ANOVA

Put your name here

Put the date here

Introduction

ANOVA stands for "Analysis of Variance". In this module, we will study the most basic form of ANOVA, called "one-way ANOVA". We've already considered the one-sample and two-sample t-tests for means. ANOVA is what you do when you want to compare means for three or more groups.

Instructions

Presumably, you have already created a new project and downloaded this file into it. Please knit the document and work back and forth between this R Markdown file and the PDF output as you work through this module.

When you are finished with the assignment, knit to PDF one last time, proofread the PDF file **carefully**, export the PDF file to your computer, and then submit your assignment.

Sometimes you will be asked to add your own R code. That will appear in this document as a code chunk with a request for you to add your own code, like so:

```
## Add code here to [do some task]...
```

Be sure to remove the line `## Add code here to [do some task]...` when you have added your own code.

Sometimes you will be asked to type up your thoughts. That will appear in the document as follows:

Please write up your answer here.

Again, please be sure to remove the line "Please write up your answer here" when you have written up your answer. In these areas of the assignment, please use contextually meaningful full sentences/paragraphs (unless otherwise indicated) and proper spelling, grammar, punctuation, etc. This is not R code, but rather a free response section where you talk about your analysis and conclusions. If you need to use some R code as well, you can use inline R code inside the block between `\begin{answer}` and `\end{answer}`, or if you need an R code chunk, please go outside the `answer` block and start a new code chunk.

Load Packages

We load the standard `mosaic` package and the `quantreg` package for the `uis` data. The `broom` package gives us tidy output.

```r
library(quantreg)
data(uis)
library(broom)
library(mosaic)
```

Research question

The `uis` dataset from the `quantreg` package contains data from the UIS Drug Treatment Study. Is a history of IV drug use associated with depression?[1]

Data preparation and exploration

To talk about the ANOVA procedure, we'll use the `IV` and `BECK` variables from the `uis` data set. We need to convert `IV` to a factor variable first. Then we'll put both variables into a new data frame for convenience.

```
IV <- factor(uis$IV, levels = c(1, 2, 3),
             labels = c("Never", "Previous", "Recent"))
IV_BECK <- data.frame(IV, BECK = uis$BECK)
```

Let's look at the three groups in our data defined by the `IV` variable. These are people who have never used IV drugs, those who have previously used IV drugs, and those who have recently used IV drugs. The following table shows how many people are in each group.

```
table(IV_BECK$IV)
```

```
##
##    Never Previous   Recent
##      223      109      243
```

We're interested in depression as measured by the Beck Depression Inventory.

Exercise

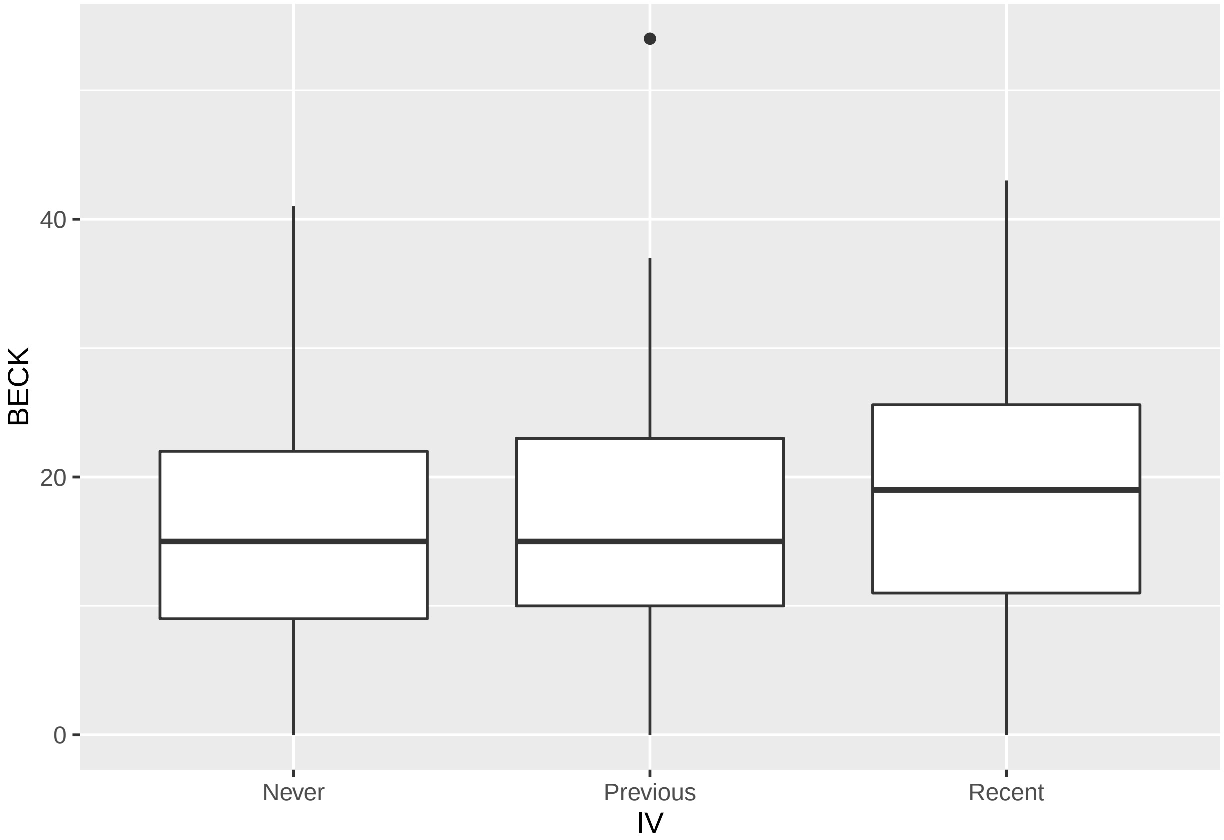
Google the Beck Depression Inventory. Write a short paragraph about it and how it purports to measure depression.

