# Intro to Social Science Data Analysis

# Lecture 9: Overview of Statistical Inference (II)

**Christopher Gandrud**

October 22, 2012

## Assignment 2 General Feedback

- **Always** describe your data source.
- What do the numbers mean? Remember the units.
- What year(s) is the data from?

## Assignment 2 General Feedback

- **Always** describe your data source.
- What do the numbers mean? Remember the units.
- What year(s) is the data from?

## Assignment 2 General Feedback

- **Always** describe your data source.
- What do the numbers mean? Remember the units.
- What year(s) is the data from?

# Assignment 3

**Due:** Friday 16 November

Have a data set with three variables of the following type:

- 1 numeric variable,
- 1 dummy variable,
- 1 multinomial variable.

There should be more than 50 observations per variable & variable category.

# Assignment 3

**Due:** Friday 16 November

Have a data set with three variables of the following type:

- 1 numeric variable,
- 1 dummy variable,
- 1 multinomial variable.

There should be more than 50 observations per variable & variable category.

# Assignment 3

**Due:** Friday 16 November

Have a data set with three variables of the following type:

- 1 numeric variable,
- 1 dummy variable,
- 1 multinomial variable.

There should be more than 50 observations per variable & variable category.

Find the answers to these questions.

Numeric Continuous Variable

- What do you predict the population mean of this variable is?
- Create two groups of this variable based on the dummy variable. Are the population means of these two groups likely to be different?

Dummy Variable

- What are the likely population proportions of the two values of this variable?

Multinomial Variable

- What are the likely population proportions of the values of this variable?

Find the answers to these questions.

Numeric Continuous Variable

- What do you predict the population mean of this variable is?
- Create two groups of this variable based on the dummy variable. Are the population means of these two groups likely to be different?

Dummy Variable

- What are the likely population proportions of the two values of this variable?

Multinomial Variable

- What are the likely population proportions of the values of this variable?

Find the answers to these questions.

Numeric Continuous Variable

- What do you predict the population mean of this variable is?
- Create two groups of this variable based on the dummy variable. Are the population means of these two groups likely to be different?

Dummy Variable

- What are the likely population proportions of the two values of this variable?

Multinomial Variable

- What are the likely population proportions of the values of this variable?

Find the answers to these questions.

Numeric Continuous Variable

- What do you predict the population mean of this variable is?
- Create two groups of this variable based on the dummy variable. Are the population means of these two groups likely to be different?

Dummy Variable

- What are the likely population proportions of the two values of this variable?

Multinomial Variable

- What are the likely population proportions of the values of this variable?

# 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?

# 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 including $\mathrm{Z}$ scores, $t$ distributions, & p-values,
- Comparing 2 population means,

Next class:

- 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 including $\mathrm{Z}$ scores, $t$ distributions, & p-values,
- Comparing 2 population means,

Next class:

- 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 including $\mathrm{Z}$ scores, $t$ distributions, & p-values,
- Comparing 2 population means,

Next class:

- 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 including $\mathrm{Z}$ scores, $t$ distributions, & p-values,
- Comparing 2 population means,

Next class:

- Making inferences with population proportions,
- Inferential statistics with categorical variables.

# Hypothesis Testing Setup

Imagine that we have a sample of 200 Cherry 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 93.9 minutes with a standard deviation of 15.6

