Hypothesis Tests for Numerical Data

Ryan Miller

Introduction

- The previous lecture introduced the t-distribution
  - You can view the t-distribution as a modified version of the Standard Normal curve with thicker tails
  - These tails account for the extra uncertainty involved in estimating σ, the standard deviation of the population, using s, the sample standard deviation

Introduction

- The previous lecture introduced the t-distribution
  - You can view the t-distribution as a modified version of the Standard Normal curve with thicker tails
  - These tails account for the extra uncertainty involved in estimating σ, the standard deviation of the population, using s, the sample standard deviation
- This lecture will cover two important topics
  - Using the t-distribution to conduct hypothesis tests
  - Alternative tests when the assumptions of the t-distribution are not met

Wetsuits and the Olympics

- At the 2008 Beijing Olympics, 25 different swimming world records were broken
  - This was the most since 1976, when goggles were first used in competition

Wetsuits and the Olympics

- At the 2008 Beijing Olympics, 25 different swimming world records were broken
  - This was the most since 1976, when goggles were first used in competition
- Of these 25 new records, 23 were set by swimmers using a wetsuit known as the *LZR Racer*, a suit produced by Speedo whose design involved scientists at NASA
  - The led to an existential crisis in competitive swimming, culminating in a ban on certain types of suits in competition
- But how convincing is the evidence that the LZR Racer truly provides an unfair advantage?
  - What alternative explanations might exist for 23 of 25 records being set by swimmers who wore LZR Racers?

Wetsuits and the Olympics

- At the 2008 Beijing Olympics, 25 different swimming world records were broken
  - This was the most since 1976, when goggles were first used in competition
- Of these 25 new records, 23 were set by swimmers using a wetsuit known as the *LZR Racer*, a suit produced by Speedo whose design involved scientists at NASA
  - The led to an existential crisis in competitive swimming, culminating in a ban on certain types of suits in competition
- But how convincing is the evidence that the LZR Racer truly provides an unfair advantage?
  - What alternative explanations might exist for 23 of 25 records being set by swimmers who wore LZR Racers?
  - Since these data are *observational*, it could be that all of the best swimmers were wearing this suit, so an *experimental* study is warranted

Wetsuits

- The `wetsuits` data contains the results of an experiment involving 12 competitive swimmers
  - Each swam 1500m for time under two conditions: wearing a high-tech wetsuit, or wearing a placebo suit identical in appearance
  - It was randomly determined which condition the participant experienced first
- The columns `Wetsuit` and `NoWetsuit` record the respective velocities (in m/s) over the 1500m swim

```
wet <- read.csv("https://remiller1450.github.io/data/Wetsuits.csv")
summary(wet[,1:2])
```

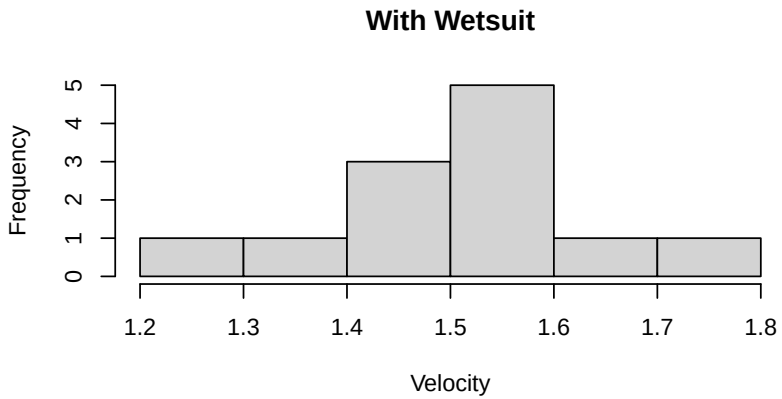
```
##     Wetsuit         NoWetsuit
##  Min.   :1.220   Min.   :1.120
##  1st Qu.:1.458   1st Qu.:1.365
##  Median :1.520   Median :1.445
##  Mean   :1.507   Mean   :1.429
##  3rd Qu.:1.570   3rd Qu.:1.505
##  Max.   :1.750   Max.   :1.640
```