Please write up your answer here.

A useful graph is a side-by-side boxplot.

```
ggplot(IV_BECK, aes(x = IV, y = BECK)) +
    geom_boxplot()
```

[1] More information about the UIS study here: https://www.umass.edu/statdata/statdata/data/uissurv.txt



This boxplot shows that the distribution of depression scores in each group is similar. There are some small differences, but it's not clear if these differences are statistically significant.

The mean Beck score is calculated with the `mean` command. If we don't use the tilde, we'll just get one overall mean for the whole sample, often called the "grand mean":

```
mean(IV_BECK$BECK)
```

```
## [1] 17.36743
```

But if we use the tilde, we can separate this out by `IV` group:

```
mean(BECK ~ IV, data = IV_BECK)
```

```
##    Never Previous   Recent
## 15.94996 16.64201 18.99363
```

The F distribution

When assessing the difference between groups, there are two numbers that are important.

The first is called the "mean square between groups" (MSG). It measures how far away each group mean is away from the overall grand mean for the whole sample. For example, for those who never used IV drugs, their mean Beck score was 15.95. This is 1.42 points below the grand mean of 17.37. On the other hand, recent IV drug users had a mean Beck score of nearly 19. This is 1.63 points above the grand mean. MSG is calculated by taking these differences for each group, squaring them to make them positive, weighting them by the sizes of each group (larger groups should obviously count for more), and dividing by the "group degrees of freedom" $df_G = k - 1$ where k is the number of groups. The idea is that MSG is a kind of "average variability" among the groups. In other words, how far away are the groups from the grand mean (and therefore, from each other)?

The second number of interest is the "mean square error" (MSE). It is a measure of variability within groups. In other words, it measures how far away data points are from their own group means. Even under the assumption of a null hypothesis that says all the groups should be the same, we still expect some variability. Its calculation also involves dividing by some degrees of freedom, but now it is $df_E = n - k$.

All that is somewhat technical and complicated. We'll leave it to the computer. The key insight comes from considering the ratio of MSG and MSE. We will call this quantity F:

$$F = \frac{MSG}{MSE}.$$

What can be said about this magical F? Under the assumption of the null hypothesis, we expect some variability among the groups, and we expect some variability within each group as well, but these two sources of variability should be about the same. In other words, MSG should be roughly equal to MSE. Therefore, F ought to be close to 1.

