CM 5.5 - Two-Way ANOVA

Math 530/630

Logistics

- A complete knitted `html` file is due on Sakai by the beginning of the next class.

Two-way ANOVA

Today we are going to look at a two-way ANOVA, which involves still just one DV but now two IVs. If we look at all factors *and* their interactions, we get a two-way factorial ANOVA. In this design, we estimate two simple effects: the main effect of factor A (with levels $j = 1, \ldots, a$), and the main effect of factor B (with levels $k = 1, \ldots, b$). In addition, we estimate the interaction effect, AxB.

So, what are the null and alternative hypotheses we can test with a two-way factorial ANOVA?

The H_0 (null hypothesis) family:	The H_1 (alternative hypothesis) family:
H_{0_A}: no effect of factor A	H_{1_A}: significant effect of factor A
H_{0_B}: no effect of factor B	H_{1_B}: significant effect of factor B
$H_{0_{A*B}}$: no interaction between factors A and B	$H_{1_{A*B}}$: significant interaction between factors A and B

Sum of Squares

This makes the sums of squares considerably more complicated. Now, we need to break down the SS_{model} into three separable sums of squares: SS_A, SS_B, and SS_{AB}.

$$SS_{total} = \sum_{i}^{n}\sum_{j}^{a}\sum_{k}^{b}(y_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2$$

With degrees of freedom $abn - 1$.

$$SS_{model} = n\sum_{j}^{a}\sum_{k}^{b}(\bar{y}_{\bullet jk} - \bar{y}_{\bullet\bullet\bullet})^2$$

With degrees of freedom $ab - 1$.

SS_A	SS_B	SS_{AB}	$SS_{residual}$
$nb\sum_j^a(\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$	$na\sum_k^b(\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet})^2$	$SS_{model} - SS_A - SS_B$	$SS_{total} - SS_{model}$

df_A	df_B	df_{AB}	$df_{residual}$
$a - 1$	$b - 1$	$(a - 1)(b - 1)$	$ab(n - 1)$

Example - Ethnicity and Gender

This dataset includes real data on state expenditures for N=1000 individuals from the State of California's Department of Developmental Services. This department provides services and supports to individuals with developmental disabilities including intellectual disability, cerebral palsy, epilepsy, autism, and other disorders. The dataset was created and analyzed for an alleged case of discrimination privileging White non-Hispanics over Hispanics in the state's expenditures (note that some data have been altered to protect the rights and privacy of individuals). Based on initial analyses, it appeared that discrimination existed. Our task will be to evaluate this conclusion. This dataset includes expenditures (**exp**; the key DV), **age** (0-95 years), gender (**sex**; 2 levels: Male or Female), race/ethnicity (**eth**; 7 levels: White not Hispanic, Hispanic, Asian, Black, American Indian, Multi-race, and Other). Here, expenditures reflect spending by the state per individual for services such as respite care for families, psychological services, medical expenses, transportation, and costs related to housing such as rent for disabled adults.

You'll need these packages:

```
library(readr)
library(dplyr)
library(janitor)
library(ggplot2)
```

First, load in the dataset and let's take a look.

```
cadds <- read_csv(here::here("data", "cadds.csv"), col_types = cols(
  sex = col_factor(levels = NULL),
  eth = col_factor(levels = NULL)
))
glimpse(cadds)
```

```
## Observations: 1,000
## Variables: 5
## $ subject <dbl> 10210, 10409, 10486, 10538, 10568, 10690, 10711, 10778...
## $ age     <dbl> 17, 37, 3, 19, 13, 15, 13, 17, 14, 13, 13, 14, 15, 17,...
## $ sex     <fct> 2, 1, 1, 2, 1, 2, 2, 1, 2, 1, 2, 1, 2, 1, 1, 2, 2, 1, ...
## $ exp     <dbl> 2113, 41924, 1454, 6400, 4412, 4566, 3915, 3873, 5021,...
## $ eth     <fct> White not Hispanic, White not Hispanic, Hispanic, Hisp...
```

Univariate Statistics: Ethnicity (Factor A)

```
cadds %>%
  tabyl(eth)
```

```
##                 eth   n percent
##  White not Hispanic 401   0.401
##            Hispanic 376   0.376
##               Black  59   0.059
##          Multi Race  26   0.026
##               Asian 129   0.129
##     American Indian   4   0.004
##               Other   2   0.002
##     Native Hawaiian   3   0.003
```

We have some groups with pretty small n's. Specifically, there are only n=4 American Indians, n=26 Multi-race, n=3 Native Hawaiians, and n=2 "Other". We have two choices: lump (with another category)

or dump (from the entire analysis). For simplicity, I am going to rename "White not Hispanic" as "WNH", leave "Hispanic" as is, and lump all other designations into an "Other" category.

```
cadds <- cadds %>%
  mutate(eth3 = case_when(
    eth == c("White not Hispanic") ~ "WNH",
    eth == c("Hispanic") ~ "Hispanic"
  )) %>%
  mutate(eth3 = ifelse(is.na(eth3), "Other", eth3),
         eth3 = as.factor(eth3)
  )

cadds %>%
  tabyl(eth3)
```

```
##      eth3   n percent
##  Hispanic 376    0.38
##     Other 223    0.22
##       WNH 401    0.40
```

