Analysis of Categorical Data (cont.)

NRES 710

Last compiled: 2024-10-01

## Review

I asked you all to read Ruxton & Beauchamp (2008). My goal was mostly to provide you all with some background context and considerations related to post-hoc tests before this lecture. We’ll mention some key points from this paper in class today.

Last class we discussed **‘Analysis of Variance (ANOVA)’** – the classic test which partitions the total sum of squares in Y into both the error and the differences between the groups. This does the same thing as regression, but now we have categorical X variables.

* **Continuous Y**
* **Categorical X with >2 groups**

If you only had two groups, you would just use a t-test – although we have learned that both t-tests and ANOVA are really the same analysis using a linear model.

Most of the lecture today is going to focus on historical approaches to using post-hoc tests. These studies often were experimental in nature, perhaps in agricultural settings, and involved setting up an experimental manipulation of something in the field. Those experiments often involved collecting data and then running an ANOVA.

The ANOVA result provide a **single p-value – the significance of variable as a whole**. Today we will look at data describing the mass of an animal across four seasons as an example. The single ANOVA p-value will give us the significance of this entire ‘season’ variable. Historically speaking, researchers would often respond to that p-value in two ways:

* **p > 0.05 – none of the category groups are statistically significantly different**
  + In this case, you would no longer pursue any further statistical testing.
  + I don’t like this approach! You don’t know whether your lack of result was due to a lack of sample size or due to a lack of effect size.
  + But this path was often followed to avoid doing unnecessary tests and to **minimize the Type I error rate**. (And also why we do ANOVA instead of jumping into a bunch of pairwise t-tests.)
* **p < 0.05 – at least two groups are statistically significantly different**
  + In this case, you would then try to figure out which groups were the different ones using **post-hoc tests**.

## Post-hoc tests

**Post-hoc test – “after the fact” test**

* There is some confusion about post-hoc tests, as many people think that the purpose of post-hoc tests is to test for all pairwise comparisons. In fact, we don’t need a post-hoc test to do this, because we can easily test for all pairwise comparisons in our linear model using dummy-coded variables.
* **Purpose – artificially inflate p-values in order to maintain the experiment-wide Type I error rate**.
* Post-hoc tests do not changes estimates or differences (s) between groups. Instead, they only affect p-values and confidence intervals, because these are both calculated using the same underlying distribution.
* Because I don’t much care for p-values, I don’t really like post-hoc tests.
* But, I feel like it’s important to introduce them in this class, given their large history in the field of statistics, and that you likely have or will encounter them.

So you do an ANOVA, and then you need to do a post-hoc tests. How do you do that? First step, is to choose which post-hoc test to run. There are many!

**Q:** What are post-hoc tests that you are aware of??

Tukey test, Fisher’s Least Significance Difference test (aka, Fisher’s LSD), Dunnett, Schauffe, many others, etc.

Ruxton & Beauchamp (2008) identified 10 different post-hoc tests used in their survey.

The reason why there are so many post-hoc tests is folks are trying to find an optimal **balance between the Type I error rate and Power**.

* Type I error is when we incorrectly reject the null hypothesis.
* Power is our ability to reject the null when a significant difference does truly exists.

Type I error rate and Power are inversely related. For example, let’s say we use a p-value of 0.01 as a cutoff. This decreases the chance that we will commit Type I error from 0.05 to 0.01, and we will be very unlikely to commit Type I error. But, we lose Power to detect significant differences when they likely exist! This balance represents a tradeoff between “getting things wrong” and “learning about differences”.

In an idea world where we do a post-hoc test, we would have a Type I error rate of 0.05 and also have some theoretically-determined maximum power. Statisticians at universities trying to be important, and get tenure, try to develop their own post-hoc test to do this…

A commonly used post-hoc test is **Tukey’s post-hoc test** – has pretty good power! But it has a higher Type I error rate (0.06, 0.07). This causes weird circumstances where you had a marginally significant ANOVA (e.g., 0.049), but Tukey’s test fails to detect significant differences between pairwise comparisons.