# 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**).

# The language of hypothesis testing

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

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

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

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

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

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.*

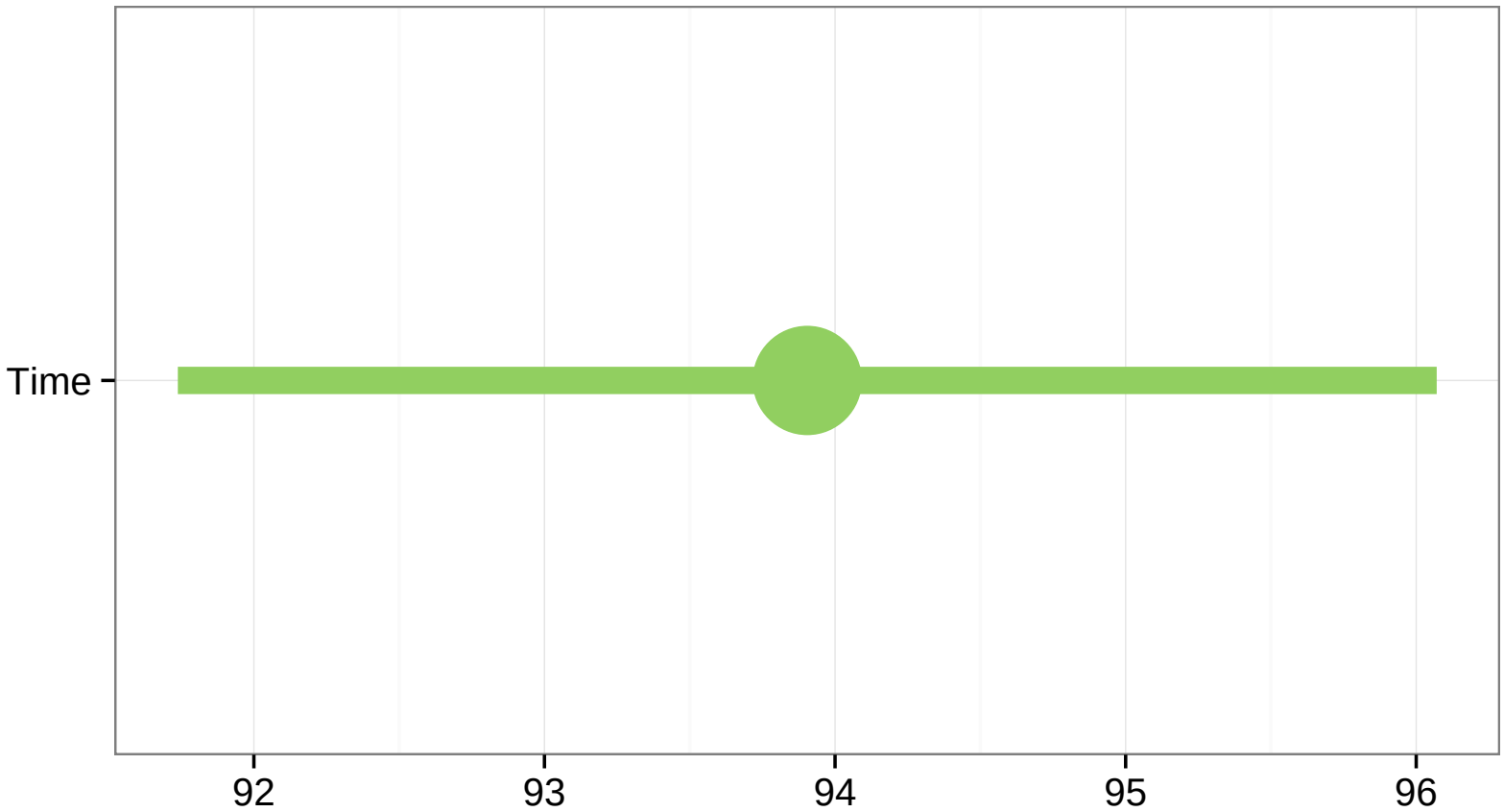
# Evidence

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

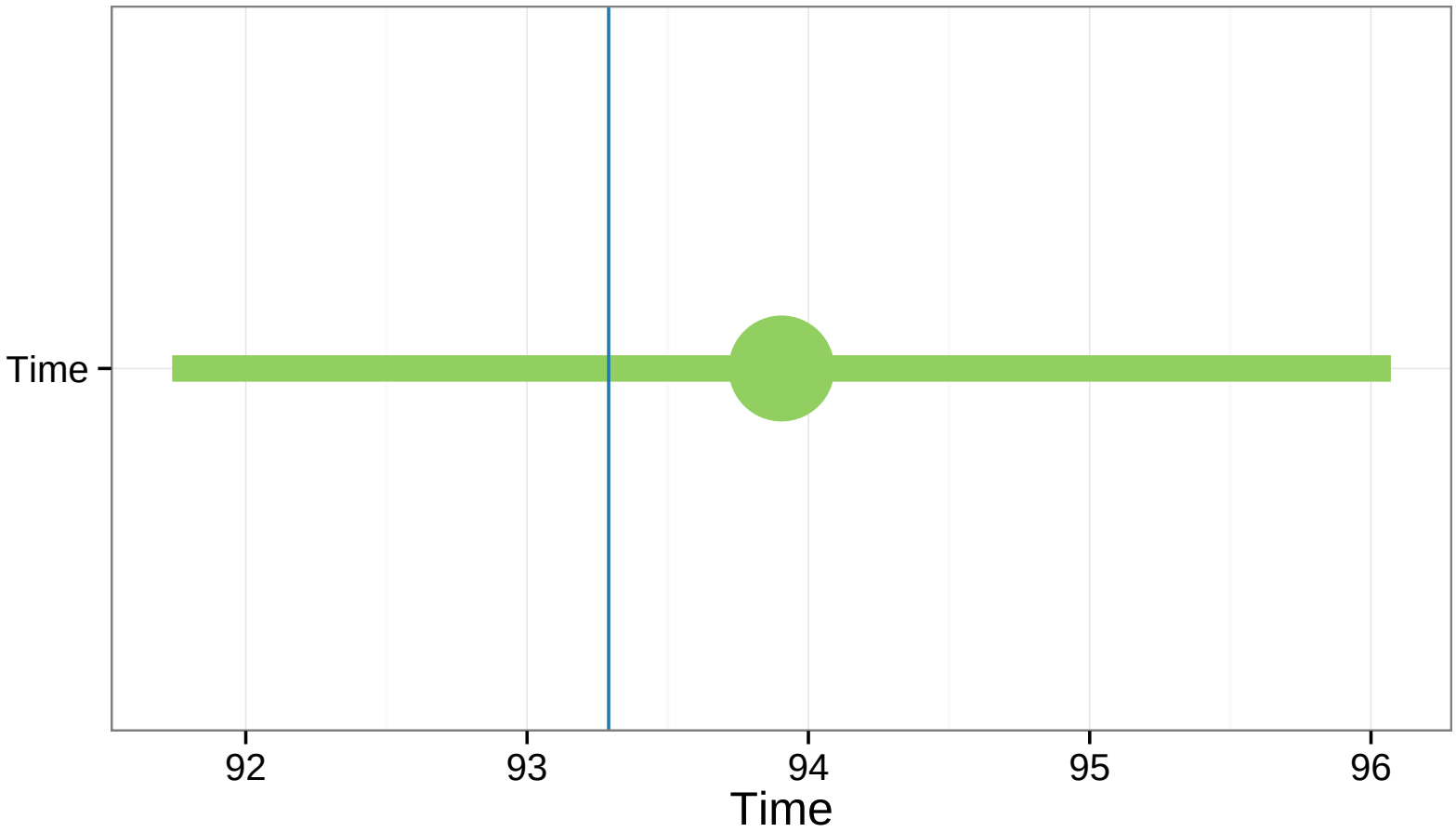
# Evidence

Confidence intervals!

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



# 95% Confidence Interval 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 in favour of $H_A$ |
| **Real World** | $H_0$ True | okay | Type 1 Error |
| | $H_A$ True | Type II Error | okay |

(Diaz et al. 2001, 160)