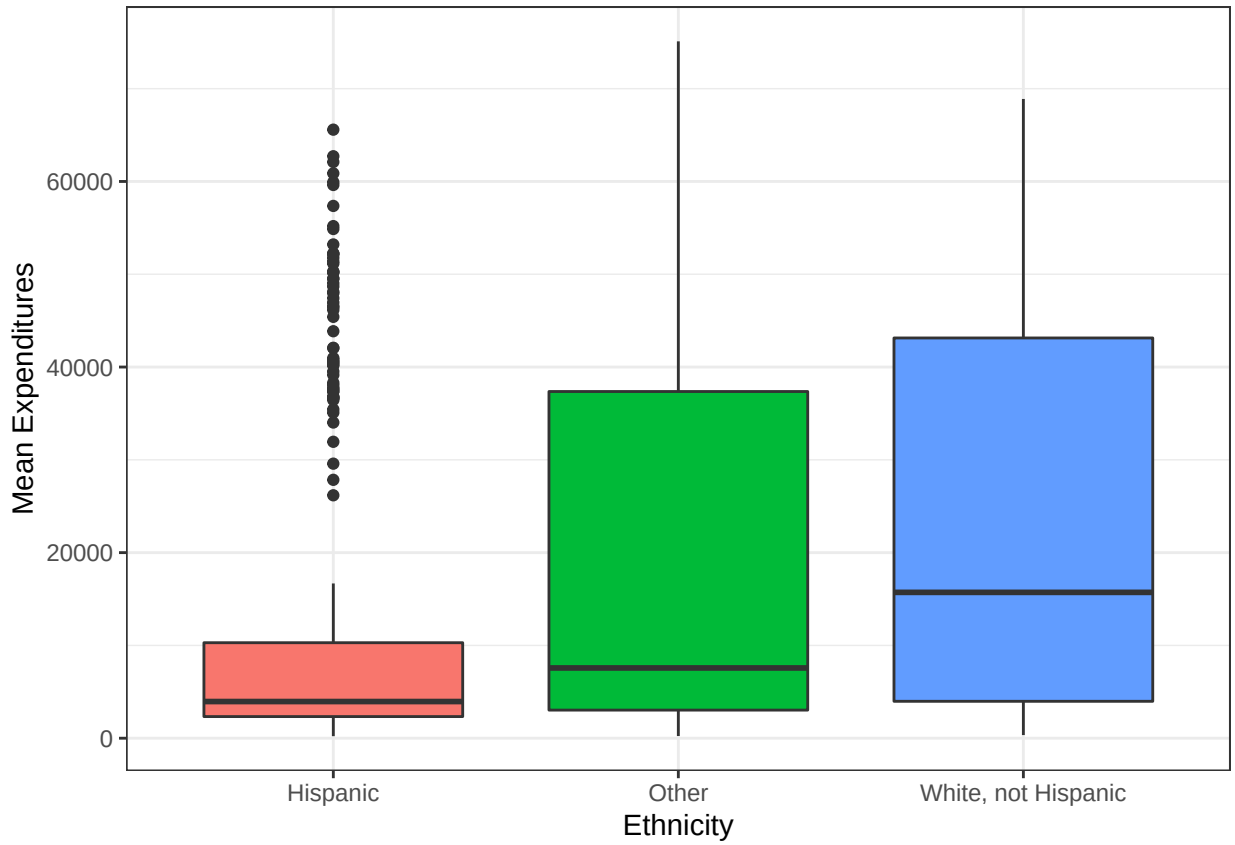
OK, that is better. Note that we still have widely varying group sizes, which means that the ANOVA we conduct will be an unbalanced (unequal group sizes), two-way (two IVs), between-groups (each group is an independent sample) univariate (one DV) ANOVA. Make sure you understand each of these labels thoroughly. Now, let's look at some descriptives for our j=3 groups:

```
cadds %>%
  group_by(eth3) %>%
  summarise(mean = mean(exp, na.rm =TRUE),
            sd = sd(exp, na.rm = TRUE),
            n = n())
```

```
## # A tibble: 3 x 4
##   eth3       mean     sd     n
##   <fct>     <dbl>  <dbl> <int>
## 1 Hispanic 11066. 15630.   376
## 2 Other    17944. 19458.   223
## 3 WNH      24698. 20604.   401
```