All of the above tests try to balance this tradeoff, and all of them vary at it. And whichever one you choose is **arbitrary**. When you have a marginally significant result, you might start chasing pairwise significant results by using different post-hoc tests. Let’s try Tukeys! Oh, that didn’t work. Let’s try Schauffe’s! Nope, that didn’t work either. And then you try another, and it gives you two significant comparisons.

This causes us to go on a ‘fishing expedition’, which, in my opinion, is not a very good approach to science. We should use statistics to test specific *a priori* hypotheses that we have developed for a good reason – not go chasing after results we have no reason to suspect exist in reality. Ruxton & Beauchamp (2008) made comments to this effect a few times.

The strange thing is that no matter which post-hoc test you used, your results are all the same: the effects (betas) between groups have never changed, because post-hoc tests have not changed that. The confidence intervals and p-values will be similar. Was this marginally different inference worth your time and effort?

The main point I want to make here: don’t get too hung up on post-hoc tests. Instead, let’s focus on the estimates of effects, the confidence intervals, and whether or not those are statistically significant.

## Tukey’s HSD Test

**Tukey’s Honest Significant Difference (HSD) Test** – harder to get significant differences, but less likely to commit Type I error.

To explore this, let’s use some data that I made in R. These data simulate the body mass of Greater Sage-grouse (*Centrocurcus urophasianus*). You can download it [here](lecture_9_seasons.csv), and a script to simulate it is included at the bottom of the page.

* **Y-variable – mass**
* **X-variable – season (4 groups)**

Picture: Bert Filemyr

Let’s take a look at it now.

# Load and examine the data  
datum <- read.csv("lecture\_9\_seasons.csv")  
head(datum)

The data have the mass of animals measured in four different season: spring, summer, fall, and winter. We have a categorical X-variable “Season”, and also dummy-coded variables for each of the individual seasons.

### The number of comparisons

Our goal is to estimate if mass of the animal is different between the seasons. How many comparisons might we need to make?

Here’s a trick to do it!

Fall  
Spring  
Summer  
Winter

1. Draw a line connecting all the neighbors.
2. Draw a line connecting all the two-neighbors away.
3. Draw a line connecting all the three-neighbors away.
4. Etc.
5. Then add up all the lines!

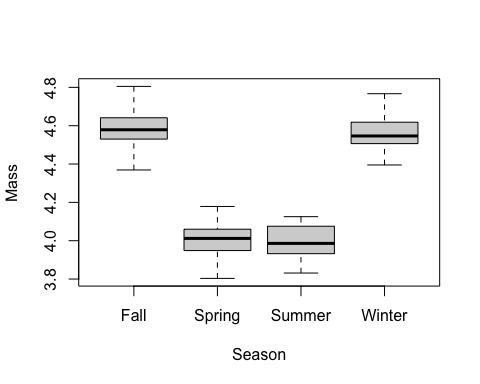
In this case, we might have to do **6 pairwise comparisons**.

### Linear model

Let’s write out our linear model:

**Review each of the betas**

* – average mass in the fall
* – difference in mass between spring and fall
* – difference in mass between summer and fall
* – difference in mass between winter and fall
* If we want to know other differences, we have to change the references and re-run.

What are the ‘true’ values?

* **= 4.6**
* **= = -0.6**
* **= -0.05**

### Analysis in R

Let’s analyze these data in R:

# Plot the data  
datum$Season <- factor(datum$Season)  
plot(Mass ~ Season, data = datum)

# Fit a linear model  
results <- lm(Mass ~ Season, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Mass ~ Season, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.2189 -0.0588 -0.0051 0.0577 0.2170   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.5880 0.0209 219.40 <2e-16 \*\*\*  
## SeasonSpring -0.5738 0.0296 -19.40 <2e-16 \*\*\*  
## SeasonSummer -0.5931 0.0296 -20.06 <2e-16 \*\*\*  
## SeasonWinter -0.0274 0.0296 -0.93 0.36   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0935 on 76 degrees of freedom  
## Multiple R-squared: 0.907, Adjusted R-squared: 0.904   
## F-statistic: 248 on 3 and 76 DF, p-value: <2e-16

Because we only have one variable (‘Season’), the p-value in the bottom right is the same as the p-value we would get running an ANOVA. This gives us the significance of the season variable as a whole: it tells us that at least two groups within this variale are different from each other.