# Quantifying Error Probabilities

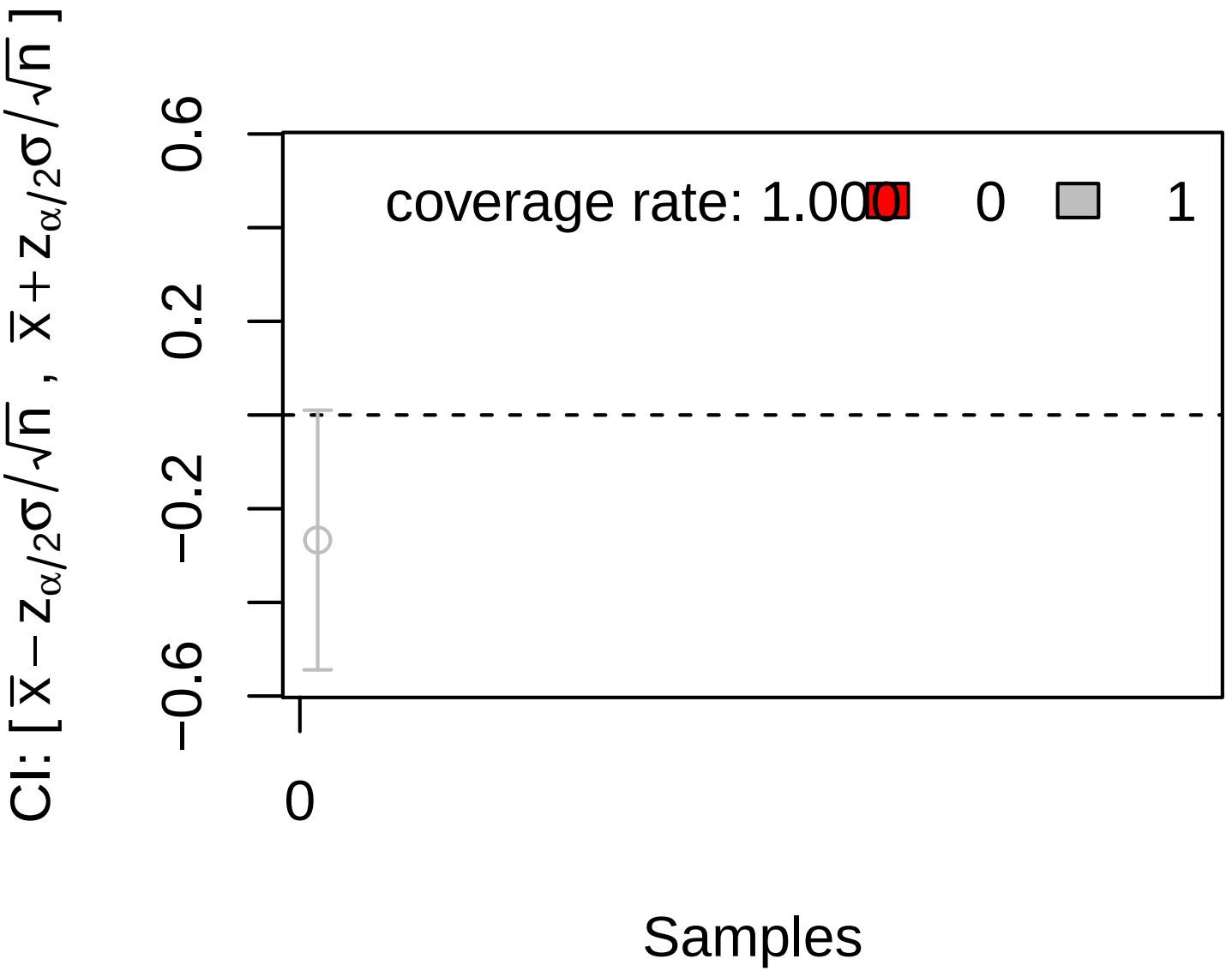
We previously used a 95% confidence interval to test the Null Hypothesis.

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

# Remember: Confidence Interval Simulation



# Higher Confidence

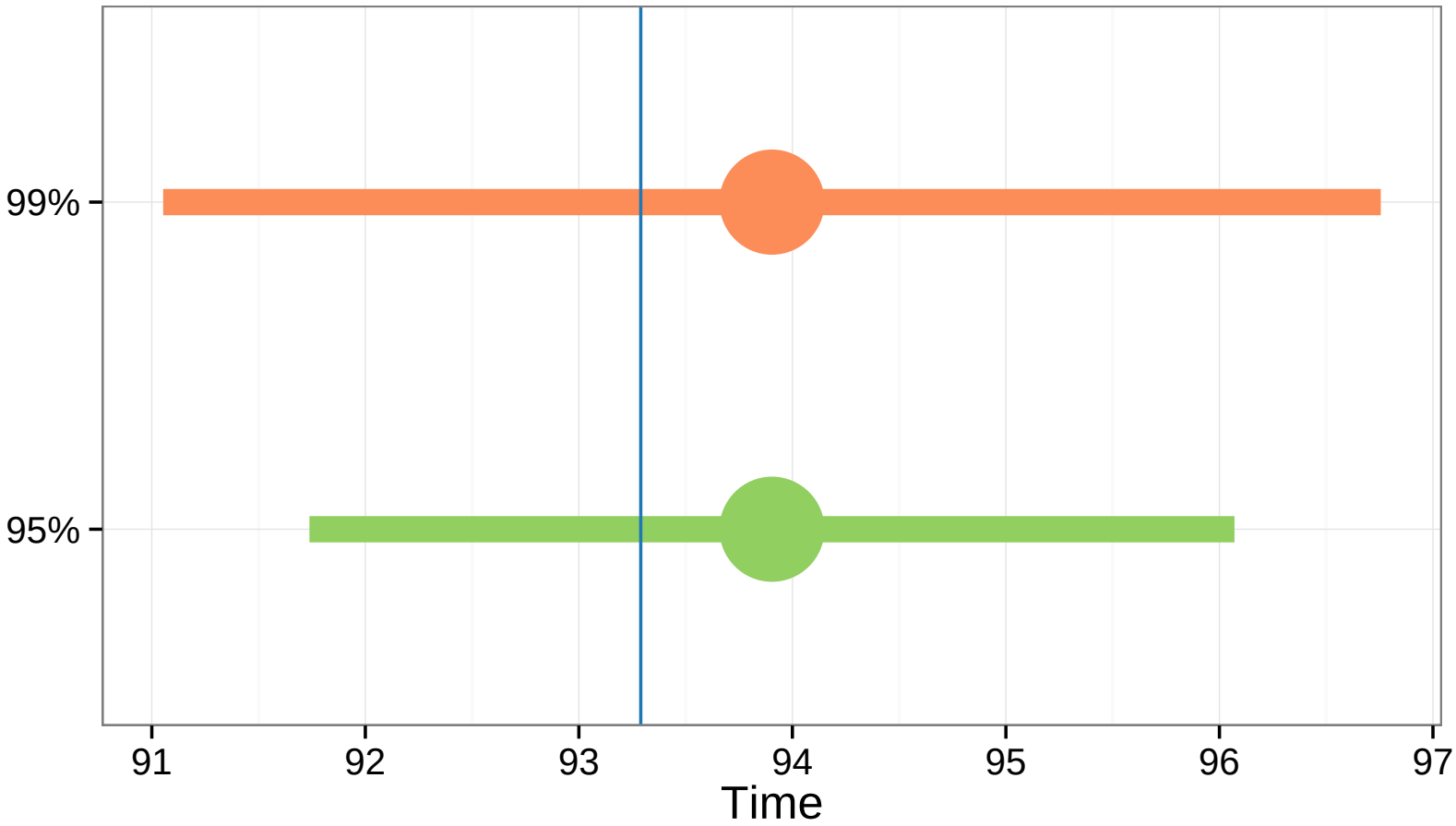
If we use a **higher significance level** we will be more confident that we correctly rejected or failed to reject the null hypothesis.