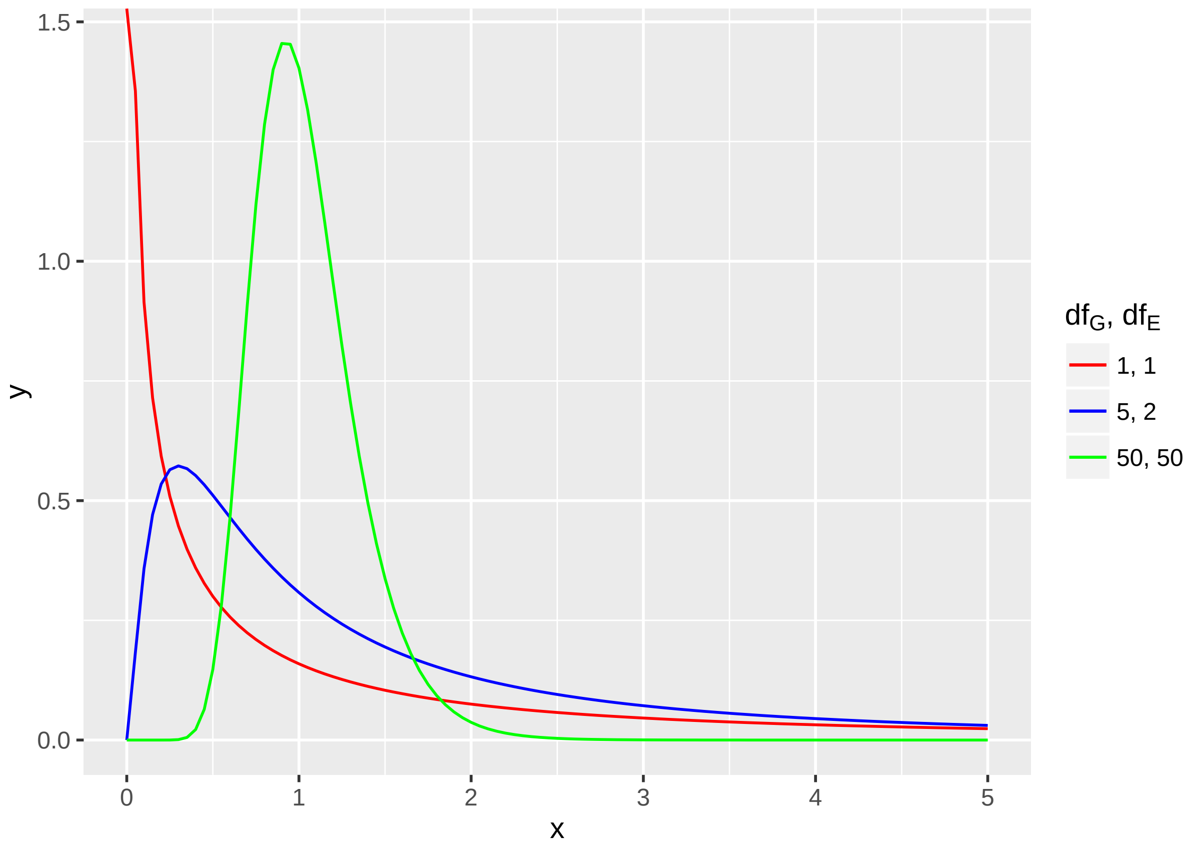
It's not particular interesting if F is less than one. That just means that the variability between groups is small and the variability of the data within each group is large. That doesn't allow us to conclude that there is a difference between groups. However, if F is really large, that means that there is much more variability between the groups than there is within each group. Therefore, the groups are far apart and there is evidence of a difference between groups.

MSG and MSE are measures of variability, and that's why this is called "Analysis of Variance".

The F distribution is the correct sampling distribution model. Like a t model, there are infinitely many different F models because degrees of freedom are involved. But unlike a t model, the F model has *two* degrees of freedom, df_G and df_E. Both of these numbers affect the precise shape of the F distribution.

For example, here is picture of a few different F models. (Don't worry about the syntax here. You won't need to do this on your own.)

```r
ggplot(data.frame(x = c(0, 5)), aes(x)) +
    stat_function(fun = df, args = list(df1 = 1, df2 = 1),
                  aes(color = "1, 1")) +
    stat_function(fun = df, args = list(df1 = 5, df2 = 2),
                  aes(color = "5, 2" )) +
    stat_function(fun = df, args = list(df1 = 50, df2 = 50),
                  aes(color = "50, 50")) +
    scale_color_manual(name = expression(paste(df[G], ", ", df[E])),
                       values = c("1, 1" = "red",
                                  "5, 2" = "blue",
                                  "50, 50" = "green"),
                       breaks =  c("1, 1", "5, 2", "50, 50"))
```



Assumptions

What conditions can we check to justify the use of an F model for our sampling distribution? In addition to the typical "Random" and "10%" conditions that ensure independence, we also need to check the "Nearly normal" condition for each group, just like for the t tests. A new assumption is the "Constant Variance" assumption, which says that each group should have the same variance in the population. This is impossible to check, although we can use our sample as a rough guide. If each group has about the same spread, that is some evidence that such an assumption might hold in the population as well. Also, ANOVA is pretty robust to this assumption, especially when the groups are close to the same size. Even when the group sizes are unequal (sometimes called "unbalanced"), some say the variances can be off by up to a factor of 3 and ANOVA will still work pretty well. So what we're looking for here are gross violations, not minor ones.

Let's go through the rubric with commentary.

Exploratory data analysis

Use data documentaton (help files, code books, Google, etc.), the str command, and other summary functions to understand the data.

[Type `library(quantreg)` then `?uis` at the Console to read the help file. You have already Googled the Beck Depression Score.]

```
str(uis)
```

```
## 'data.frame':    575 obs. of  18 variables:
##  $ ID    : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ AGE   : num  39 33 33 32 24 30 39 27 40 36 ...
##  $ BECK  : num  9 34 10 20 5 ...
##  $ HC    : num  4 4 2 4 2 3 4 4 2 2 ...
##  $ IV    : num  3 2 3 3 1 3 3 3 3 3 ...
##  $ NDT   : num  1 8 3 1 5 1 34 2 3 7 ...
##  $ RACE  : num  0 0 0 0 1 0 0 0 0 0 ...
##  $ TREAT : num  1 1 1 0 1 1 1 1 1 1 ...
##  $ SITE  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LEN.T : num  123 25 7 66 173 16 179 21 176 124 ...
##  $ TIME  : num  188 26 207 144 551 32 459 22 210 184 ...
##  $ CENSOR: num  1 1 1 1 0 1 1 1 1 1 ...
##  $ Y     : num  5.24 3.26 5.33 4.97 6.31 ...
##  $ ND1   : num  5 1.11 2.5 5 1.67 ...
##  $ ND2   : num  -8.047 -0.117 -2.291 -8.047 -0.851 ...
##  $ LNDT  : num  0.693 2.197 1.386 0.693 1.792 ...
##  $ FRAC  : num  0.6833 0.1389 0.0389 0.7333 0.9611 ...
##  $ IV3   : num  1 0 1 1 0 1 1 1 1 1 ...
```