When we examining the significance of effects within the model, we can see that mass in the Spring is different from Fall and that mass in the Summer is different from Fall.

**Trick question:** Is Winter different from Fall?

Yes, it is! We know that Winter is different from Fall, because we made the data and we simulated it to have an effect of -0.05. However, we are not able to detect that effect with this statistical analysis. Just because it was not statistically significant does not mean there is no biological significance.

**Q:** If we wanted to be sure to detect this, what would we have to do? **Increase our sample size.** But otherwise: the betas are all pretty close to Truth

**Q:** How would we test for the difference between Spring and Summer? **Change the reference.**

# Fit a linear model  
results2 <- lm(Mass ~ relevel(Season, ref="Summer"), data = datum)  
summary(results2)

##   
## Call:  
## lm(formula = Mass ~ relevel(Season, ref = "Summer"), data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.2189 -0.0588 -0.0051 0.0577 0.2170   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 3.9949 0.0209 191.04 <2e-16  
## relevel(Season, ref = "Summer")Fall 0.5931 0.0296 20.06 <2e-16  
## relevel(Season, ref = "Summer")Spring 0.0193 0.0296 0.65 0.52  
## relevel(Season, ref = "Summer")Winter 0.5658 0.0296 19.13 <2e-16  
##   
## (Intercept) \*\*\*  
## relevel(Season, ref = "Summer")Fall \*\*\*  
## relevel(Season, ref = "Summer")Spring   
## relevel(Season, ref = "Summer")Winter \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0935 on 76 degrees of freedom  
## Multiple R-squared: 0.907, Adjusted R-squared: 0.904   
## F-statistic: 248 on 3 and 76 DF, p-value: <2e-16

Now we can see other differences: summer and spring, summer and winter.

We just have to re-run this one more time to get the final comparison that we need: winter and spring.

### Tukey’s HSD

Let’s start by examining the help file for ‘TukeyHSD()’.

# Help file  
help(TukeyHSD)  
  
# Tukey requires an ANOVA output  
results3 <- aov(Mass ~ Season, data = datum) # or  
results3 <- aov(results)  
summary(results3)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Season 3 6.50 2.168 248 <2e-16 \*\*\*  
## Residuals 76 0.66 0.009   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# This analysis is so simple that it doesn't even have a summary file. Run it directly.  
TukeyHSD(results3)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = results)  
##   
## $Season  
## diff lwr upr p adj  
## Spring-Fall -0.57385 -0.65153 -0.49616 0.0000  
## Summer-Fall -0.59313 -0.67082 -0.51545 0.0000  
## Winter-Fall -0.02736 -0.10504 0.05032 0.7915  
## Summer-Spring -0.01929 -0.09697 0.05839 0.9145  
## Winter-Spring 0.54649 0.46880 0.62417 0.0000  
## Winter-Summer 0.56577 0.48809 0.64346 0.0000

Let’s compare that to our original ‘lm()’ results:

# Compare original lm() to TukeyHSD() results  
summary(results)

##   
## Call:  
## lm(formula = Mass ~ Season, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.2189 -0.0588 -0.0051 0.0577 0.2170   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.5880 0.0209 219.40 <2e-16 \*\*\*  
## SeasonSpring -0.5738 0.0296 -19.40 <2e-16 \*\*\*  
## SeasonSummer -0.5931 0.0296 -20.06 <2e-16 \*\*\*  
## SeasonWinter -0.0274 0.0296 -0.93 0.36   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.0935 on 76 degrees of freedom  
## Multiple R-squared: 0.907, Adjusted R-squared: 0.904   
## F-statistic: 248 on 3 and 76 DF, p-value: <2e-16

* The difference between spring and fall is: -0.57
* The difference between summer and fall is: -0.59
* Point being: Tukey’s HSD does not change the betas!

The ‘diff’ is the difference between the groups. The first columns tells us those exact differences: e.g., “Spring - Fall”.

We can use the ‘lwr’ and ‘upr’ to calculate 95% CI, as we have done before.

Let’s look at the p-values.

* Winter-to-Fall was 0.358, but Tukey’s says it is 0.79.
* Etc.