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

The 99% Confidence Interval of the Mean:

$$CI_{99\%} = \bar{x} \pm 2.58 * SE_{\bar{x}}$$

# Comparing Confidence Intervals



# Confidence Intervals and Z Scores

Where does the 2.58 come from?

Where does the 1.96 come from (for 95% confidence intervals)?

# Confidence Intervals and Z Scores

They are the $\mathrm{Z}$ score for the confidence level.

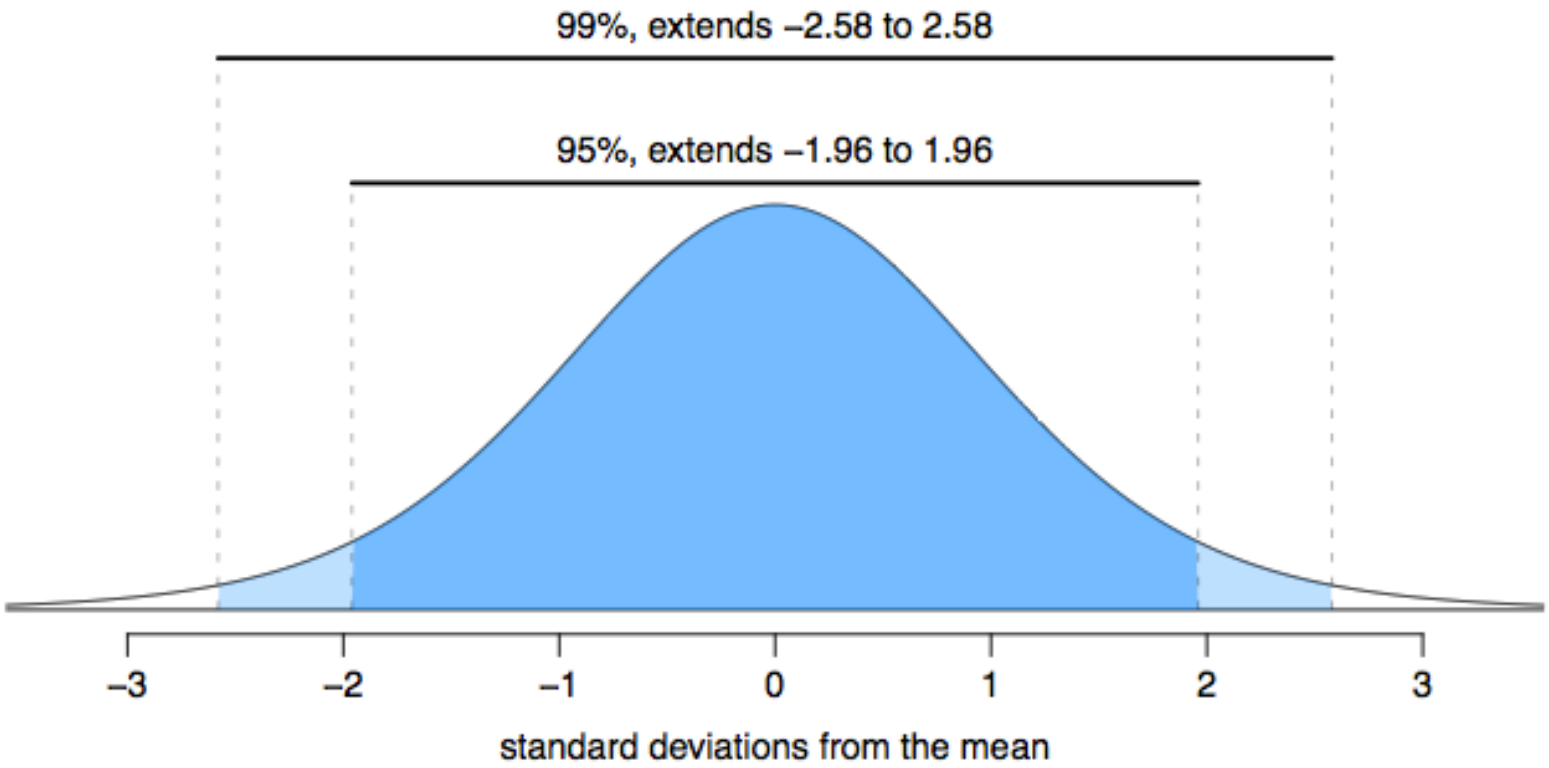
$\mathrm{Z}$ scores are the number of standard deviations above or below the mean that an observation falls.

Because of the Central Limit Theorem, if

- the sample observations are **independent**,
- the sample size is **large** (about $n \geq 50$),
- the distribution is **not extremely skewed**.

then we can assume that...