Prepare the data for analysis.

We need `IV` to be a factor variable. We put `IV` and `BECK` into a separate data frame.

```r
IV <- factor(uis$IV, levels = c(1, 2, 3),
             labels = c("Never", "Previous", "Recent"))
IV_BECK <- data.frame(IV, BECK = uis$BECK)
```

Commentary: we already did this above, but we include it here for completeness.

Make tables or plots to explore the data visually.

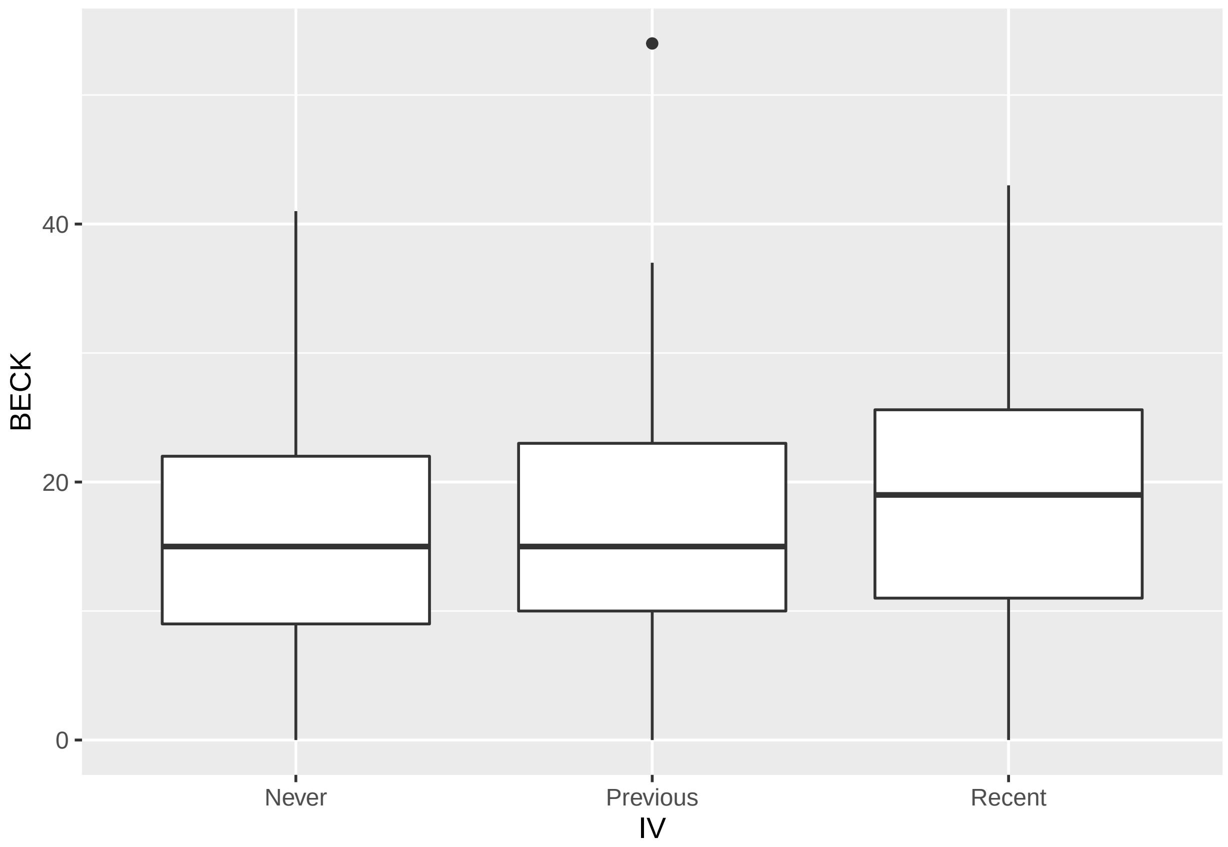
The following table shows how many people are in each group.

```r
table(IV_BECK$IV)
```

