Analysis of Categorical Data

NRES 710

Last compiled: 2024-09-23

## Summary

Up to this point, we have been operating under the idea that we have been examining the relationship between a **continuous Y variable** and a **continuous X variable**.

What if we no longer have a continuous X-variable, but instead have a **categorical X-variable**?

Let’s assume that we have our usual continuous Y-variable, but now our X-variable is categorical. And the X-variable is the most simple type of categorical variable: it is binomial (two categories):

* **Continuous Y, Categorical X (binomial)**

**Q:** How would we typically analyze data with a categorical X and a continuous, normally-distributed Y?

**t-Test!** Named after the ‘Student t distribution’.

## t-Tests

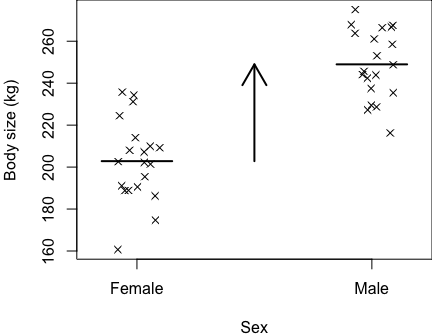
For example, let’s say we are interested in testing for body size differences between two sexes of **elephant seal** (*Mirounga leonina*).

Photo: Luke Verburgt

* **= male**
* **= female**
* **= size**

Our data might look like this (*Brian plot on board*):

Both the female and male data are normally distributed. I’m putting a slight ‘jitter’ on these points, so they can be visualized easier by the naked eye.

We want to know: how much larger are males than females?

*Brian add horizontal lines for means within each point cloud*

Or, what is the difference between the mean of females and males?

*Brian an arrow between the two means*

Our scientific effort might focus on testing the null hypothesis:

**: no difference between groups**

From a statistical perspective, we are interested in knowing whether males are significantly different than females in body size, and measuring that effect (if any).

Let’s assume that we know truth for this graph:

* **= male = 250 kg**
* **$X\_2 = female = 200 kg**
* **= 20 kg**

The standard deviation for the normal distribution of data for each sex is 20 kg.

The truth is that males are larger than females. Males are about 250 kg, females are about 200 kg, and there is some noise in our data.

Questions?

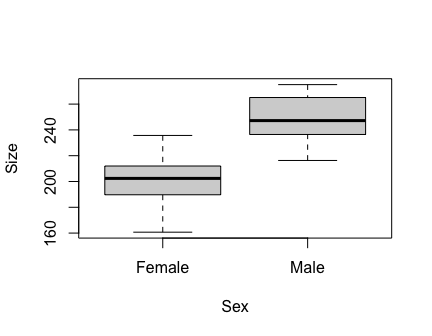
### Simple t-test in R

Let’s make these data in R, and run a t-test:

### Code for simulating data to be analyzed body size data for two sexes  
  
# Set the seed for reproducibility  
set.seed(123)  
  
# Simulate the female data  
n <- 20 # sample size  
Sex <- rep("Female", n) # female category  
Sigma <- 20 # standard deviation of the error (noise) in data  
Error <- rnorm(n, mean = 0, sd = Sigma) # simulate the error (noise)  
Size <- 200 + Error # make the response variable data  
FemaleData <- cbind(Sex, Size)  
  
# Simulate the male data  
n <- 20 # sample size  
Sex <- rep("Male", n) # female category  
Sigma <- 20 # standard deviation of the error (noise) in data  
Error <- rnorm(n, mean = 0, sd = Sigma) # simulate the error (noise)  
Size <- 250 + Error # make the response variable data  
MaleData <- cbind(Sex, Size)  
  
# Bind the data together  
datum <- data.frame(rbind(FemaleData, MaleData))  
  
# Make the 'Sex' variable a factor  
datum$Sex <- as.factor(datum$Sex)  
  
# Force to 'Size' to numeric and round to 1 digit after the decimal  
datum$Size <- round(as.numeric(datum$Size), 1)  
  
# Save the data  
write.csv(datum, "lecture\_7\_seal\_data.csv", row.names = FALSE)

I am going to show you all how to analyze these data using the ‘t.test()’ function in R!