# Z Scores and Confidence Intervals




Source: Diaz et al. (2011, 154)

# What if the Central Limit Theorem's Assumptions Aren't Met?

- Dependent data: See Sainani (2010): `http://www.stanford.edu/~kcobb/hrp259/correlateddata.pdf`.
- The larger the sample, the more lenient you can be about skewness.
- If you have a smaller sample size and it is skewed, you can use a version of the Wilcox test (in R `Wilcox.test`) for hypothesis testing. See Crawley (2005, Ch. 5)
- If you have a small sample (about $n < 50$) that is not highly skewed, you can use the $t$ distribution to calculate confidence intervals. In R use `t.test`. See Diaz et al. (2011, Ch. 6).

# What if the Central Limit Theorem's Assumptions Aren't Met?

- Dependent data: See Sainani (2010): `http://www.stanford.edu/~kcobb/hrp259/correlateddata.pdf`.
- The larger the sample, the more lenient you can be about skewness.
- If you have a smaller sample size and it is skewed, you can use a version of the Wilcox test (in R `Wilcox.test`) for hypothesis testing. See Crawley (2005, Ch. 5)
- If you have a small sample (about $n < 50$) that is not highly skewed, you can use the $t$ distribution to calculate confidence intervals. In R use `t.test`. See Diaz et al. (2011, Ch. 6).

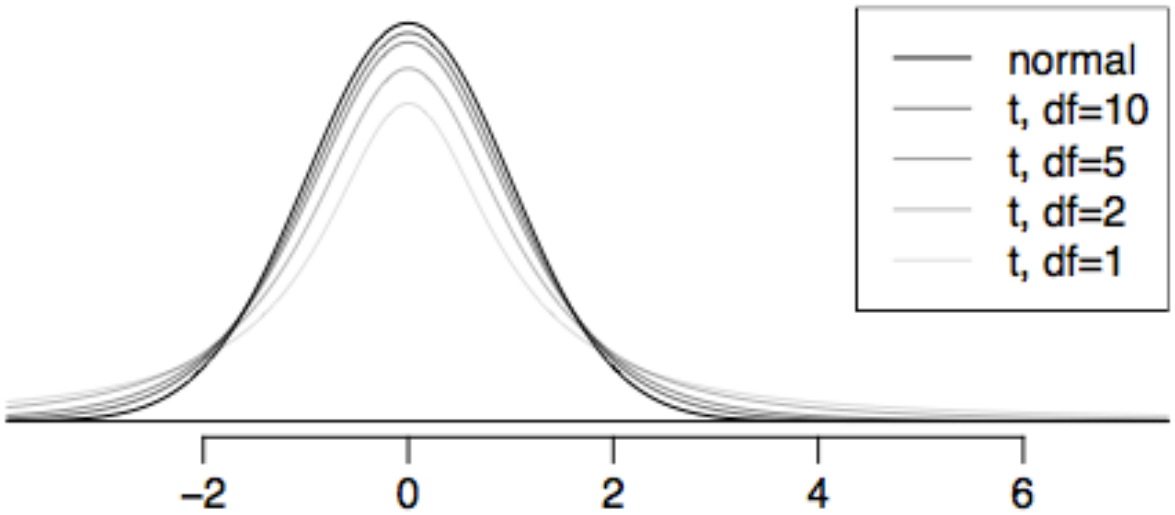
# What if the Central Limit Theorem's Assumptions Aren't Met?

- Dependent data: See Sainani (2010): `http://www.stanford.edu/~kcobb/hrp259/correlateddata.pdf`.
- The larger the sample, the more lenient you can be about skewness.
- If you have a smaller sample size and it is skewed, you can use a version of the Wilcox test (in R `Wilcox.test`) for hypothesis testing. See Crawley (2005, Ch. 5)
- If you have a small sample (about $n < 50$) that is not highly skewed, you can use the $t$ distribution to calculate confidence intervals. In R use `t.test`. See Diaz et al. (2011, Ch. 6).

# What if the Central Limit Theorem's Assumptions Aren't Met?

- Dependent data: See Sainani (2010): `http://www.stanford.edu/~kcobb/hrp259/correlateddata.pdf`.
- The larger the sample, the more lenient you can be about skewness.
- If you have a smaller sample size and it is skewed, you can use a version of the Wilcox test (in R `Wilcox.test`) for hypothesis testing. See Crawley (2005, Ch. 5)
- If you have a small sample (about $n < 50$) that is not highly skewed, you can use the $t$ distribution to calculate confidence intervals. In R use `t.test`. See Diaz et al. (2011, Ch. 6).

# The $t$ distribution with various Degrees of Freedom




df = Degrees of Freedom $(n-1)$

Source: Diaz et al. (2011, 244)

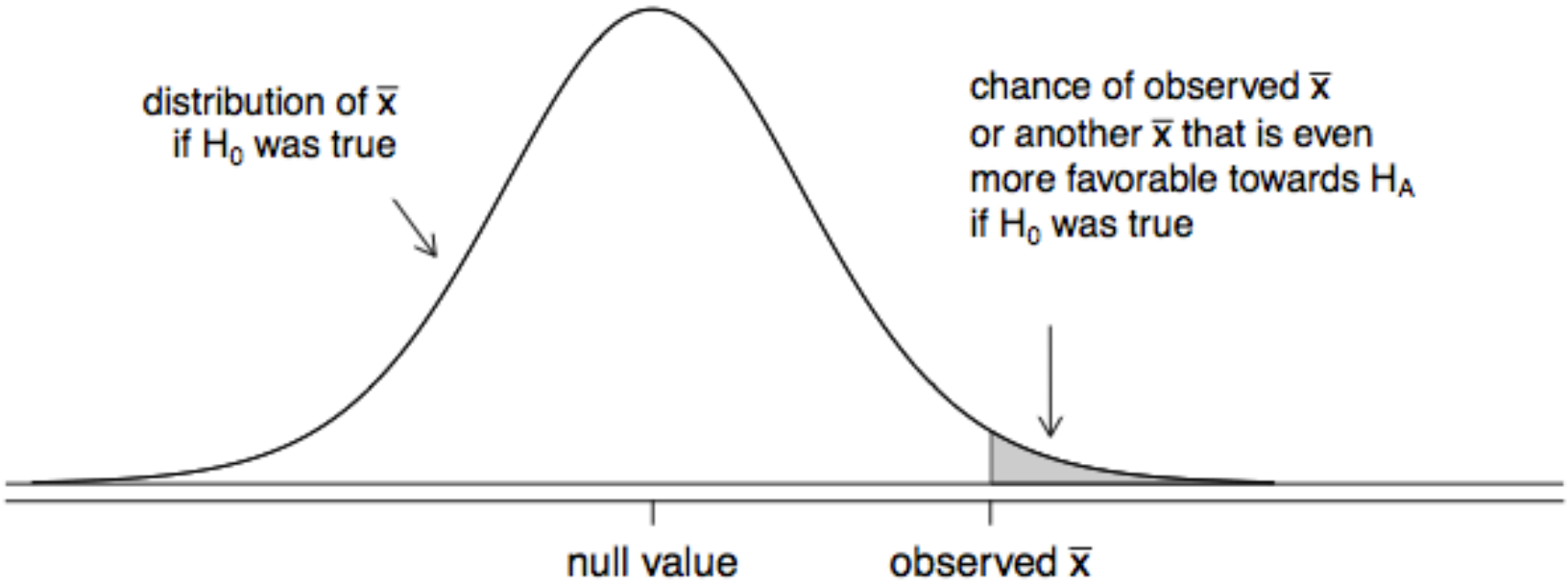
Some researchers like to quantify the strength of the evidence against the Null Hypothesis with a tool called the **p-value**.