Tukey’s causes **p-values get larger**.

Let’s look at confidence intervals:

TukeyHSD(results3)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = results)  
##   
## $Season  
## diff lwr upr p adj  
## Spring-Fall -0.57385 -0.65153 -0.49616 0.0000  
## Summer-Fall -0.59313 -0.67082 -0.51545 0.0000  
## Winter-Fall -0.02736 -0.10504 0.05032 0.7915  
## Summer-Spring -0.01929 -0.09697 0.05839 0.9145  
## Winter-Spring 0.54649 0.46880 0.62417 0.0000  
## Winter-Summer 0.56577 0.48809 0.64346 0.0000

confint(results)

## 2.5 % 97.5 %  
## (Intercept) 4.54636 4.62966  
## SeasonSpring -0.63275 -0.51495  
## SeasonSummer -0.65203 -0.53423  
## SeasonWinter -0.08626 0.03154

Compare Spring-Fall from both outputs.

# Spring to Fall  
(-0.63274618 - -0.5149456)/2 # Linear model

## [1] -0.0589

(-0.65152888 - -0.49616295)/2 #Tukey HSD

## [1] -0.07768

Tukey causes **confidence intervals to become wider**.

**Takehome:** Both the p-values and confidence intervals have become artificially inflated.

### Cautionary tale

Sometimes people take a categorical variable and they turn it into a continuous variable. This can be dangerous! Take a look:

# Look at the second to last column  
head(datum)

They then run a model with that as the predictor variable.

# Run a 'lm()' with that  
results4 <- lm(Mass ~ SeasonN, data = datum)  
summary(results4)

##   
## Call:  
## lm(formula = Mass ~ SeasonN, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.370 -0.109 -0.001 0.122 0.361   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.8709 0.0411 118.4 <2e-16 \*\*\*  
## SeasonN -0.2326 0.0150 -15.5 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.15 on 78 degrees of freedom  
## Multiple R-squared: 0.755, Adjusted R-squared: 0.751   
## F-statistic: 240 on 1 and 78 DF, p-value: <2e-16

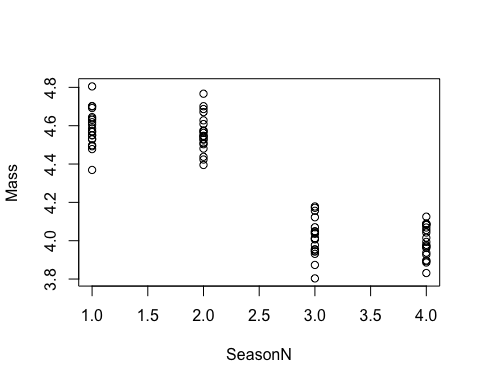
Season is significant! They write that in their paper and send it in for publication and they are done.

But they made a lot of errors here.

* First, they focused too much on p-values.
* If they would have been thinking about effects between groups (betas), they might have realized the issue.

**Q:** What did R do?

It ran a regression with X as a continuous variable in addition to the continuous Y. This assumes there is a linear relationship between 1, 2, 3, 4 and mass. That might be valid… but it isn’t!

# Scatterplot  
plot(Mass ~ SeasonN, data = datum)

**Q:** How do we know it ran it as a regression?

For one, it only gave us a single beta. (Iit only used 1 degree of freedom, but should have been 3).

A post-hoc test… won’t work:

# Scatterplot  
#TukeyHSD(aov(results4)) # doesn't work!  
# Error in `TukeyHSD.aov()`:  
# ! no factors in the fitted model

R is trying to tell us that it wants ‘Season’ to be a ‘factor’.

**factor – a categorical X-variable**

**covariate – an X-variable in a statistical model**

Two terms we should be aware of.

## Summary

We need to know that there are these things called ‘post-hoc tests’.

They try to maintain the ‘experiment-wide error rate’ (0.05 Type 1 error rate), and are still used a lot, particularly in classic agricultural or fisheries journals if you have done a manipulative experiment.

As you build more complicated models that have more complicated models (with both categorical and continuous variables), post-hoc tests won’t work.

Bonferonni corrections (basic math) can be used if absolutely necessary.