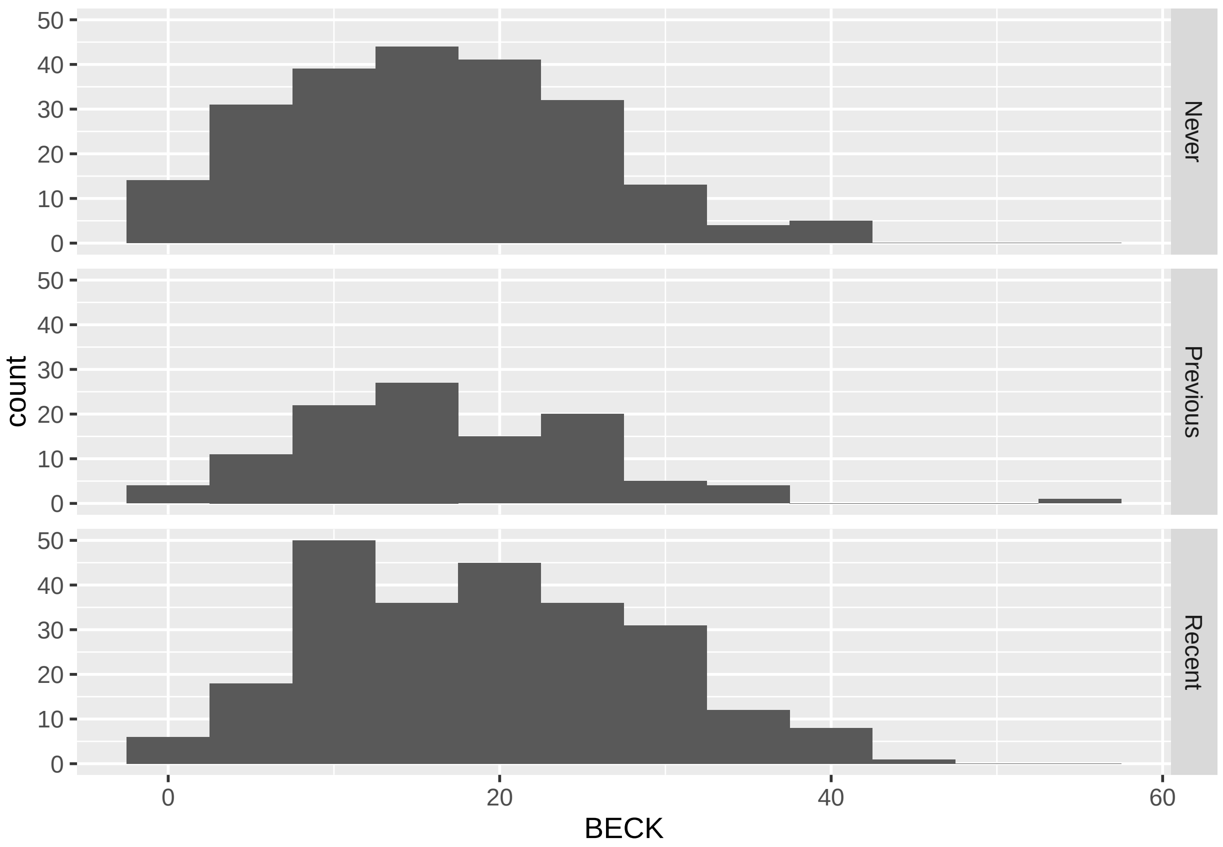
```
## 
##    Never Previous   Recent 
##      223      109      243
```

Here are two graphs that are appropriate for one categorical and one numerical variable: a side-by-side boxplot and a stacked histogram.

```r
ggplot(IV_BECK, aes(x = IV, y = BECK)) +
    geom_boxplot()
```