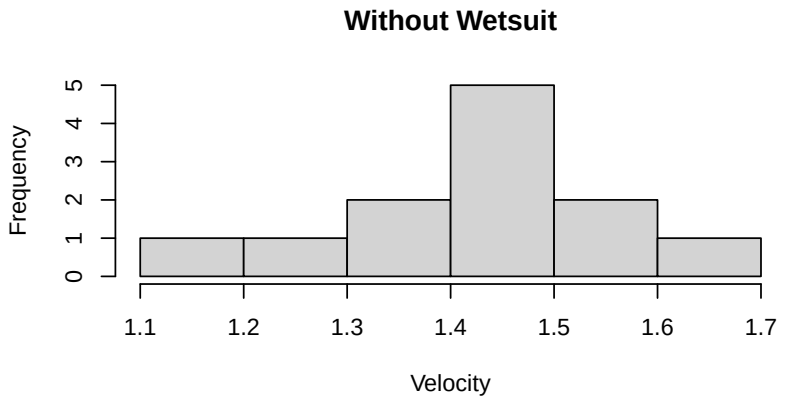
Wetsuits

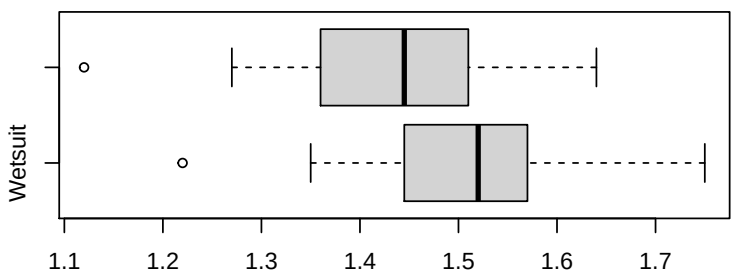
- We can use a *two-sample t-test* to compare the mean velocity with the wetsuit to the mean velocity without the wetsuit
  - $H_0 : \mu_1 = \mu_2$, or equivalently, $H_0 : \mu_1 - \mu_2 = 0$
- In order for a hypothesis test based upon the t-distribution to be appropriate, one of two conditions must be met:
  - The samples must be approximately Normally distributed
  - Or, the sample sizes must be relatively large ($n_1 \geq 30$ and $n_2 \geq 30$)

Wetsuits

The sample sizes ($n_1 = 12$ and $n_2 = 12$) aren't very large, but a graphical display of the data indicates the Normal assumption appears reasonable







The two-sample t-test

- Based upon results from the previous lecture, we expect the following T-value to follow a t-distribution:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df}$$

- Under $H_0 : \mu_1 - \mu_2 = 0$, and everything else is estimated from the sample data
- However, the degrees of freedom are somewhat complicated...

Degrees of Freedom (two-sample t-test)

- Gosset's approach assumes the standard deviation both groups is the same (ie: $\sigma_1 = \sigma_2$)
  - Thus, there's only one source of extra uncertainty (estimating the common standard deviation)
  - This approach is known as **Student's t-test**, it uses $n_1 + n_2 - 2$ degrees of freedom (one degree of freedom lost for each sample mean)

Degrees of Freedom (two-sample t-test)

- Gosset's approach assumes the standard deviation both groups is the same (ie: $\sigma_1 = \sigma_2$)
  - Thus, there's only one source of extra uncertainty (estimating the common standard deviation)
  - This approach is known as **Student's t-test**, it uses $n_1 + n_2 - 2$ degrees of freedom (one degree of freedom lost for each sample mean)
- A second approach was developed by BL Welch that assumes $\sigma_1 \neq \sigma_2$
  - Thus, there are two sources of uncertainty
  - This approach is know as **Welch's t-test**, and the degrees of freedom calculation is complicated (fortunately `R` will do it for us)

Practice

1) In `R`, perform a two-sample t-test using summary statistics from the `wetsuits` dataset to calculate a T-value, and then using `pt()` to find p-value
2) Then, use the `t.test()` function to repeat the same test. Pay attention to whether Student's or Welch's test is used by default.