### Load in the data  
datum <- read.csv("lecture\_7\_seal\_data.csv")  
  
# Examine the data  
head(datum)  
tail(datum)  
str(datum)

## 'data.frame': 40 obs. of 2 variables:  
## $ Sex : chr "Female" "Female" "Female" "Female" ...  
## $ Size: num 189 195 231 201 203 ...

# Force 'Sex' to be a factor  
datum$Sex <- as.factor(datum$Sex)  
  
# Plot the data to examine it!  
plot(Size ~ Sex, data = datum)

# Box and whiskers plot!  
# Bold black line = median  
# Edges of box: 75% and 25% quartiles  
# Bars: 95% limits  
# Points: outliers, or 5% of data outside of the 95% intervals  
  
# Analyze this using the 't.test()' function in R:  
help(t.test)  
results <- t.test(Size ~ Sex, data = datum)  
  
# Examine the summary  
summary(results)

## Length Class Mode   
## statistic 1 -none- numeric   
…

# Nothing pops up. Some functions in R don't have summary functions for them because they are so simple!  
  
# Just ask for the object  
results

##   
## Welch Two Sample t-test  
##   
## data: Size by Sex  
## t = -8.1, df = 37, p-value = 1e-09  
## alternative hypothesis: true difference in means between group Female and group Male is not equal to 0  
## 95 percent confidence interval:  
## -57.73 -34.56  
## sample estimates:  
## mean in group Female mean in group Male   
## 202.8 249.0

All the results come from the object itself!

* t-statistic
* degrees of freedom
* p-value – testing the null hypothesis that there is no difference in the means between the two groups
* 95% confidence intervals – on the difference between the two groups
* average size of female group
* average size of male group

**Q:** What is the one thing it doesn’t give us that we might want to see??

It does not provide an **estimate** of the size difference between two groups!! This is not helpful (and somewhat silly)!

* It gives us confidence intervals around the estimate of the difference between the two groups, but it does not give us the estimate. Really weird.

We can calculate it by doing some math… subtract the smaller group from the larger group:

# Difference between groups  
248.9 - 202.8

## [1] 46.1

This is how much larger the males are than females.

It provides us *confidence limits*, but it does not provide us confidence interval. We can calculate those with:

# Use confidence limits of difference between groups to calculate confidence intervals  
(57.73 - 34.55)/2

## [1] 11.59

# Males are 46.1 kg (+/-11.6; 95% CI) heavier in size than females (p = 1.101e-09).

**Q:** Why is the 95% confidence interval that R provided negative?

R calculated this effect as doing females minus males, whereas I calculated it as males minus females. We just need to be savvy to make sure we think about the outputs and ensure that everything makes sense. Use common sense.

**Questions?**

## t-tests are just a regression

One thing that always confused me about learning statistics as an undergraduate and graduate student was that all of the usual analyses (regression, t-test, ANOVA, ANCOVA, etc.) are all taught as ‘different tests’. It creates many more boxes and relationships that you have to memorize – different things to categorize in your brain as different and thus requires more effect to memorize.

Here’s a little secret… t-tests are just a **regression**! And can be analyzed as such with ‘lm()’!

**Q:** What’s our linear model again??

We can use the **linear model** to analyze the seal size data!

Questions you might ask yourself:

* Why teach this as a whole-new test, when it’s the same mathematical formula we have looked at all along?
* And, since sex is not a number, how would this work?

It works through the magical process of **‘dummy-coding’**!

**Dummy-coding – a process to convert categories to 1s and 0s.** For example:

**Sex:** **Male:**

Male 1  
Male 1  
Female 0  
Female 0

In the dummy-coded variable, you can arbitrarily decide which category gets 1s and which category gets 0s. That’s up to you. It will only influence whether the effects are estimated as positive or negative, depending on the directionality.