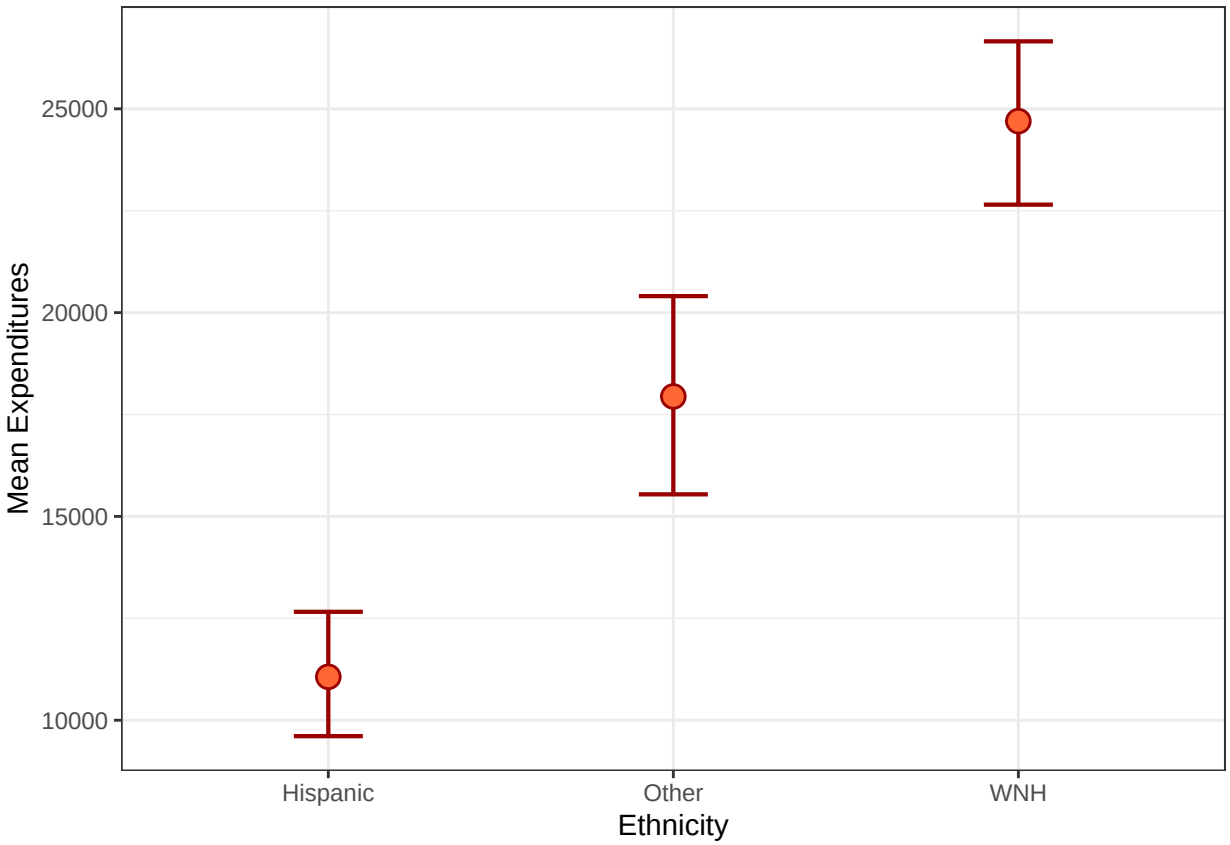
As far as data visualizations go, a boxplot and line graph of the means are helpful at this stage. I might be worried about heterogeneity of variances based on the boxplots and the standard deviations in each group as well, but I'm going to keep trucking forward at this point.

```
ethbox <- ggplot(cadds, aes(eth3, exp,fill=eth3))+
  geom_boxplot()+
  scale_x_discrete(labels=c("Hispanic","Other","White, not Hispanic"))+
  guides(fill=FALSE)+
  labs(x = "Ethnicity", y = "Mean Expenditures")+
  theme_bw()
ethbox
```



Let's also plot mean expenditures, and add error bars to reflect the 95% bootstrapped confidence interval around each estimate of the mean.

```
ethline <- ggplot(cadds, aes(eth3, exp))+
#  stat_summary(fun.y = mean, geom = "line", size = 1, aes(group=1), colour = "black") +
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar", width = 0.2, size = 0.75, colour = "#990000")
  stat_summary(fun.y = mean, geom = "point", size = 4, colour = "#990000") +
  stat_summary(fun.y = mean, geom = "point", size = 3, colour = "#FF6633") +
  labs(x = "Ethnicity", y = "Mean Expenditures")+
  theme_bw()
ethline
```



All evidence here seems to tell the same story- can you see why some evidence suggests discrimination in the form of higher expenditures for White non-Hispanics?

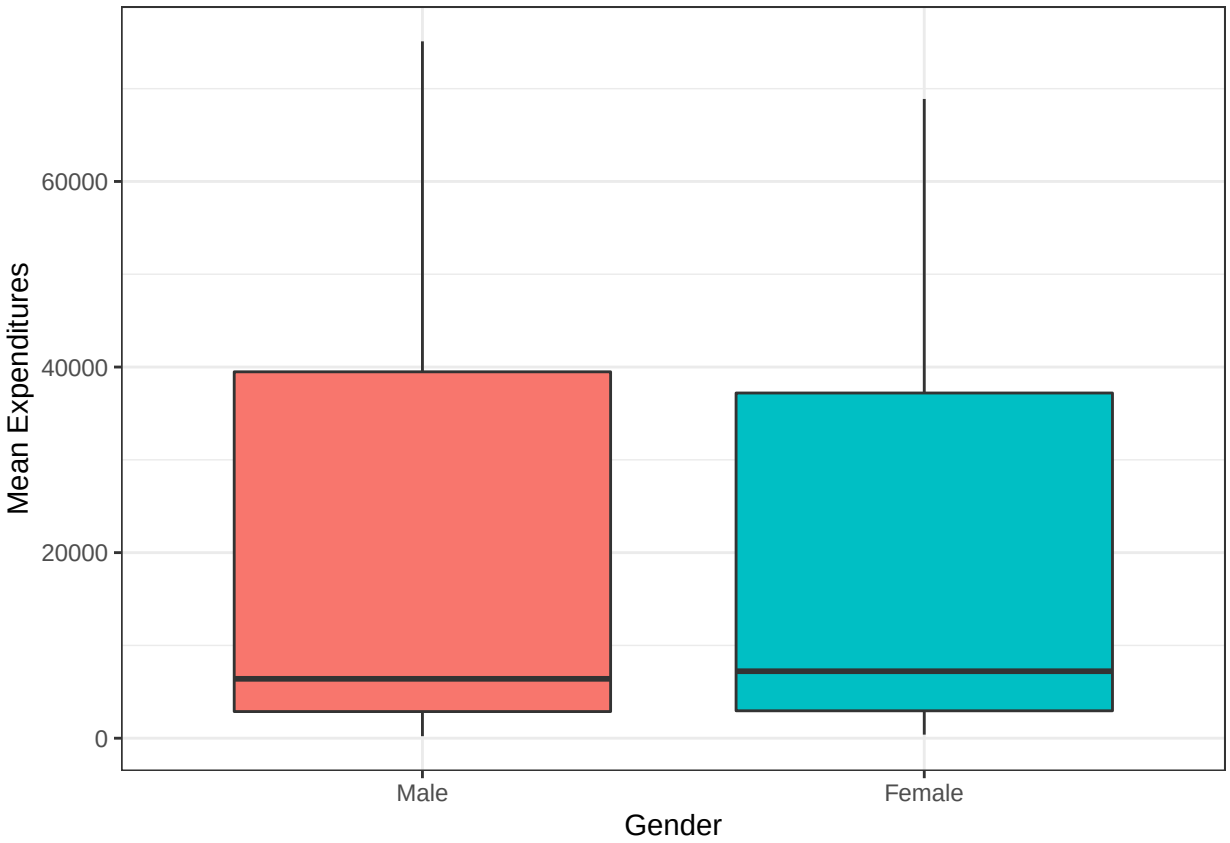
Univariate Statistics: Gender (Factor B)

