3+ Sample testing with continuous outcomes

Skipping around

Reminder, if you are following the textbook, to stick with continuous outcomes, we are going to skip ahead to chapter 24 (ANOVA) and chapter 23(regression) and bring in some outside information about non-parametric testing.

Summary of Continouous outcomes so far - the flavors of T

In R, the t.test function will allow you to conduct any of the t-tests we have covered so far

- One sample T test comparing a sample mean to a hypothesized null (if we know σ and have a large sample we could also consider a Z test) t.test(data, mu=value)
- Two sample T test comparing samples from independent populations t.test(continousvar~categoricalvar) or t.test(data1, data2)
- Two sample T test comparing paired (non-independent) groups of observations t.test(continuousvar~categoricalvar, paired=TRUE) or t.test(data1, data2, paired=TRUE)

Three sample testing

Now that we have looked at how to compare one sample to a value and means between two samples, let's extend this to a case where we have 3 samples or 3 groups that we are interested in comparing.

Today's lecture

- Introducing ANOVA - what is the null hypothesis of this test?
- Visualizing 3 sample data
- Using ANOVA to test for a difference
- Example

ANOVA

ANOVA

ANOVA or analysis of variance can be thought of as an extension of the two-sample t-test to three or more samples.

If we had 3 groups to compare, we could do this by using the two-sample t test multiple times This would result in $\binom{3}{2}$ comparisons:

$\mu_1 \neq \mu_2$ $\mu_2 \neq \mu_3$ $\mu_1 \neq \mu_3$

The problem with that approach is that we would end up with 3 p-values, one for each test performed. That doesn't tell us how likely it is that three sample means are spread apart as far as they are. If we are comparing more than 3 groups, this problem is compounded by creating even more comparisons.

We need a method that allows us to have an overall measure of confidence in all our conclusions about comparisons. This is a common problem of multiple comparisons.

ANOVA

Here we are testing a null hypothesis that all the means are the same, for 3 samples this would be

H_0: $\mu_1 = \mu_2 = \mu_3$

Our alternative hypothesis is that at least one of the means is not equal to the others

Even though your hypothesis involves means, the test compares the variability between groups to the variability within groups

ANOVA

Analysis of variance (ANOVA) is also referred to as the F test.

The ANOVA is based on two kinds of variability: - The variability among sample means or how much the individual group means vary around the overall mean - The variability within groups, how much do individual observation values vary around the group mean

If the variability within the k different populations is small relative to the variability among their respective means, this suggests that the population means are in fact different.

ANOVA

Loosely expressed:

$$F = \left(\frac{variation among sample means}{variation among observations in the same sample} \right)$$

Note: You will NOT need to know the full formula for the F-test, you will need to know how to do one in R

Data

What would the data look like in a data frame?

Data

What would the data look like in a data frame?

- One "grouping" variable (categorical)
- One continuous response variable

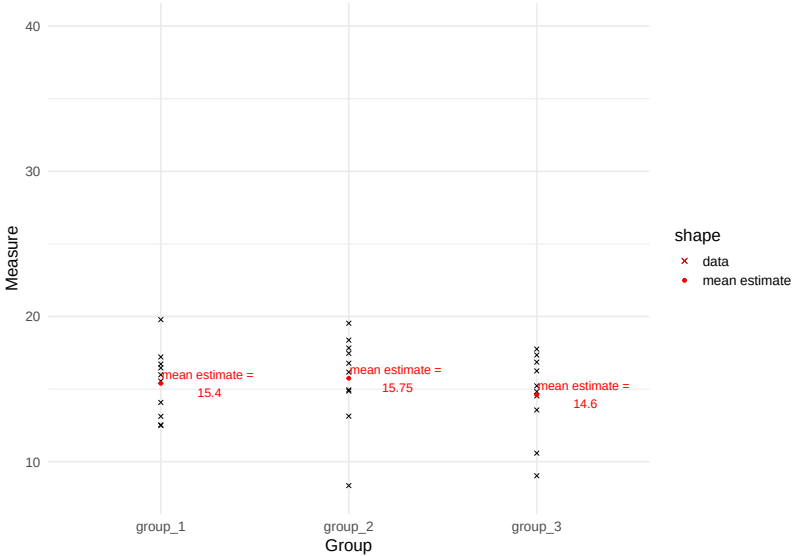
ANOVA asks if there is an association between the grouping variable and the response variable.