Practice - solution (part 1)

```
## Sample Stats
xbar1 <- mean(wet$Wetsuit)
xbar2 <- mean(wet$NoWetsuit)
s1 <- sd(wet$Wetsuit)
s2 <- sd(wet$NoWetsuit)

## T-value
t_val <- (xbar1 - xbar2)/sqrt(s1^2/12 + s2^2/12)
t_val
```

```
## [1] 1.368791
```

```
## p-value
2*pt(t_val, df = 22, lower.tail = FALSE)
```

```
## [1] 0.1848798
```

Practice - solution (part 2)

```r
## Welch's Test
t.test(x = wet$Wetsuit, y = wet$NoWetsuit)
```

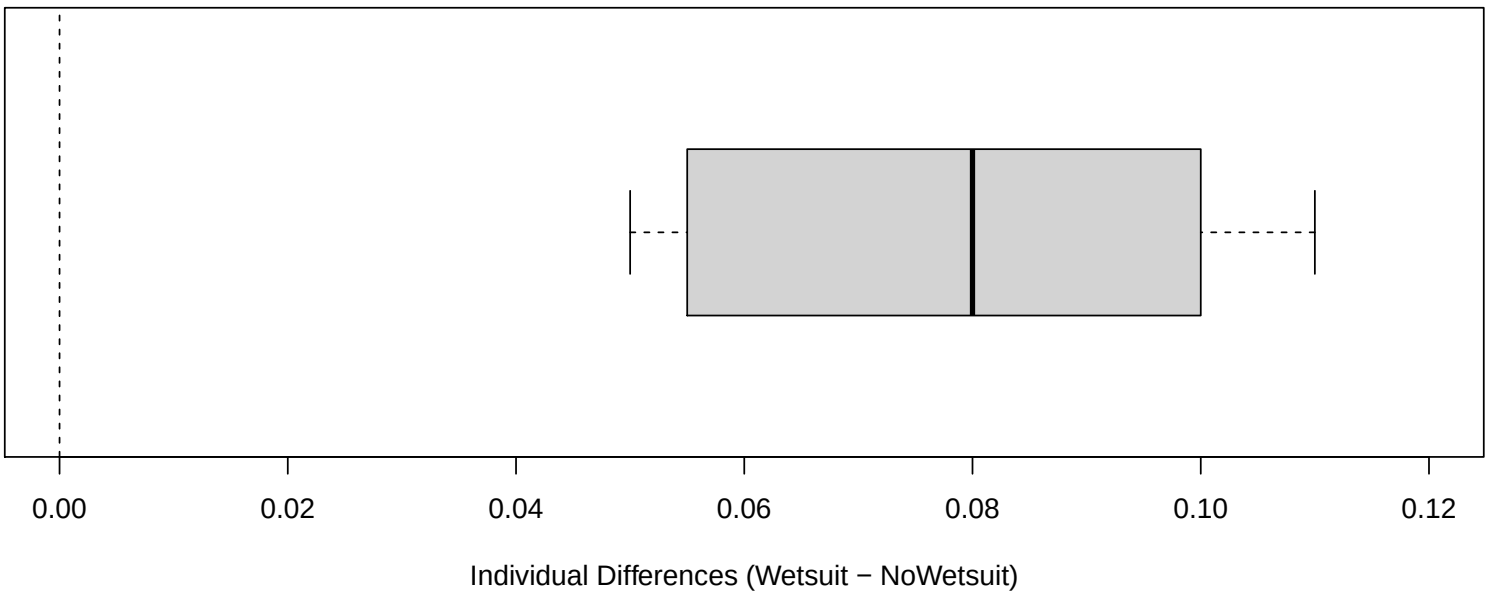
```
##
##  Welch Two Sample t-test
##
## data:  wet$Wetsuit and wet$NoWetsuit
## t = 1.3688, df = 21.974, p-value = 0.1849
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03992937  0.19492937
## sample estimates:
## mean of x mean of y
##  1.506667  1.429167
```

```r
## Student's Test
t.test(x = wet$Wetsuit, y = wet$NoWetsuit,
       var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  wet$Wetsuit and wet$NoWetsuit
## t = 1.3688, df = 22, p-value = 0.1849
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03992124  0.19492124
## sample estimates:
## mean of x mean of y
##  1.506667  1.429167
```

Interpretting the Results

- Both Welch's and Student's tests result in p-values of 0.1849
  - We conclude that there is insufficient evidence supporting the notion that high-tech wetsuits provide an advantage
- But how is that possible when *all* of the 12 study participants swam faster with the wetsuit...