Let's check out gender now.

```
cadds %>%
  group_by(sex) %>%
  summarise(mean = mean(exp, na.rm =TRUE),
            sd = sd(exp, na.rm = TRUE),
            n = n())
```

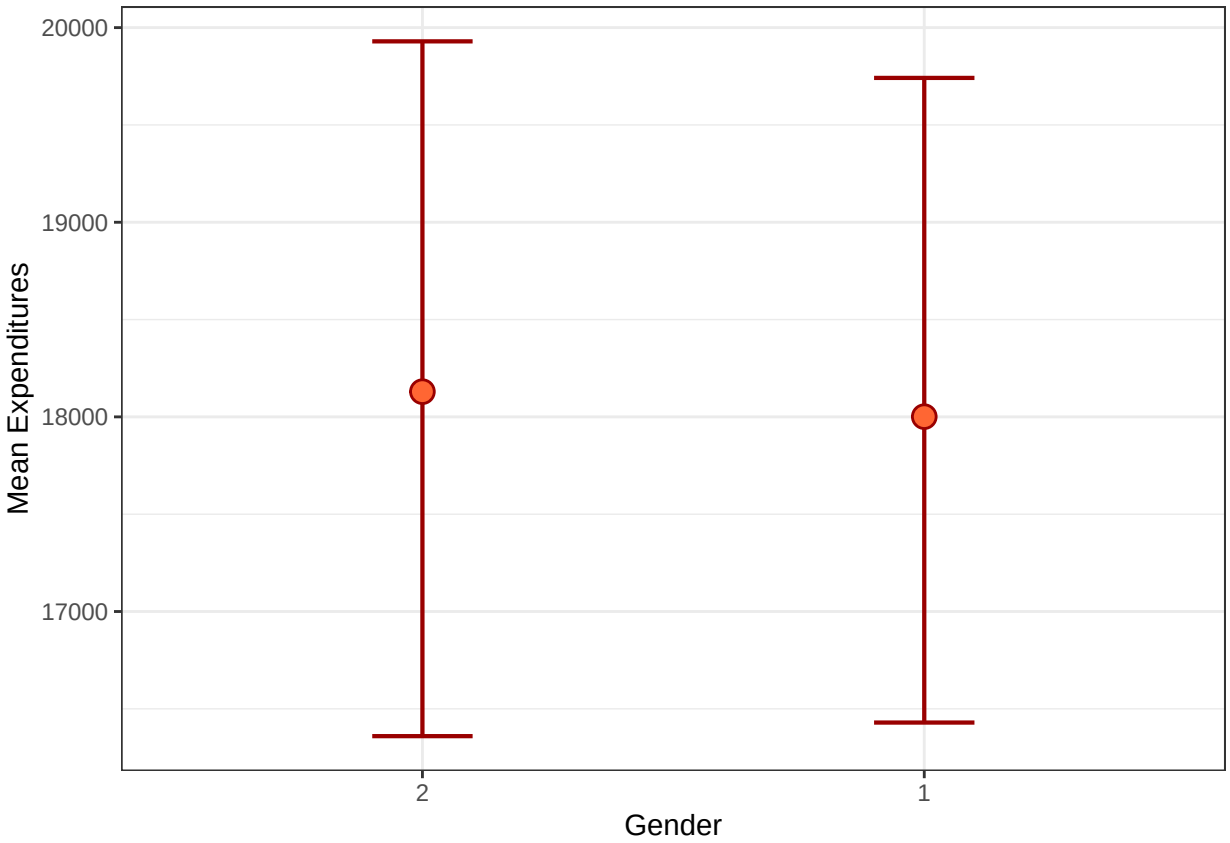
```
## # A tibble: 2 x 4
##   sex     mean     sd     n
##   <fct>  <dbl>  <dbl> <int>
## 1 2     18130. 20020.   503
## 2 1     18001. 19068.   497
```

```
genbox <- ggplot(cadds, aes(sex, exp,fill=sex))+
  geom_boxplot()+
  scale_x_discrete(labels=c("Male","Female"))+
  guides(fill=FALSE)+
  labs(x = "Gender", y = "Mean Expenditures")+
  theme_bw()
genbox
```



Let's also plot mean expenditures, and add error bars to reflect the 95% bootstrapped confidence interval around each estimate of the mean.

```
genline <- ggplot(cadds, aes(sex, exp))+
  #stat_summary(fun.y = mean, geom = "line", size = 1, aes(group=1), colour = "#FF6633") +
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar", width = 0.2, size = 0.75, colour = "#990000")
  stat_summary(fun.y = mean, geom = "point", size = 4, colour = "#990000") +
  stat_summary(fun.y = mean, geom = "point", size = 3, colour = "#FF6633")+
  labs(x = "Gender", y = "Mean Expenditures")+
  theme_bw()
genline
```



Again, all evidence here seems to tell the same story, but it is a very boring one. I'm going to guess that expenditures are similar for males and females in California.