## 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*.

In our previous **example**, we would want to know the probability of seeing a mean of at least 94.5 if the true population mean is actually 93.29 (the population mean for 2006).

# Visualizing the p-value



# Example

In our previous **example**, we would want to know the probability of seeing a mean of at least 94.5 if the true population mean is actually 93.29 (the population mean for 2006).

# 1st: Test Statistic

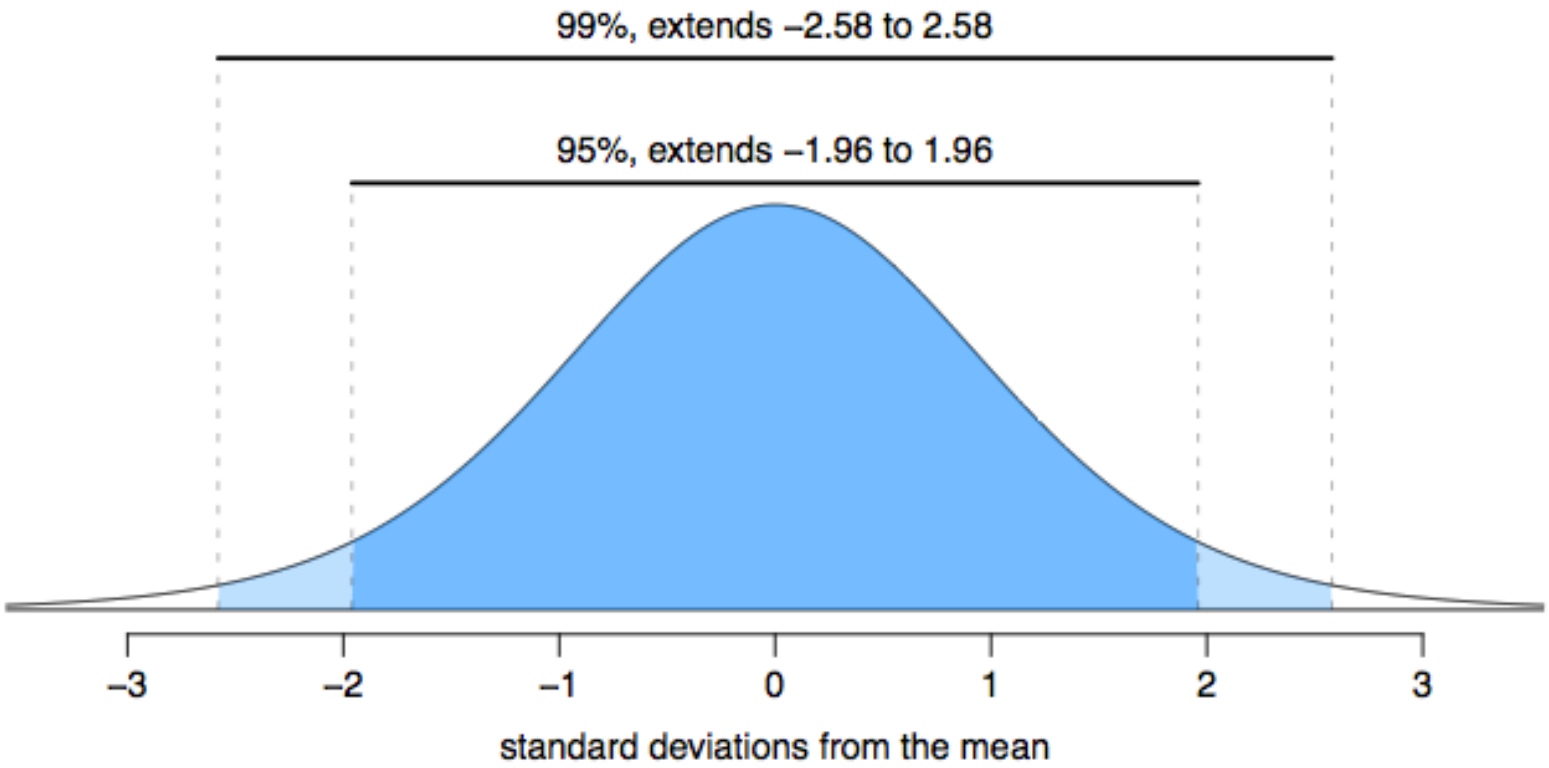
We first calculated the Z score **test statistic** of the sample mean $\bar{x} = 94.5$

$$\text{Z} = \frac{\bar{x} - \text{null value}}{SE_{\bar{x}}}$$

In R:

```
# Load library
library(plotrix)

# Find Z score
Z <- (mean(Run10Samp$time) - 93.29)/
      std.error(Run10Samp$time)

Z

## [1] 0.5559
```

# $\mathbb{Z}$ Scores and Confidence Intervals




Source: Diaz et al. (2011, 154)

# 2nd: Find p-values

To find the p-value for the 0.6 use `pnorm`.

```
P1 <- 1 - pnorm(Z)
P1

## [1] 0.2891
```

So, if the null hypothesis was true we would have a 0.2891 probability of observing a mean of at least 0.6 larger than the 2006 mean.