Paired Data

- Our initial analysis (two-sample t-test) ignored the fact that these data are *paired*
  - Paired study designs offer a major statistical advantage - each subject serves as their own control, thereby blocking out variability between individuals
  - Put differently, we shouldn't be treating these data as two separate groups, instead we should be looking for *within subject* differences

Paired Data

- Our initial analysis (two-sample t-test) ignored the fact that these data are *paired*
  - Paired study designs offer a major statistical advantage - each subject serves as their own control, thereby blocking out variability between individuals
  - Put differently, we shouldn't be treating these data as two separate groups, instead we should be looking for *within subject* differences
- The implication is we should have used a one-sample t-test on the paired differences in swim velocity

$$\frac{\bar{x}_d - \mu_d}{s_d / \sqrt{n_{\text{pairs}}}} \sim t_{df=n-1}$$

Practice

1) Analyze the wetsuit data properly by creating a new variable called `diff`, and then performing a one-sample t-test using the summary statistics of this variable and the `pt()` function
2) Compare your results to using the `t.test()` function on the original data with the argument `paired = TRUE`

Practice - solution (part 1)

```
diff <- wet$Wetsuit - wet$NoWetsuit
xbar <- mean(diff)
s <- sd(diff)

t_val <- xbar/(s/sqrt(12))
t_val
```

```
## [1] 12.31815
```

```
2*pt(t_val, df = 11, lower.tail = FALSE)
```

```
## [1] 8.885414e-08
```

Practice - solution (part 2)

```r
## Using the paired argument
t.test(x = wet$Wetsuit, y = wet$NoWetsuit, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  wet$Wetsuit and wet$NoWetsuit
## t = 12.318, df = 11, p-value = 8.885e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06365244 0.09134756
## sample estimates:
## mean of the differences
##                  0.0775
```

```r
## Doing a true one-sample test
t.test(x = wet$Wetsuit - wet$NoWetsuit, mu = 0)
```

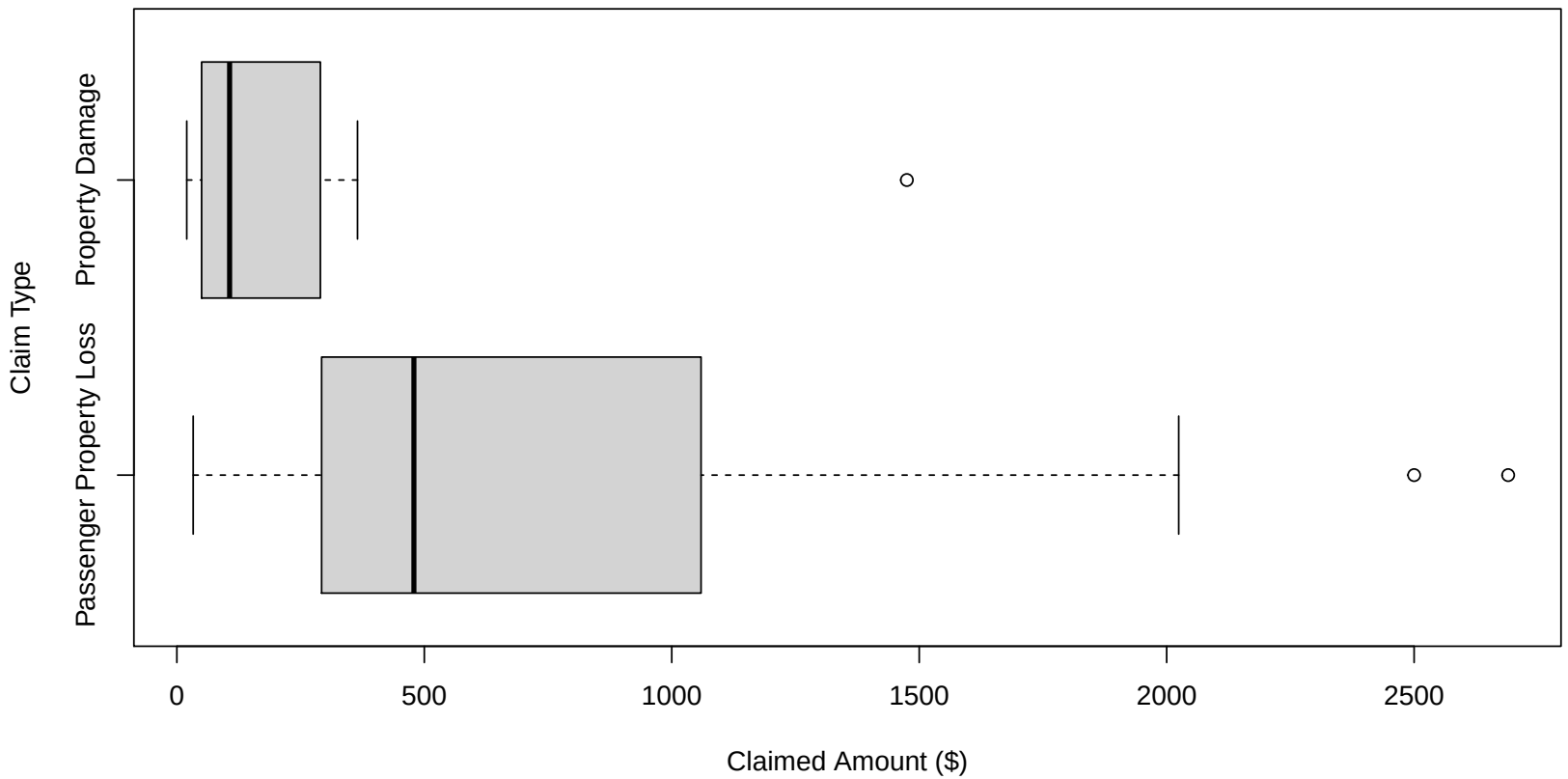
```
##
##  One Sample t-test
##
## data:  wet$Wetsuit - wet$NoWetsuit
## t = 12.318, df = 11, p-value = 8.885e-08
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.06365244 0.09134756
## sample estimates:
## mean of x
##    0.0775
```