Two-way ANOVA (Factor A, B, and their interaction A*B)

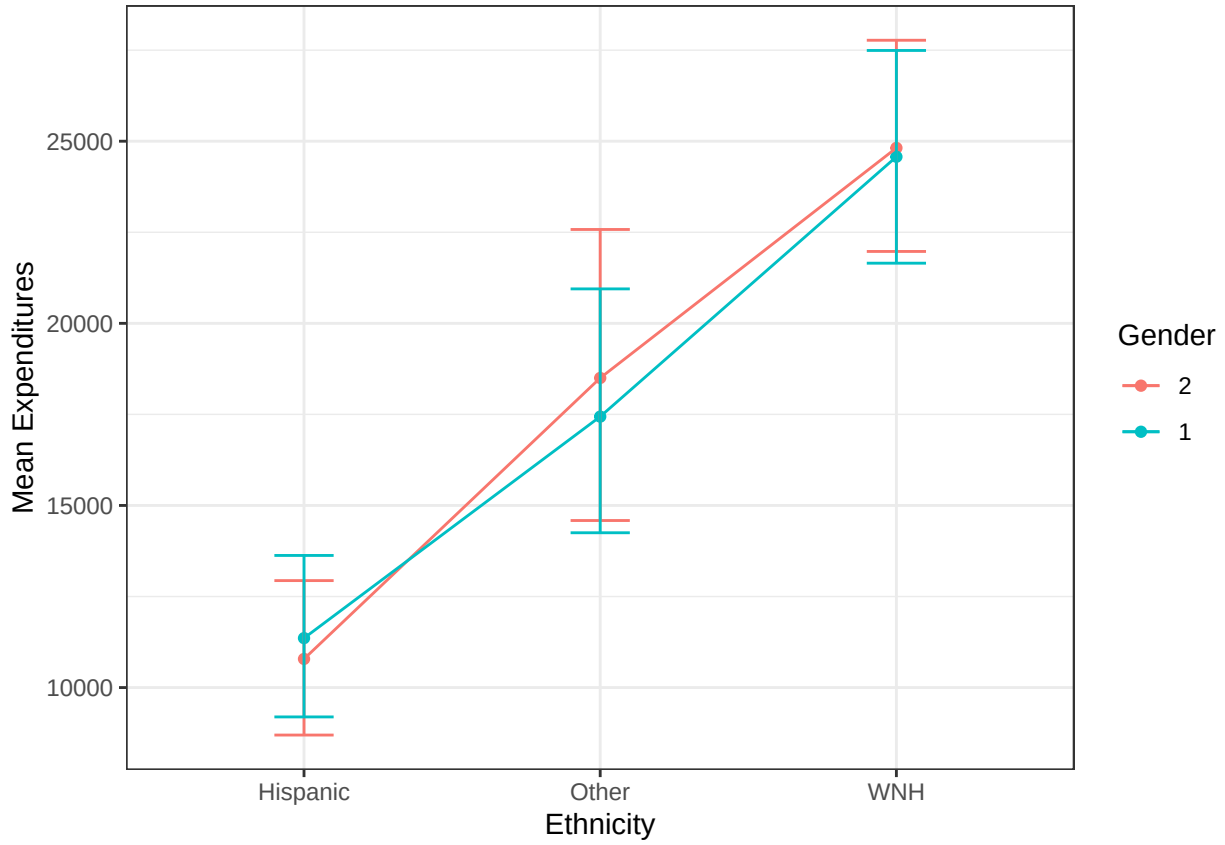
Now we will analyze both factors, ethnicity and gender, simultaneously. First, just as we did at the univariate level, we can look at the means. Now, the means in each cell are called cell means.

```
cadds %>%
  group_by(sex, eth3) %>%
  summarise(mean = mean(exp, na.rm =TRUE),
            sd = sd(exp, na.rm = TRUE),
            n = n())
```

```
## # A tibble: 6 x 5
## # Groups:   sex [2]
##   sex   eth3       mean     sd     n
##   <fct> <fct>     <dbl>  <dbl> <int>
## 1 2     Hispanic 10786. 15153.   192
## 2 2     Other    18500  20624.   106
## 3 2     WNH      24816. 21368.   205
## 4 1     Hispanic 11357. 16148.   184
## 5 1     Other    17439. 18413.   117
## 6 1     WNH      24574. 19828.   196
```