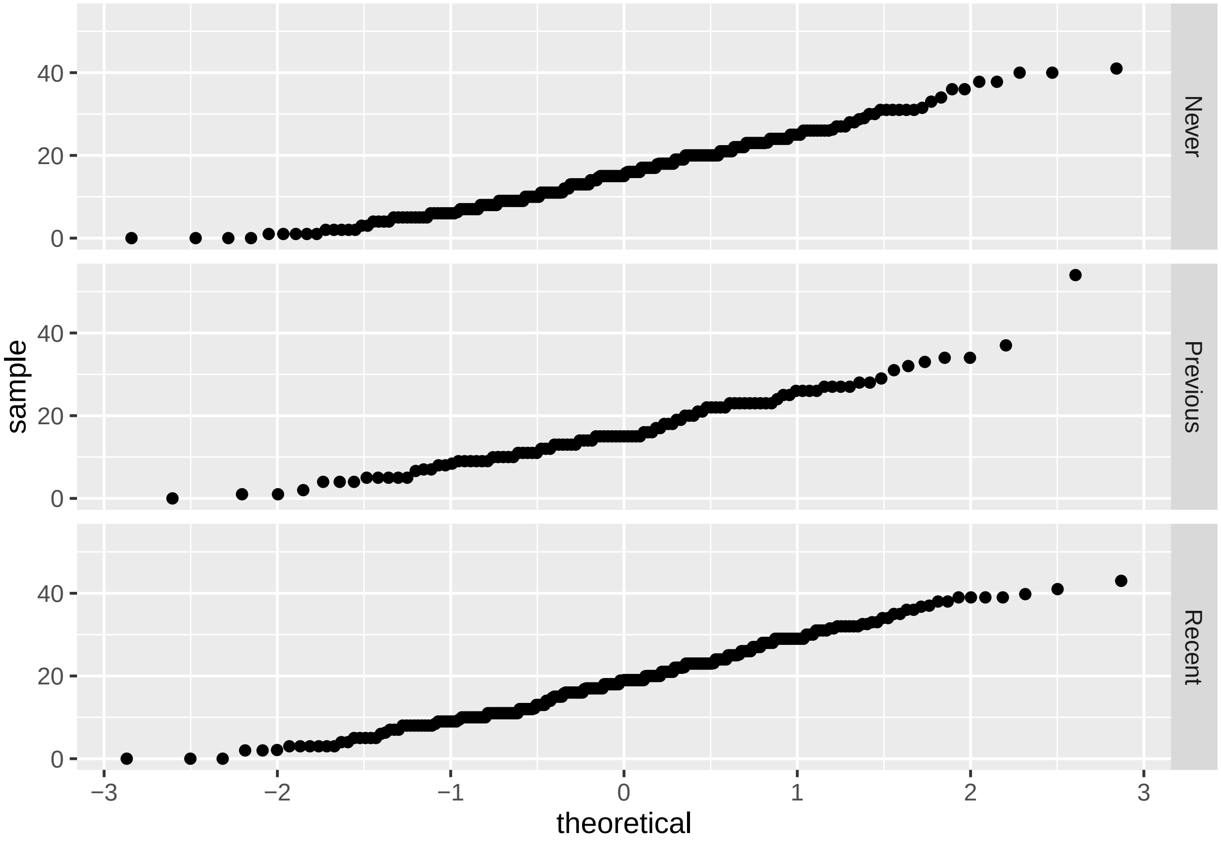
```
ggplot(IV_BECK,  aes(x = BECK)) +
    geom_histogram(binwidth = 5) +
    facet_grid(IV ~ .)
```



Both graphs show that the distribution of depression scores in each group is similar.

The distributions look reasonably normal, but we can also check the QQ plots:

```r
ggplot(IV_BECK, aes(sample = BECK)) +
    geom_qq() +
    facet_grid(IV ~ .)
```



There is one outlier in the "Previous" group, but with sample sizes as large as we have in each group, it's unlikely that this outlier will be influential. So we'll just leave it in the data and not worry about it.

Here are the group means:

```
mean(BECK ~ IV, data = IV_BECK)
```

```
##    Never Previous   Recent
## 15.94996 16.64201 18.99363
```

Commentary: We also did most (but not all) of this above. Again, we include everything in the rubric so you'll have a complete and thorough example.

Hypotheses

Identify the sample (or samples) and a reasonable population (or populations) of interest.

The sample consists of people who participated in the UIS drug treatment study. Because the UIS studied the effects of residential treatment for drug abuse, the population is, presumably, all drug addicts.

Express the null and alternative hypotheses as contextually meaningful full sentences.

H_0: There is no difference in depression levels between those who have no history of IV drug use, those who have some previous IV drug use, and those who have recent IV drug use.

H_A: There is a difference in depression levels between those who have no history of IV drug use, those who have some previous IV drug use, and those who have recent IV drug use.

Express the null and alternative hypotheses in symbols.

$H_0 : \mu_{never} = \mu_{previous} = \mu_{recent}$

There is no easy way to express the alternate hypothesis in symbols because any deviation in any of the categories can lead to rejection of the null. You can't just say $\mu_{never} \neq \mu_{previous} \neq \mu_{recent}$ because two of these categories might be the same and the third different and that would still be consistent with the alternative hypothesis.

So the only requirement here is to express the null in symbols.

Model

Identify the sampling distribution model.

We will use an F model with $df_G = 2$ and $df_E = 572$.

Commentary: Remember that

$$df_G = k - 1 = 3 - 1 = 2,$$

(k is the number of groups, in this case, 3), and

$$df_E = n - k = 575 - 3 = 572.$$

Check the relevant conditions to ensure that model assumptions are met.

- Random
  - We have no information about how this sample was collected, so we have to hope it's representative.
- 10%
  - 575 is definitely less than 10% of all drug addicts.
- Nearly normal
  - The earlier stacked histograms and QQ plots showed that each group is nearly normal. (There was one outlier in one group, but our sample sizes are quite large.)
- Constant variance
  - The spread of data looks pretty consistent from group to group in the stacked histogram and side-by-side boxplot.

Mechanics

Compute the test statistic.

```
BECK_test <- tidy(aov(BECK ~ IV, data = IV_BECK))
BECK_test
```

```
##        term  df     sumsq    meansq statistic     p.value
## 1        IV   2  1148.039 574.01934  6.721405 0.001302279
## 2 Residuals 572 48849.766  85.40169        NA          NA
```

Commentary: ANOVA is run with the aov command and the tilde notation. (BECK ~ IV means, "Calculate the means of BECK grouped by IV.")