This is bigger than the 0.05 significance level. So, we fail to reject the null hypothesis that $\mu_{x_{09}} > \mu_{x_{06}}$.

# Um. . .

But we were interested in the null hypothesis $\mu_{x_{09}} = \mu_{x_{06}}$, not $\mu_{x_{09}} > \mu_{x_{06}}$.

So far we have only looked one tail of the distribution. To examine both tails simply **multiply the one-tail p-value by 2**.

```
P2 <- 2 * (1 - pnorm(Z))

P2

## [1] 0.5782
```

We fail to reject the null hypothesis that $\mu_{x_{09}} = \mu_{x_{06}}$.

In other words, we have evidence that there is no difference in the mean finishing times from the 2006 and 2009 Cherry Blossom Run.

# Comparing Means

# Question

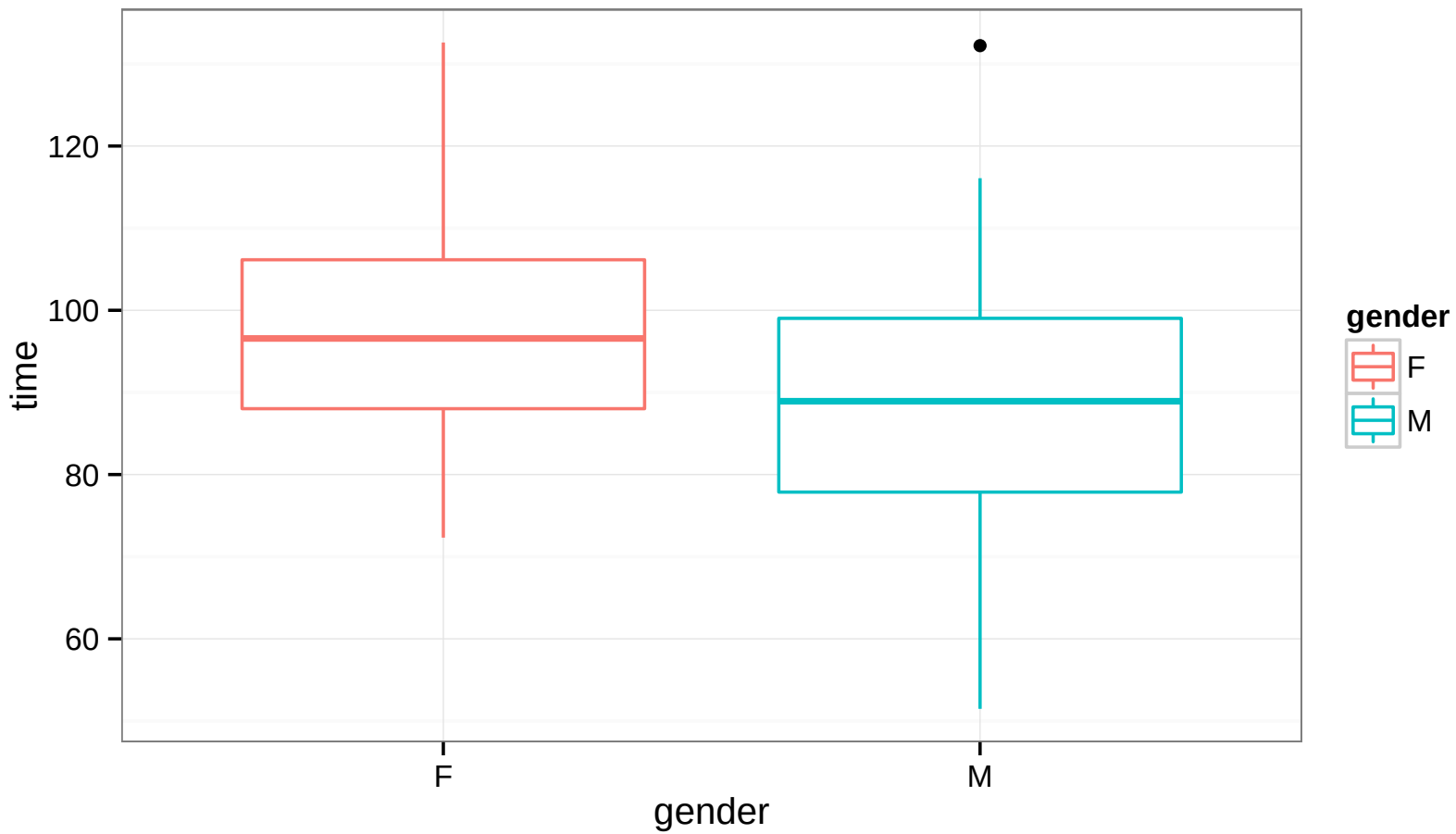
Are men's finishing times different than women's finishing time in the 2009 Cherry Blossom Run?