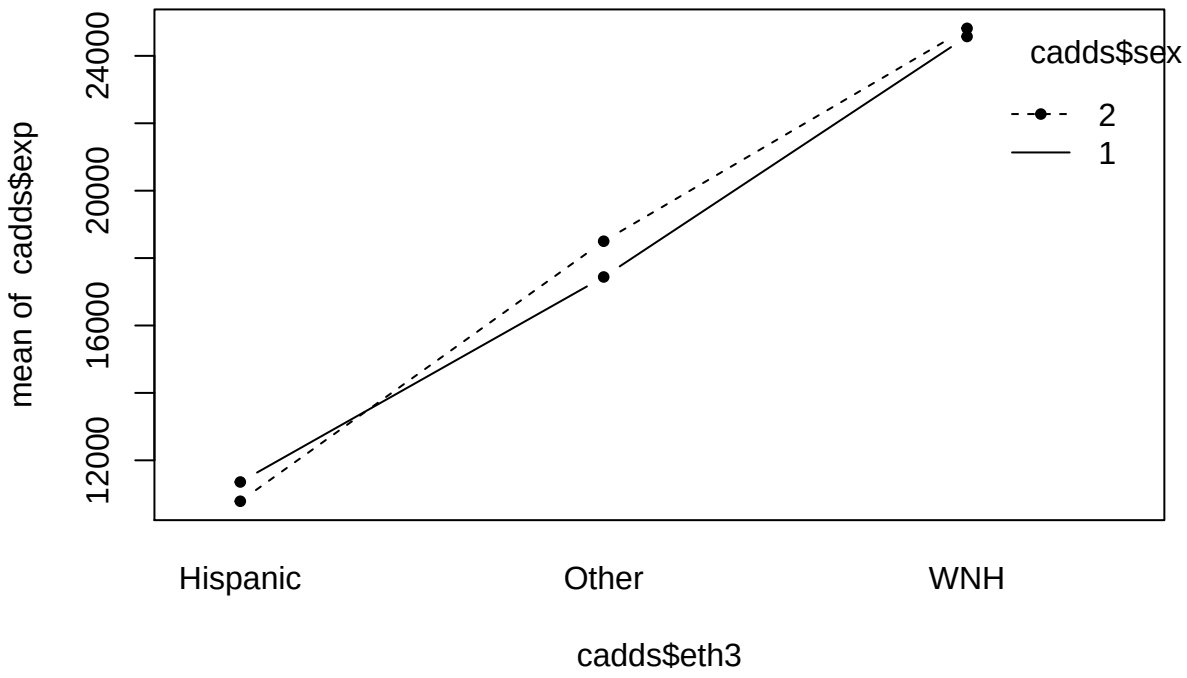
Plotting data for a 2x2 ANOVA is a little tricky, but worthwhile.

```
int <- ggplot(cadds, aes(eth3, exp, colour = sex)) +
  stat_summary(fun.y = mean, geom = "point") +
  stat_summary(fun.y = mean, geom = "line", aes(group= factor(sex))) +
  stat_summary(fun.data = mean_cl_boot, geom = "errorbar", width = 0.2) +
  labs(x = "Ethnicity", y = "Mean Expenditures", colour = "Gender") +
  theme_bw()
int
```



Another (less pretty but still functional) way:

```
interaction.plot(cadds$eth3,cadds$sex,cadds$exp,type="b",pch=20)
```



What you see here is exactly what we would expect based on the univariate statistics and plots:

1. There appears to be a main effect of ethnicity; that is, expenditures appear to be higher among White non-Hispanic individuals than among Hispanic individuals.
2. There does not appear to be a main effect of gender; that is, expenditures appear to be roughly the same for males and females.
3. There does not appear to be a significant interaction between ethnicity and gender; that is, the effect of ethnicity does not appear to depend on the level of gender and vice versa. If the interaction *were* significant, the main effect of one factor would be different for different levels of the other factor (so, if we saw that expenditures were higher among Hispanics than White non-Hispanics but only for girls). So the question to ask yourself when you look at this graph is: does including gender change how ethnicity behaves? In this case, the answer is no. An important point to remember is that if you do have a significant interaction term, this changes the way you have to interpret the main effects. A significant main effect in the context of a significant interaction is not the same as a significant main effect in the context of a non-significant interaction.

Let's run the ANOVA in R.

```
ca_aov1 <- aov(exp~sex*eth3,data=cadds) #same as: sex+eth3+sex:eth3
summary(ca_aov1)
```

```
##              Df       Sum Sq     Mean Sq F value               Pr(>F)
## sex           1      4122210     4122210    0.01                 0.91
## eth3          2  36063945832 18031972916   51.90 <0.0000000000000002 ***
## sex:eth3      2     95555178    47777589    0.14                 0.87
## Residuals   994 345376693512   347461462
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Right? **Wrong.** This is not the way to do a 2x2 ANOVA in R. Or at the very least, this is not an advisable way to run anything more complicated than a one-way ANOVA.

Types of Sums of Squares

The first issue is that the default sums of squares output in R is Type I. Turns out, there are actually four different types of sums of squares, dating back apparently to the original output available from SAS. Strangely, Type I sums of squares are sequential, meaning that the order with which you enter the IVs in your model statement in R matters. In fact, each IV is only evaluated after the previous IV as they were typed in the model. It is rare that this is the hypothesis you would like to test. Please read up on this issue here. In our example, because gender and the interaction term are both not significant we get about the same results, but it is critical to remember that R by default via *aov* provides Type I sums of squares estimates. For our purposes, because this interaction is not likely to be significant (look at my plots again), I would go with Type II sums of squares. In the absence of an interaction, I think Type II is defensible (and I have yet to see the type of SS reported).

The easiest way to conduct a Type II/III sums of squares ANOVA is to use the car package, although there is a way in base R to do this. But, before we tackle this, we need to discuss the second big issue, which is setting contrasts.

Setting Contrasts

Another issue is the default contrasts in R. You can easily check the default contrasts for any factor variable:

```
contrasts(cadds$eth3)
```

```
         Other WNH
Hispanic     0   0
Other        1   0
WNH          0   1
```

In the ANOVA framework, a contrast is a weighted sum of the group means such that:

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k$$

What the above output shows you are R's default **coding schemes** for the contrasts, but not the **contrast weights** themselves (the c_j values above). As a demonstration, append a column of one's in the first column of this matrix, then take the inverse:

```
temp <- cbind(constant=1,contrasts(cadds$eth3))
temp
```

```
         constant Other WNH
Hispanic        1     0   0
Other           1     1   0
WNH             1     0   1
```