Comments on Paired Designs

- Paired designs tend to be very powerful (notice $p \leq 0.0001$ vs. $p \approx 0.18$)
  - This is because they block out variability between individuals (ie: differences in swimming ability) and focus on variability within individuals
- This statistical advantage is accompanied by practical barriers, not every comparison can be paired
  - For example, you cannot perform two types of surgeries on the same person (Lister's experiment)

Permutation Tests

Below is a sample of 45 claims against the TSA, 23 of these claims were property loss, while 21 were property damage. The modest sample size and right-skew should make us uncomfortable



Permutation Tests

- An alternative method for statistically comparing the means of two groups is a **permutation test**, an approach related to simulation
  - The general idea is that if $\mu_1 = \mu_2$, the group labels (property loss vs. property damage) can be seen as random
  - Thus, we can *randomly reassign* these labels to simulate data we might have seen had H_0 been true

Permutation Tests

- An alternative method for statistically comparing the means of two groups is a **permutation test**, an approach related to simulation
  - The general idea is that if $\mu_1 = \mu_2$, the group labels (property loss vs. property damage) can be seen as random
  - Thus, we can *randomly reassign* these labels to simulate data we might have seen had H_0 been true
- Technically, a true permutation test considers each possible configuration of group labels, but simulation approaches that *randomly reassign* the labels yield similar results