Visualizing first

What graphical strategies have we learned to look at variability within and between groups?

A) Is there a difference between these means?

Describe why you do or do not think so.



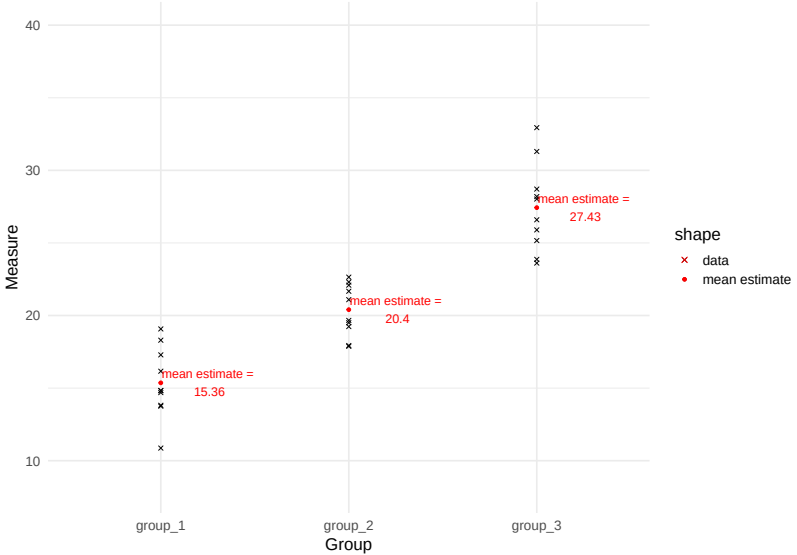
Summary of the plots

Plot (A)

- The means (red dots) were not very different across the groups. This means the variation between the group means was small.
- The distribution of the data (black Xs) was wide enough that the distribution of points for each group overlapped almost completely. This means that the variation within each group was relatively wide

B) Is there a difference between these means?

Describe why you do or do not think so.



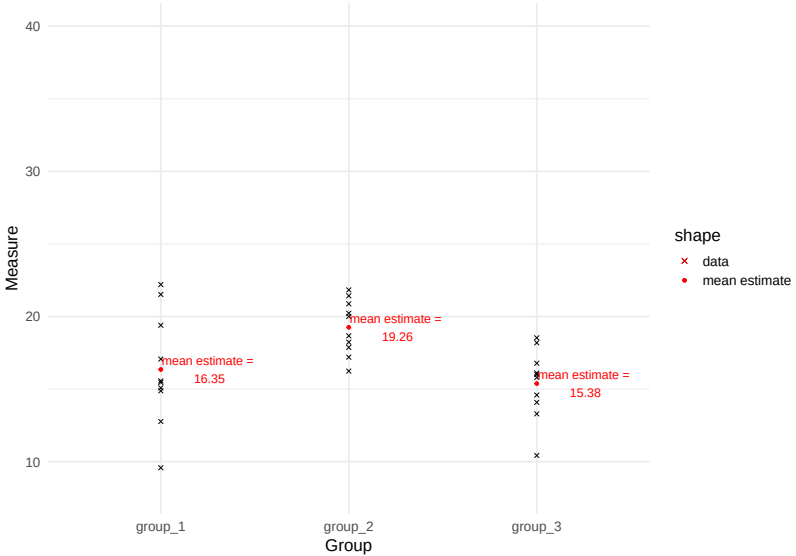
Summary of the plots

Plot (B)

- The means are quite different across the groups. The variation between the group means would be larger than in plot (A)
- The distribution of the data overlaps between groups 1 and 2 and 2 and 3, but not 1 and 3. The variation within each group is as wide as it was in Plot (A) but doesn't mask the mean differences, especially between group 1 and 3

C) Is there a difference between these means?

Describe why you do or do not think so.