```
solve(temp)
```

```
         Hispanic Other WNH
constant        1     0   0
Other          -1     1   0
WNH            -1     0   1
```

Now the rows of *this* matrix tells you the actual contrast weights (with the exception of β_0, which is not a contrast; note it is the only one that does not sum to zero). Thus, by default, the coefficients β_0, β_1, and β_2 in our linear model are contrasts:

$$\beta_0 = +1\mu_1 + 0\mu_2 + 0\mu_3$$

$$\beta_1 = -1\mu_1 + 1\mu_2 + 0\mu_3$$

$$\beta_2 = -1\mu_1 + 0\mu_2 + 1\mu_3$$

Note that if you solve each of these equations, our β_0 is simply the mean of group 1 ($\mu_1 = \bar{y_1}$), β_1 is the difference between μ_2 and μ_1 ($\bar{y_2} - \bar{y_1}$), and β_2 is the difference between μ_3 and μ_1 ($\bar{y_3} - \bar{y_1}$). You might at this point wish to ask yourself- are these the effects you want to test?

In a typical ANOVA, we are likely to be more interested in the following contrasts:

$$\beta_1 = \frac{2}{k}\mu_1 - \frac{1}{k}\mu_2 - \frac{1}{k}\mu_3 = \mu_1 - (\frac{1}{k}\mu_1 + \frac{1}{k}\mu_2 + \frac{1}{k}\mu_3)$$

$$\beta_2 = \frac{2}{k}\mu_2 - \frac{1}{k}\mu_1 - \frac{1}{k}\mu_3 = \mu_2 - (\frac{1}{k}\mu_1 + \frac{1}{k}\mu_2 + \frac{1}{k}\mu_3)$$

```
options(contrasts=c("contr.sum","contr.poly")) #this MUST be specified before aov
contrasts(cadds$eth3)
```

```
         [,1] [,2]
Hispanic    1    0
Other       0    1
WNH        -1   -1
```

```
tempnew <- cbind(constant=1,contrasts(cadds$eth3))
solve(tempnew)
```

```
         Hispanic Other   WNH
constant     0.33  0.33  0.33
             0.67 -0.33 -0.33
            -0.33  0.67 -0.33
```

Here are our new c_j values:

$$\beta_0 = .33\mu_1 + .33\mu_2 + .33\mu_3$$

$$\beta_1 = .67\mu_1 - .33\mu_2 - .33\mu_3 = \mu_1 - (.33\mu_1 + .33\mu_2 + .33\mu_3)$$

$$\beta_2 = -.33\mu_1 + .67\mu_2 - .33\mu_3 = \mu_2 - (.33\mu_1 + .33\mu_2 + .33\mu_3)$$

Note that **now** if you solve each of these equations, our β_0 is the grand mean ($\mu_{\bullet\bullet} = \bar{y_{\bullet\bullet}}$), β_1 is the difference between μ_1 and the grand mean, and β_2 is the difference between μ_2 and the grand mean. Now of course, the ANOVA output does not provide estimates for β, but as we have seen, ANOVA is based off of regression - it is just a different way of summarizing the analyses in sums of squares form. Long story short, it is important to set the contrasts using the options I specify below, because the default contrast coding scheme in R rarely reflects your null/alternative hypotheses.

```
library(car)
options(contrasts=c("contr.sum","contr.poly")) #this MUST be specified before aov
ca_aov2 <- aov(exp~sex*eth3,data=cadds)
Anova(ca_aov2,type=c("II"))
```

```
Anova Table (Type II tests)

Response: exp
               Sum Sq  Df F value              Pr(>F)
```

```
sex            3504854   1    0.01                0.92
eth3       36063945832   2   51.90 <0.0000000000000002 ***
sex:eth3      95555178   2    0.14                0.87
Residuals 345376693512 994
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And now we can finally interpret our findings. First, there is a significant F-ratio for ethnicity, indicating that expenditures were significantly different based on the individuals' ethnicity. This means that overall, if we ignore gender, ethnicity influenced expenditures. Second, there is not a significant F-ratio for gender, which tells us that if we ignore ethnicity, the gender of the individual did not influence expenditures. So, ethnicity being "equal", being male or female did not affect expenditures. Finally, the F-ratio for the interaction is also not significant, which tells us that our interpretations of the main effects are valid, and neither depend on the other factor in the model.

N.B. This is a nice command in R for easily obtaining all cell means in a table format (note that "rep"=n for that cell):

```
model.tables(ca_aov2,"means", SE=T) #unbalanced, so SEs are a problem
```

Post-hoc Contrasts

Now, we know there is a main effect of ethnicity, so we need to conduct post-hoc analyses to see where these differences lie. Let's try a few multiple comparison procedures: Benjamini-Hochberg, Bonferroni, and Tukey Honestly Significant Difference.

```
pairwise.t.test(cadds$exp,cadds$eth3, p.adjust.method="BH", pool.sd=T)
```

```
    Pairwise comparisons using t tests with pooled SD

data:  cadds$exp and cadds$eth3

      Hispanic             Other
Other 0.00002              -
WNH   < 0.0000000000000002 0.00002

P value adjustment method: BH
```

```
pairwise.t.test(cadds$exp,cadds$eth3, p.adjust.method="bonferroni", pool.sd=T)
```

```
    Pairwise comparisons using t tests with pooled SD

data:  cadds$exp and cadds$eth3

      Hispanic             Other
Other 0.00004              -
WNH   < 0.0000000000000002 0.00005

P value adjustment method: bonferroni
```

```
library(multcomp)
summary(glht(ca_aov2, linfct = mcp(eth3 = "Tukey")))
```

```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = exp ~ sex * eth3, data = cadds)

Linear Hypotheses:
                       Estimate Std. Error t value Pr(>|t|)
Other - Hispanic == 0     6898       1577    4.37  <0.0001 ***
WNH - Hispanic == 0      13623       1338   10.18  <0.0001 ***
WNH - Other == 0          6725       1558    4.32  <0.0001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Note that you can also change "eth3" to "sex" to run the Tukey HSD on the other factor.