Permutation Tests

Shown below is R code simulating the *permutation distribution*

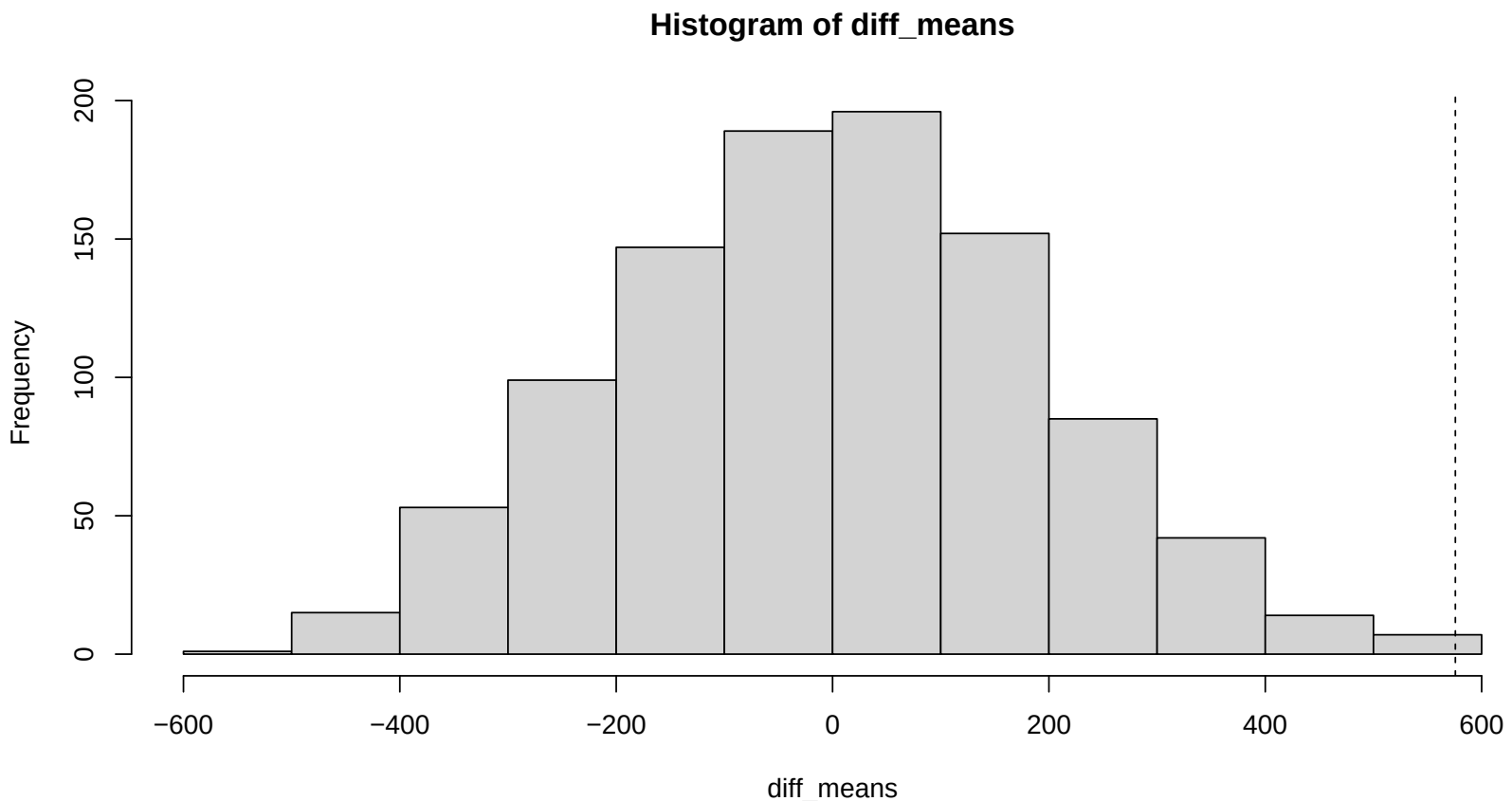
```
x <- tsa_sample$Claim_Amount
y <- tsa_sample$Claim_Type

diff_means <- numeric(1000)
for(i in 1:1000){
new_y <- sample(y)
diff_means[i] <-
  mean(x[new_y == "Passenger Property Loss"]) -
  mean(x[new_y == "Property Damage"])
}

obs_diff <- mean(x[y == "Passenger Property Loss"]) -
  mean(x[y == "Property Damage"])
```

Permutation Tests

```
hist(diff_means)
abline(v = obs_diff, lty = 2)
```



Permutation Tests

Simply tallying the proportion of permuted differences at least as extreme as the observed difference provides a p-value

```
p_val <- sum(diff_means >= obs_diff)/1000
2*p_val
```

```
## [1] 0
```

Thus, we can conclude with a high degree of statistical certainty that the mean property loss claim is higher than the mean property damage claim, despite the fact that these data did not allow us to use the t-test

Conclusion

- In this lecture we saw how the t-distribution can be used in hypothesis testing
- The one-sample t-test can be applied to a single numeric variable, or paired differences
  - It requires Normally distributed data, or a $n \geq 30$, and uses $T = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{df=n-1}$

Conclusion

- In this lecture we saw how the t-distribution can be used in hypothesis testing
- The one-sample t-test can be applied to a single numeric variable, or paired differences
  - It requires Normally distributed data, or a $n \geq 30$, and uses $T = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{df=n-1}$
- The two-sample t-test is used to compare the means of two groups
  - It requires both samples are Normally distributed, or $n_1 \geq 30$ and $n_2 \geq 30$, and uses $T = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{s_1^2/n_1+s_2^2/n_2}} \sim t_{df}$
  - Student's test assumes $\sigma_1 = \sigma_2$ and uses $df = n_1 + n_2 - 2$
  - Welch's test presumes $\sigma_1 \neq \sigma_2$ and uses a more complicated method for calculating df