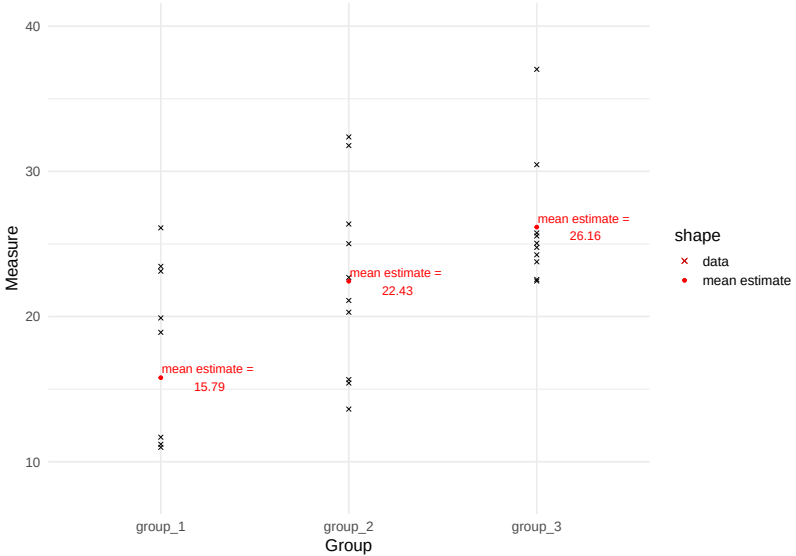
Summary of the plots

Plot (C)

- Here, the means for group 1 and 3 look similar, but the mean for group 2 appears a bit higher than the other two, though there is still overlap between the data from all the groups
- Is there evidence that at least one of the means is different?

D) Is there a difference between these means?

Describe why you do or do not think so.



Summary of the plots

Plot (D)

- Plot (D) looked like Plot (B) but with more variation within groups
- This variation makes the difference between the means harder to detect

Overall summary

- What we informally did on the previous slides was compare the variation between group means to the variation within the groups
- This focus on variation is why this test is called ANOVA: an ANalysis Of VAriance
- When the ratio of between vs. within variation is large enough then we detect a difference between the groups
- When the ratio isn't large enough we don't detect the difference.
- This ratio is our test statistic, denoted by F

Descriptive plots

Descriptive plots

What other ways to present the data visually have we learned that might be useful before we move on to testing?

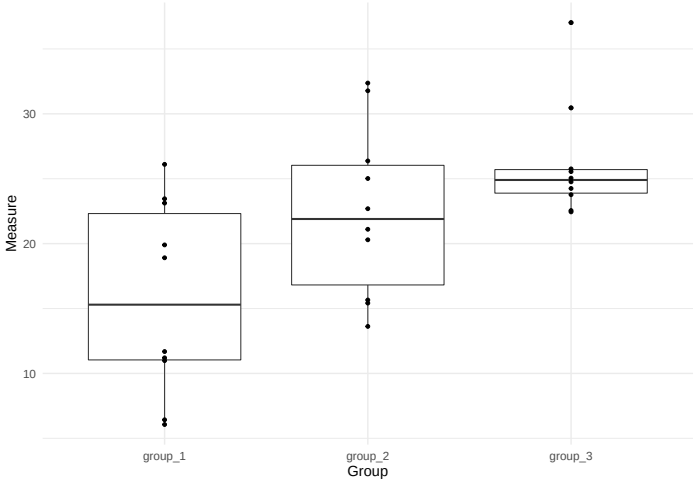
Descriptive plots

How would you want to plot these data before you conduct a test?

- Option 1: Box plot for each level of the grouping variable (with overlaid data points)

```
ggplot(diff_3_narrow, aes(x = Group, y = Measure)) +
geom_boxplot() +
geom_point() +
theme_minimal(base_size = 15)
```

Box plot



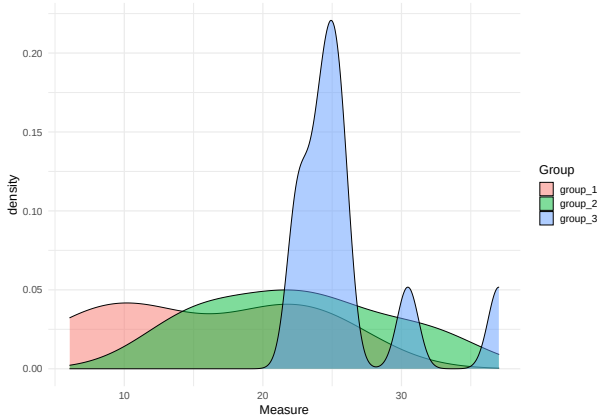
Descriptive plots

How would you want to plot these data before you conduct a test?