When you create a dummy-coded variable, I recommend making the variable name whatever category was given ‘1’ – it’s the most clear.

Now, we can replace our **sex** variable with **male**:

Let’s examine how this works mathematically. Anytime a sample is a male, it gives us a 1 here. Anytime a sample is female, it gives us a 0. Using this information, we can simplify this formula and see how it gives us the answer for both sexes.

If we want to know the size of females, we get:

If we want to know the size of males, we get:

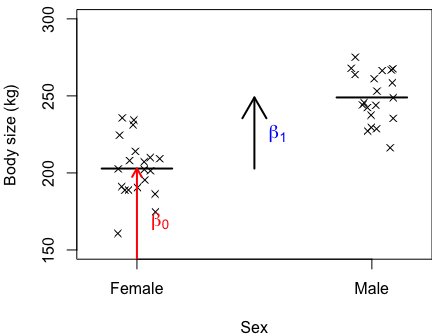
**Q:** Everyone follow what we just did here?

**Q:** Looking at these equations, what is the meaning of ?

* **Average size of females**
* **Or, more generically, the average Y of the reference group** – samples that were assigned 0.

**Q:** Looking at these equations, what is the meaning of ?

* **Difference between groups** – this is what we want to know! The thing that the null hypothesis is testing – whether this effect is different than zero or not. This is the thing the ‘t.test()’ function didn’t even measure for us… (what a shame!)

Let’s go back to our original graph.

**Q:** how might we visualize and on this graph?

If we remember that female is 0, then is the y-intercept, or the body size when X = 0.

And then is the difference between these two groups.

**Q:** How do we calculate slope again? Rise over run.

Since males are 1, then the ‘run’ is from 0 to 1 – which equals 1. The rise is divided by 1, which equals .

So is still the slope, but it’s also more simply just the difference between the groups.

**Question?**

So: why think about this as a t-test? Perhaps it would be easier to use the linear model that we have studied for three weeks, and gotten to know pretty well – and instead say that our X-variable is categorical, rather than continuous.

Let’s see what this looks like by trying this again in R.

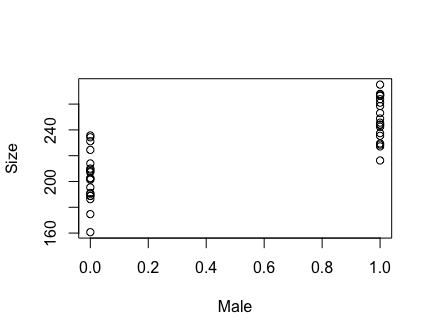
## Analysis of categorical data w/ lm()

### Code for simulating data to be analyzed body size data for two sexes  
  
# Recall our 'datum' object  
head(datum)  
tail(datum)  
  
# We need to 'dummy-code' our Sex variable, e.g., as 'Male'  
Male <- c(rep(0, n), rep(1, n))  
  
# Add 'Male' to the dataframe  
datum <- cbind(datum, Male)

Now that we have a dummy-coded variable, we can simulate our data much easier:

# Aside: we can use dummy-coded sex variable to simulate data much easier  
Size <- 200 + 50 \* Male + rnorm(n\*2, 0, Sigma)  
Size

## [1] 207.6 190.0 193.3 179.6 178.6 206.1 209.0 201.1 218.4 241.0 190.2 153.8 220.1 185.8 186.2  
## [16] 220.5 194.3 175.6 203.6 197.2 250.1 257.7 242.6 262.9 245.6 256.6 271.9 258.7 243.5 273.0  
## [31] 269.9 261.0 254.8 237.4 277.2 238.0 293.7 280.7 245.3 229.5

Plot the data

# Plot the data!  
plot(Size ~ Male, data = datum)

Note: previously R made the X-axis as a categorical (female vs. male), but now it’s continuous… R automatically makes graphs depending on it’s default interpretation of the data. It reads characters/letters as ‘categorical’ and numbers and continuous variables. Since we swapped this to be dummy-coded with 1s and 0s, it automatically scaled the X-axis as continuous!