# Summary Descriptives

Table: Descriptive Statistics of the Sample

| | $\bar{x}$ | $s$ | $n$ |
|---|---|---|---|
| Female | 98.9 | 14.2 | 105 |
| Male | 88.3 | 15.3 | 95 |

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



Null Hypothesis:

$$\mu_m = \mu_w$$

Alternative Hypothesis:

$$\mu_m \neq \mu_w$$

An equivalent way to write this null hypothesis is:

$$\mu_m - \mu_w = 0$$

# Our Sample

The **difference in mean** times between Men and Women is:

$$\mu_m - \mu_w$$

88.3 - 98.9 = -10.6

# Standard Error of the Difference of Two Means

Standard Error of the Difference of Two Means:

$$SE_{\bar{x}_m - \bar{x}_w} = \frac{\sigma_m}{\sqrt{n_m}} + \frac{\sigma_w}{\sqrt{n_w}}$$

# Approximate the Standard Error.

Because in our sample each gender has **more than 50 observations** that are **independent**, and the is **normally distributed** we can use the **sample standard deviations** to approximate the standard error of the difference of two means:

$$SE_{\bar{x}_m - \bar{x}_w} = \frac{\sigma_m}{\sqrt{n_m}} + \frac{\sigma_w}{\sqrt{n_w}} \approx \frac{s_m}{\sqrt{n_m}} + \frac{s_w}{\sqrt{n_w}} = \frac{15.3}{\sqrt{95}} + \frac{14.2}{\sqrt{105}}$$

$$SE_{\bar{x}_m - \bar{x}_w} = 2.9565$$

## Here is the R Code for Everything We Need (1)

```
# Subset Samples
MenSubset <- subset(Run10Samp$time,
                    Run10Samp$gender == "M")
WomenSubset <- subset(Run10Samp$time,
                    Run10Samp$gender == "F")

# Means
MeanMen <- mean(MenSubset)
MeanWomen <- mean(WomenSubset)

# Mean difference
MeanDiff <- MeanMen - MeanWomen
```

## Here is the R Code for Everything We Need (2)

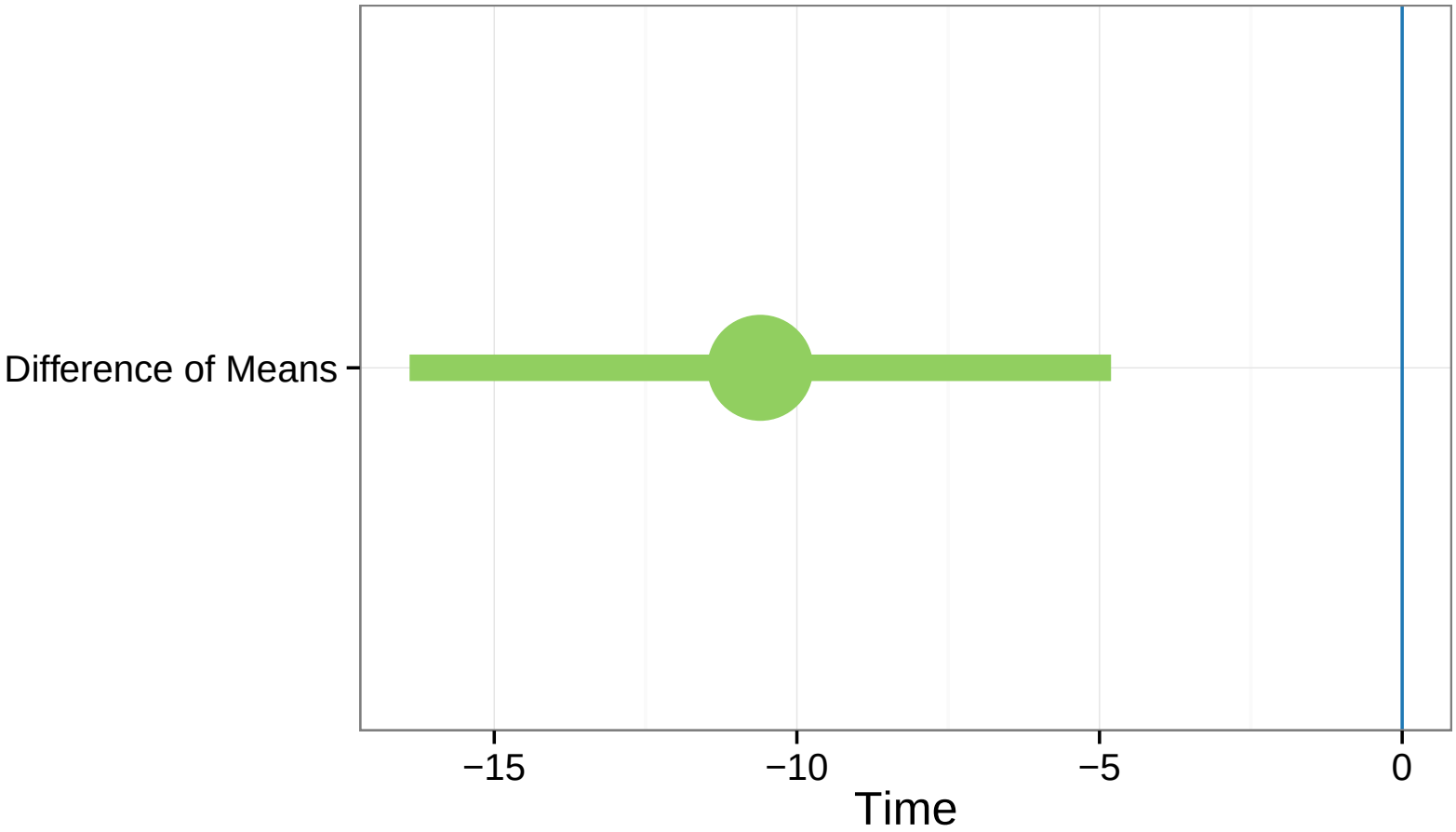
```
# Standard Errors
SEMen <- std.error(MenSubset)
SEWomen <- std.error(WomenSubset)

# Standard error of the difference of two means
SEDiff <- SEMen + SEWomen
```

Now that we have the difference in means, the standard error of the difference in means, and the sample sizes we can statistically examine whether or not the mean finishing times are different for men and women.

## Let's find the 95% Confidence interval of the Difference of the Two Means

```
# Lower bound of the confidence interval
Lower <- MeanDiff - 1.96 * SEDiff

# Upper bound of the confidence interval
Upper <- MeanDiff + 1.96 * SEDiff

Lower

## [1] -16.4

Upper

## [1] -4.81
```

# Better



# Substantive Interpretation

So, we can reject the null hypothesis that $\mu_m - \mu_w = 0$

I.e. we have evidence that women's and men's finishing times in the 2009 Cherry Blossom Race were different.

# References I

Crawley, Michael J. 2005. Statistics: An Introduction Using R. Chichester: John Wiley Sons. Ltd.

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