This is the first example we've seen where the tidy output of a hypothesis test has been more than one row. The table of values here is called an ANOVA table. Although it looks foreign, you actually know what almost all of these numbers mean, although by a different name.

The first row contains everything related to the groups (as indicated by the IV term). The 2 degrees of freedom listed here are what we were calling df_G. The next number, sumsq, or the "sum of squares", is just adding up all the squared differences between the group means and the grand mean, weighted by the size of each group. You can easily see that this sum (1148.039) divided by the degrees of freedom (2) gives you meansq (574.01934) which we were calling MSG.

The second row says "residuals", which is just a technical term for the distances between the data points and their group means. We have seen the number 572 before; this is df_E. The sum of squares is now measuring the squared distance from each data value to its group mean, the same idea as above but for data points instead of whole groups. Again, check for yourself that sumsq (48849.766) divided by the degrees of freedom (572) gives you meansq (85.40169), which is just MSE.

Moving to the right in the table, what is the statistic? This is the value of F. Recall that

$$F = \frac{MSG}{MSE}.$$

Check for yourself: $574.01934/85.40169 = 6.721405$.

If we need to isolate this number, it no longer works simply to type BECK_test$statistic. Try it to see why:

```
BECK_test$statistic
```

```
## [1] 6.721405       NA
```

There are two entries here because the tidy output has two rows, but there is only one F score. (NA is the R code for a missing value.) We can grab the first entry like this:

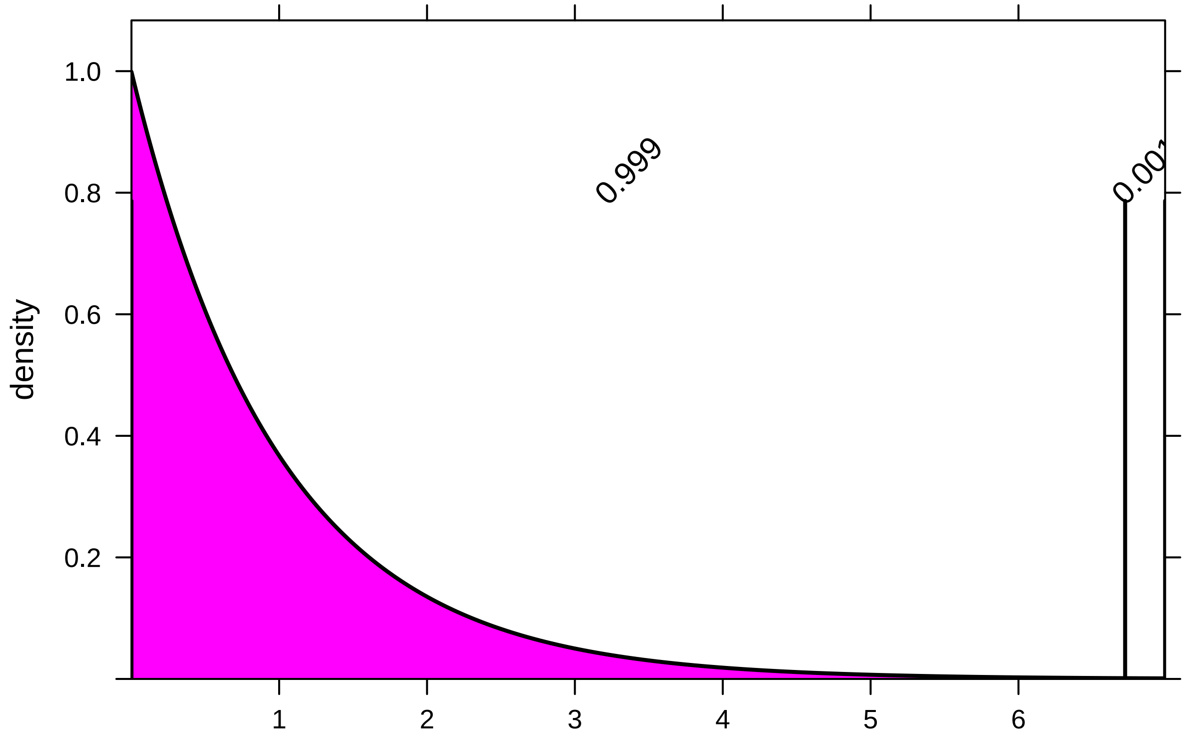
```
BECK_test$statistic[1]
```

```
## [1] 6.721405
```

The F score is 6.721405.

Plot the null distribution.

```
pdist("f", df1 = BECK_test$df[1], df2 = BECK_test$df[2],
      q = BECK_test$statistic[1])
```



```
## [1] 0.9986977
```

Commentary: As the F distribution has two parameters corresponding to df_G and df_E, we have to feed both of them into the `pdist` command. As before, the tidy output has two numbers in the *df* column, so we have to use `[1]` and `[2]` to grab them.