- Option 2: Density plot for each level of the grouping variable

```
ggplot(diff_3_narrow, aes(x = Measure)) +
geom_density(aes(fill = Group), alpha = 0.5) +
theme_minimal(base_size = 15)
```

Density plot



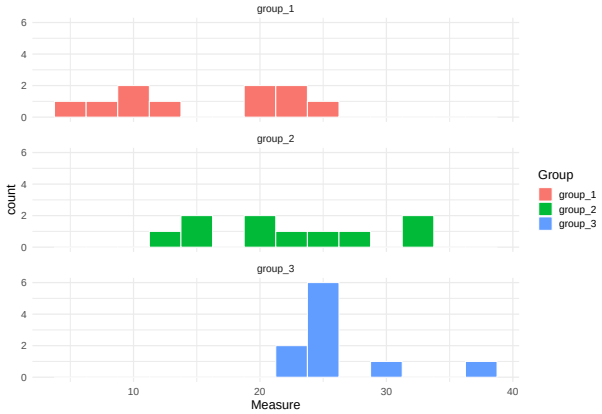
Descriptive plots

How would you want to plot these data before you conduct a test?

- Option 3: Histogram for each level of the grouping variable

```
ggplot(diff_3_narrow, aes(x = Measure)) +
geom_histogram(aes(fill = Group), col = "white", binwidth = 2.5) +
theme_minimal(base_size = 15) +
facet_wrap(~ Group, nrow = 3)
```

Histograms with facet wrap



Testing with ANOVA

The hypotheses

Null hypothesis

$H_0 : \mu_1 = \mu_2 = ...\mu_K$, where K is the number of levels of the grouping variable

- Can you also state the null hypothesis in words?

The hypotheses

Alternative hypothesis

H_a : not all μ_1, μ_2,..., μ_K are equal

- In words: Not all means are the same. Or, at least one of the means differs from the others.

The test statistic (ANOVA F Statistic)

$$F = \frac{\text{variation among group means}}{\text{variation among individuals in the same group}}$$

- Numerator is, fundamentally, the variance of the sample means
- Denominator is, fundamentally, an average of the group variances.
- The F statistic follows an F distribution
- Computation details are at the end of the book chapter (these computation details will not be tested)

The F distribution

- Skewed right
- Take only positive values
- The F distribution depends on the number of means being compared and the sample size for each of the groups
- Let k be the number of groups being compared and $N_{Total} = n_1 + n_2 + ... + n_k$ (the total sample size across all the groups)
- Then the F statistic follows an F distribution with $k - 1$ degrees of freedom in the numerator and $N_{Total} - k$ degrees of freedom in the denominator
- The p-value of the ANOVA F statistic is always the area to the right of the test statistic

ANOVA in R: use `aov()`, then `tidy()` it up!

- `aov()` stands for analysis of variance

The general syntax for the ANOVA is:

aov(outcomevariable ~ groupvaraible, data=dataset)

We will save the output of this as an object and then use tidy(object) to get the output we want.

reference: https://broom.tidyverse.org/reference/anova_tidiers.html

ANOVA in R: use `aov()`, then `tidy()` it up!

We will focus on two parts of the output from this package:

- `statistic` is the F test statistic, the ratio of the variation between means vs. the variation within groups.
- `p.value` is the p-value for the test.

p of an f statistic in R

You can check that you can calculate the p-value from the F distribution. Remember, that you need to specify a degrees of freedom for the numerator and for the denominator:

pf(value, df1=numerator degrees of freedom, df2= denominator degrees of freedom, lower.tail=F)

This general pattern of syntax should look familiar by now. . . .

Conditions for using ANOVA

Conditions for ANOVA

Condition 1: k independent SRSs, one from each of k populations.

- The most important assumption, because this method, like the others in Part III of the course, depends on the premise of having taken a random sample.

Conditions for ANOVA

Condition 2: Each of the k populations has a Normal distribution with an unknown mean μ_i.

- This assumption is less necessary
- The ANOVA test is robust to non-Normality.
- Remember that the ANOVA is based on comparing the differences of sample means
- What did the CLT tell us about variability of sample means when the samples are not normally distributed?

Conditions for ANOVA

Condition 3: All the populations have the same standard deviation σ, whose value is unknown.

- Hardest condition to satisfy and check
- If this condition is not satisfied ANOVA is often okay if the sample sizes are large enough and if they are similar across the groups
- Can use `group_by()` and `summarize()` to calculate the sample SDs to see if they're similar and indicative that the population parameters are too
- General rule: we want the largest sample standard deviation to be less than twice as largest as the smallest one. I.e., $s_{max}/s_{min} < 2$

Conditions for ANOVA