```
TukeyHSD(ca_aov2, "eth3", ordered = FALSE) #another option
plot(TukeyHSD(ca_aov2, "eth3"))
```

All of these converge to the same conclusion: expenditures among White non-Hispanics are greater than all other groups, and expenditures among Hispanics are the lowest.

Unequal Variances

With unequal variances with just one IV factor, we could run Welch's F test (oneway.test), which is the alternative to Welch's t test (t.test with var.equal=F) when we have a factor with more than two levels. Unfortunately, there are not many options for unequal variances when we have more than one IV. The *car* package does offer this "white.adjust" option, which if set to true will use a heteroscedasticity-corrected coefficient covariance matrix (see White, 1980, "A heteroskedastic consistent covariance matrix estimator and a direct rest of heteroskedasticity"). For now, I will introduce this as an option for the omnibus ANOVA, although there may be better non-NHST (e.g., Bayesian or mixed models) options for such a scenario.

```
Anova(ca_aov2,type=c("II"), white.adjust=T)
```

Follow-up contrasts between groups could be done using pairwise Welch's t-tests. This option cannot be used when you have paired samples (i.e., pool.sd must be false if paired=TRUE).

```
pairwise.t.test(cadds$exp,cadds$eth3, p.adjust.method="BH", pool.sd=F)
pairwise.t.test(cadds$exp,cadds$sex, p.adjust.method="BH", pool.sd=F)
```

On your own

You will conduct an ANOVA to examine whether ethnicity and age affect expenditures. This ANOVA will be a 2 (2 levels of ethnicity: White non-Hispanic and Hispanic) x 5 (5 levels of age group: under 6, 6-12, 13-17, 18-21, and over 22). So, first, drop the "other" ethnicity category, and create an age group variable as follows:

```
ca2 <- subset(cadds, !eth3=="Other")
ca2$eth3 <- droplevels(ca2$eth3)
library(Hmisc)
ca2$agegroup <- cut2(ca2$age, c(6,13,18,22))
table(ca2$agegroup,ca2$eth3)
```

```
        Hispanic WNH
[ 0, 6)       44  20
[ 6,13)       91  46
[13,18)      103  67
[18,22)       78  69
[22,95]       60 199
```

1. Plot age group on the x-axis and expenditures on the y-axis using boxplots. Needs for disabled individuals typically increase with age, resulting in higher expenditures at older ages. Is this what you observe?

2. Now, add ethnicity as the "fill" factor, making a separate boxplot for each ethnic group a different color. Do you see anything to confirm what we previously found was the main effect of ethnicity? Remember, we found that *at any level of the factor "gender"*, expenditures were higher among White non-Hispanic versus Hispanic individuals. Do you see the same pattern here with respect to age group rather than gender? Based on our previous analyses, what did you expect to see, and what surprises you?

3. Examine the means, sds, and n's as I did in the previous example. Look very carefully at each. Do you see any patterns, in particular in the sample sizes? Knowing that Hispanic children tend to be diagnosed later than White non-Hispanic children (and thus should be under-represented at younger ages), do these numbers surprise you?

4. Create the interaction plot (x-axis=ethnicity; y-axis=expenditures; color=age group) and interpret. Do you see evidence for systematic bias in expenditures in favor of White non-Hispanics at any age group? That is, is the typical Hispanic receiving fewer funds (expenditures) than the typical White non-Hispanic?

5. Conduct an ANOVA and interpret. What do you conclude? How do these effects differ from our previous analyses of ethnicity and gender? Do your previous conclusions change based on your analyses? Follow up with any post-hoc tests you think are necessary to validate your conclusions.

6. Why is the overall average for all individuals significantly different indicating ethnic discrimination of Hispanics? To answer this question, you should consider the *linear structural model parameterization* for the two-way ANOVA:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$$

where there are a levels of factor A, b levels of factor B, and μ_{jk} represents cell means for any combination of level A_j with B_k.

$$\mu = \sum_j \sum_k \frac{\mu_{jk}}{ab}$$

$$\mu_{j\bullet} = \sum_k \frac{\mu_{jk}}{b}$$

$$\mu_{\bullet k} = \sum_j \frac{\mu_{jk}}{a}$$

$$\alpha_j = \mu_{j\bullet} - \mu$$
$$\beta_k = \mu_{\bullet k} - \mu$$
$$(\alpha\beta)_{jk} = \mu_{jk} - (\mu - \alpha_j - \beta_k)$$

Specifically, consider ethnicity to be factor A (j=1 [Hispanic] to 2 [White not Hispanic]), and age group to be factor B (k=1 to 5). So a=2 and b=5.

$$\mu = \sum_j \sum_k \frac{\mu_{jk}}{10}$$
$$\mu_{j\bullet} = \sum_k \frac{\mu_{jk}}{5}$$
$$\mu_{\bullet k} = \sum_j \frac{\mu_{jk}}{2}$$
$$\alpha_j = \mu_{j\bullet} - \mu$$
$$\beta_k = \mu_{\bullet k} - \mu$$
$$(\alpha\beta)_{jk} = \mu_{jk} - (\mu - \alpha_j - \beta_k)$$

What is $\mu_{1\bullet} = \sum_k \mu_{1k}/5$? The numerator is a pooled mean estimate, that given unequal group sizes, is weighted by the size of each k group: $\sum_k \mu_{1k} = \mu_{11} + \mu_{12} + \ldots + \mu_{1b}$, where for example $\mu_{11} = \frac{\sum y_{i1}}{n_{11}}$.

> Note that to answer this question adequately you do not *need* any math, proofs, formulas, etc. necessarily- a "right" answer can just be a few logical sentences.