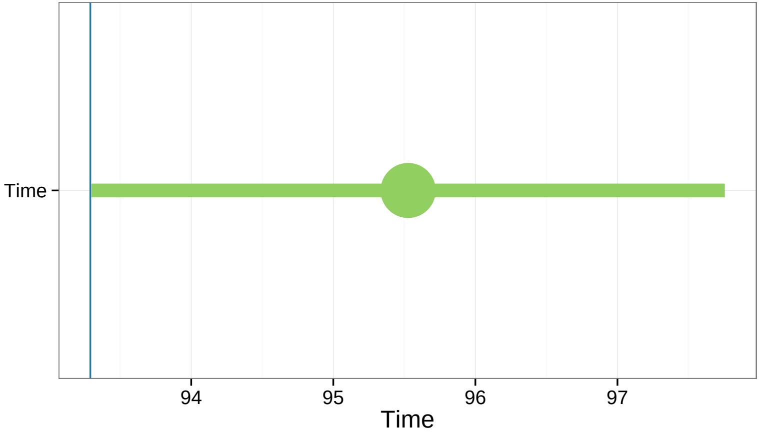
What kind of evidence can we use to either reject or fail to reject the null hypothesis?

Confidence intervals!

95% Confidence Inverval for 2009 Mean Finishing Times ($n = 200$)



95% Confidence Inverval for 2009 Mean Finishing Times ($n = 200$) Compared to the Null Hypothesis



Fail to Reject

The 2006 population mean is inside of the 95% confidence interval of the 2009 sample estimate.

Therefore, we *fail to reject* the null hypothesis that the 2009 Cherry Blossom Run mean finishing time is different from the 2006 mean finishing time.

Be Careful: Hypothesis testing is far from perfect.

		Test Conclusion	
		Do Not Reject H_0	Reject Infavour of H_A
Real World	H_0 True	okay	Type 1 Error
	H_A True	Type II Error	okay

(Diaz et al. 2001, 160)

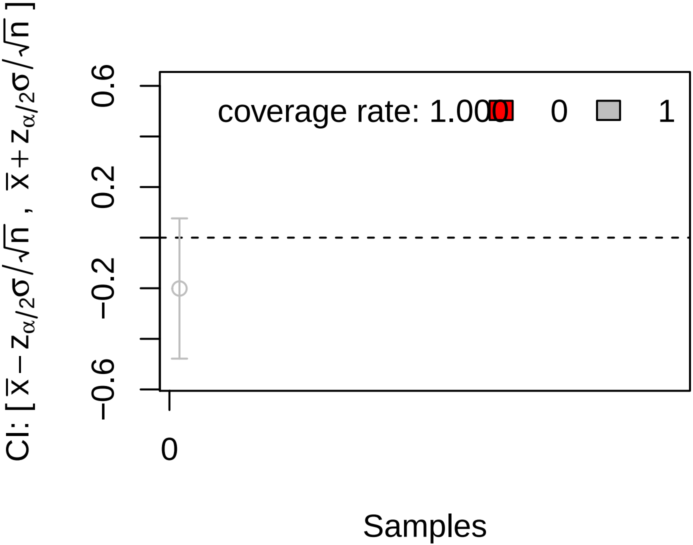
We previously used a 95% confidence interval to test the Null Hypothesis.

This means that 5% of the time we will **incorrectly** reject H_0 due to **sampling variation**.

2.5% of the time the confidence interval will be **too high**.
2.5% of the time the confidence interval will be **too low**.

This is called the 95% **significance level** or sometimes $\alpha = 0.05$.

Remember: Confidence Interval Simulation



Higher Confidence

If we use a **higher significance level** we will be more confident that we correctly rejected or failed to reject the null hypothesis.

For example, at the 99% significance level we will incorrectly reject the H_0 1% of the time.

Some researchers like to quantify the strength of the evidence against the Null Hypothesis with a tool called the **p-value**.

p-value

The probability of seeing data at least as favourable to the alternative hypothesis as our current data, *if the null hypothesis is true*.

Using p-values to comparing means:

Are men's finishing times different than women's finishing time in the 2009 Cherry Blossom Run?

First the Descriptive Statistics

```
#### Summary of Men's Times #### Subset sample to
#### include only men's Times
MenSubset <- subset(Run10Samp$time, Run10Samp$gender ==
    "M")
# Mean
Mean <- mean(MenSubset)
# Standard Deviation
SD <- sd(MenSubset)
# Number of observations
N <- length(MenSubset)
# Create Gender Variable
Gender <- "Male"
# Combine
GenderData <- data.frame(Gender, Mean, SD, N)
```

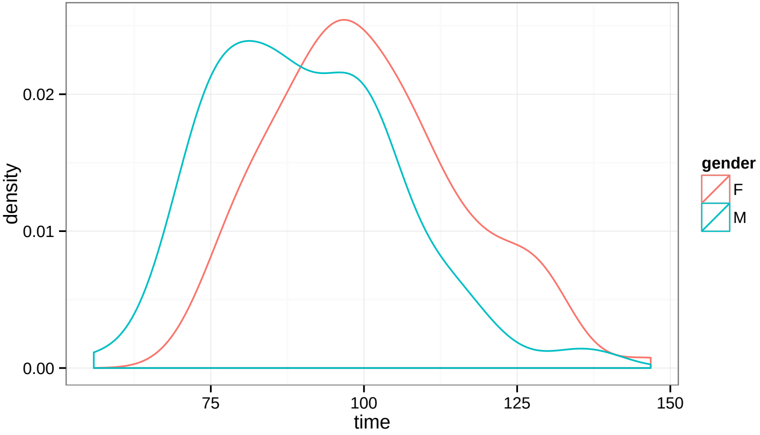
```r
#### Summary of Women's Times #### Subset sample
#### to include only women's Times
WomenSubset <- subset(Run10Samp$time, Run10Samp$gender ==
    "F")
# Mean
Mean <- mean(WomenSubset)
# Standard Deviation
SD <- sd(WomenSubset)
# Number of observations
N <- length(WomenSubset)
# Create Gender Variable
Gender <- "Female"
# Combine
GenderDataF <- data.frame(Gender, Mean, SD, N)

# Combine into one data frame
GenderData <- data.frame(rbind(GenderData, GenderDataF))
```

Summary Descriptives

Table: Descriptive Statistics of the Sample

	$\bar{x}$	s	n
Female	100.7	15.3	101
Male	90.2	15.1	99

```r
# Compare densities of Men/Women Times
ggplot(Run10Samp, aes(time)) +
        geom_density(aes(
          line = gender, color = gender)) +
        theme_bw()
```



Null Hypothesis: $\mu_{men} = \mu_{women}$

Alternative Hypothesis: $\mu_{men} \neq \mu_{women}$

An equivalent way to write this null hypothesis is:
$\mu_{men} - \mu_{women} = 0$

References I

Diaz, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2011. OpenIntro Statistics. 1st ed. http://www.openintro.org/stat/downloads.php.