Let’s now try to run our t-test using ‘lm()’:

# Examine the old t-test results  
results

##   
## Welch Two Sample t-test  
##   
## data: Size by Sex  
## t = -8.1, df = 37, p-value = 1e-09  
## alternative hypothesis: true difference in means between group Female and group Male is not equal to 0  
## 95 percent confidence interval:  
## -57.73 -34.56  
## sample estimates:  
## mean in group Female mean in group Male   
## 202.8 249.0

248.9 - 202.8 # effect of being male

## [1] 46.1

# Use lm() to run regression with dummy-coded X-data  
results2 <- lm(Size ~ Male, data = datum)  
summary(results2)

##   
## Call:  
## lm(formula = Size ~ Male, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.14 -12.65 -0.43 12.80 32.86   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 202.83 4.04 50.16 < 2e-16 \*\*\*  
## Male 46.15 5.72 8.07 9.2e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.1 on 38 degrees of freedom  
## Multiple R-squared: 0.632, Adjusted R-squared: 0.622   
## F-statistic: 65.1 on 1 and 38 DF, p-value: 9.23e-10

**Q:** What do we see??

* Average size of females: 202.8 from t-test, compared to intercept 202.8 from regression
* Effect of being male: 46.1 from t-test, 46.1 from regression

The numbers are the same!

Turns out that when you run ‘t.test()’, it does dummy-coded behind the scenes, and then just runs a ‘lm()’.

# Confidence intervals  
confint(results2)

## 2.5 % 97.5 %  
## (Intercept) 194.65 211.02  
## Male 34.57 57.72

Confidence intervals also ~match up between t-test and lm results.

What if we run a regression but with a categorical X-variable…?

# Regression with a categorical variable  
results3 <- lm(Size ~ Sex, data = datum)  
summary(results3)

##   
## Call:  
## lm(formula = Size ~ Sex, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.14 -12.65 -0.43 12.80 32.86   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 202.83 4.04 50.16 < 2e-16 \*\*\*  
## SexMale 46.15 5.72 8.07 9.2e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.1 on 38 degrees of freedom  
## Multiple R-squared: 0.632, Adjusted R-squared: 0.622   
## F-statistic: 65.1 on 1 and 38 DF, p-value: 9.23e-10

This works also! Because R will automatically turn that categorical X-variable into a dummy-coded continuous variable. In this case, it became 1 for ‘SexMale’ and 0 for ‘SexFemale’. It chose ‘Female’ as the reference group, and male becomes the group females are being compared to.

**Q:** Why did it choose ‘Female’ as the reference? Alphabetical order.

So, you don’t even have to do the dummy-coding – R can do that for you! (Although I still like to, to be sure I know what’s going on.)

## Conclusions

What this means is: we don’t actually have to memorize “Do I do a t-test here, or a regression here, for these data…?” It doesn’t matter!

The lightbulb turned on for me with statistics when I realized that we don’t have to remember all the different situations for which you would run a t-test, ANOVA, regression, ANCOVA, etc…

Instead, ***every analysis can be done with some form of a linear model***.

For the rest of the semester, we will continue to revisit our tried-and-true formula for the linear model:

And then we will tweak it just a little bit, week by week, until we use this basic formula for the entire class. It will be a little more complicated each week, but it will be easier to understand because we will build on it slowly, carefully, and hopefully in a logical way!

We will learn how to tweak this formula to analyze our data as the data become more complicated.

**Note:** We can’t use our same canned statement for reporting results like we used for linear regression. Next class we will learn how to report our results when X-variable is categorical. And then we will extend the formula to accommodate situations for when we have more than two groups. And then we will keep moving on from there!

[–go to next lecture–](lecture_8.html)