Calculate the P-value.

```
BECK_test$p.value[1]
```

```
## [1] 0.001302279
```

Commentary: Same issue here: the P-value is located in the first entry of `BECK_test$p.value`.

Note that this is, by definition, a one-sided test. Extreme values of F are the ones that are far away from 1, and only those values in the right tail are far from 1.

Conclusion

State the statistical conclusion.

We reject the null hypothesis.

State (but do not overstate) a contextually meaningful conclusion.

There is sufficient evidence that there is a difference in depression levels between those who have no history of IV drug use, those who have some previous IV drug use, and those who have recent IV drug use.

Identify the possibility of either a Type I or Type II error and state what making such an error means in the context of the hypotheses.

If we've made a Type I error, that means that there really isn't a difference between the three groups, but our sample is an unusual one that did detect a difference.

Exercise

Everything we saw earlier in the exploratory data analysis pointed toward failing to reject the null. All three groups look very similar in all the plots, and the means are not all that far from each other. So why did we get such a tiny P-value and reject the null? In other words, what is it about our data that allows for small effects to be statistically significant?

Please write up your answer here.

If you were a psychologist working with drug addicts, would the statistical conclusion (rejecting the null and concluding that there was a difference between groups) be of clinical importance to you? In other words, if there is a difference, is it of practical significance and not just statistical significance?

Please write up your answer here.

There is not really a confidence interval for ANOVA. We are not hypothesizing about the value of any particular parameter, so there's nothing to estimate with a confidence interval. However, there is something of interest to estimate. That leads us to the...

Post hoc analysis

If it has been determined that there is a difference between groups, it's natural to ask which group or groups are most different from the others. For example, we could have ten groups, nine that are identical, and one that is radically different. Rejecting the null means we have evidence of a difference overall, but it doesn't specify where the difference lies. Wouldn't it be nice to know which group or groups were different enough to trigger a statistically significant result?

One easy way to do this is to compute Tukey Honest Significant Differences[2] with the `TukeyHSD` command.

The `TukeyHSD` command doesn't like the tidy output; it wants the raw output from the `aov` command. So let's give Tukey what he wants. We don't want to make Tukey angry.

[2] John Tukey was a famous statistician.

```
BECK_test_untidy <- aov(BECK ~ IV, data = IV_BECK)
TukeyHSD(BECK_test_untidy)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = BECK ~ IV, data = IV_BECK)
##
## $IV
##                      diff        lwr      upr     p adj
## Previous-Never  0.692054 -1.8458349 3.229943 0.7976511
## Recent-Never    3.043674  1.0299195 5.057429 0.0012039
## Recent-Previous 2.351620 -0.1517446 4.854986 0.0707718
```

This command computes “pairwise differences” which means that it compares every possible pair of groups to each other. For example, the first row indicates that the difference in means between the “Previous” group and the “Never” group is 0.692054. That number is somewhat meaningless by itself, though. Following it is a confidence interval for the “true” difference in means between the “Previous” and “Never” groups. Since this interval contains zero, we don’t have any evidence of a difference between “Previous” and “Never”. This is also reflected in a large P-value on the right side of the table.

Contrast the first row to the second row of the table, the difference between “Recent” and “Never”. Here the difference between the means, 3.043674, is much larger. Furthermore, the confidence interval completely misses zero and the P-value is very small. There is a significant difference between the “Recent” and “Never” groups. (As referenced in an exercise above, though, don’t confuse statistical significance for practical significance or clinical significance.)

The last row shows a difference that is also a bit on the larger side, but the interval does contain zero and the P-value is not quite small enough by the traditional $\alpha = 0.05$ standard.

So it seems that it’s the “Recent” group that is somewhat different from the other two.

All that’s really happening here is that the `TukeyHSD` command is running two-sample t-tests between each pair of variables, coming up with both a confidence interval and a P-value.

You may have wondered what `p adj` means. These are “adjusted” P-values. Why do they need to be adjusted? Well, if $\alpha = 0.05$, that means that we have a 5% chance of making a Type I error and saying something is significant even when it’s not. This is a 1 in 20 chance. If we then run 20 t-tests, chances are that one of them will come back “significant” even though it isn’t. That’s just how the laws of probability work.

Now, of course, we’re not running 20 tests here; we’re running three. Nevertheless, every time we run an additional test on the same data, we increase the likelihood of a Type I error. One way to counteract that is to inflate the P-values a little so that only the really, really significant differences get flagged as significant. And even though it doesn’t say it, the confidence intervals are widened a bit as well.

I’m sweeping a lot of details of this under the rug. A more advanced statistics class will deal directly with this issue of “multiple comparisons” and how to adjust for them.

Your turn

Use the same data. This time determine if heroine/cocaine use during the three months prior to admission is associated with depression. Run ANOVA according to the rubric and then perform a post hoc analysis using Tukey Honest Significant Differences.