Lecture 34: Analysis of Variance (ANOVA)

Comparing means in more than two groups

Instructor: Tomer Altman

November 19, 2025

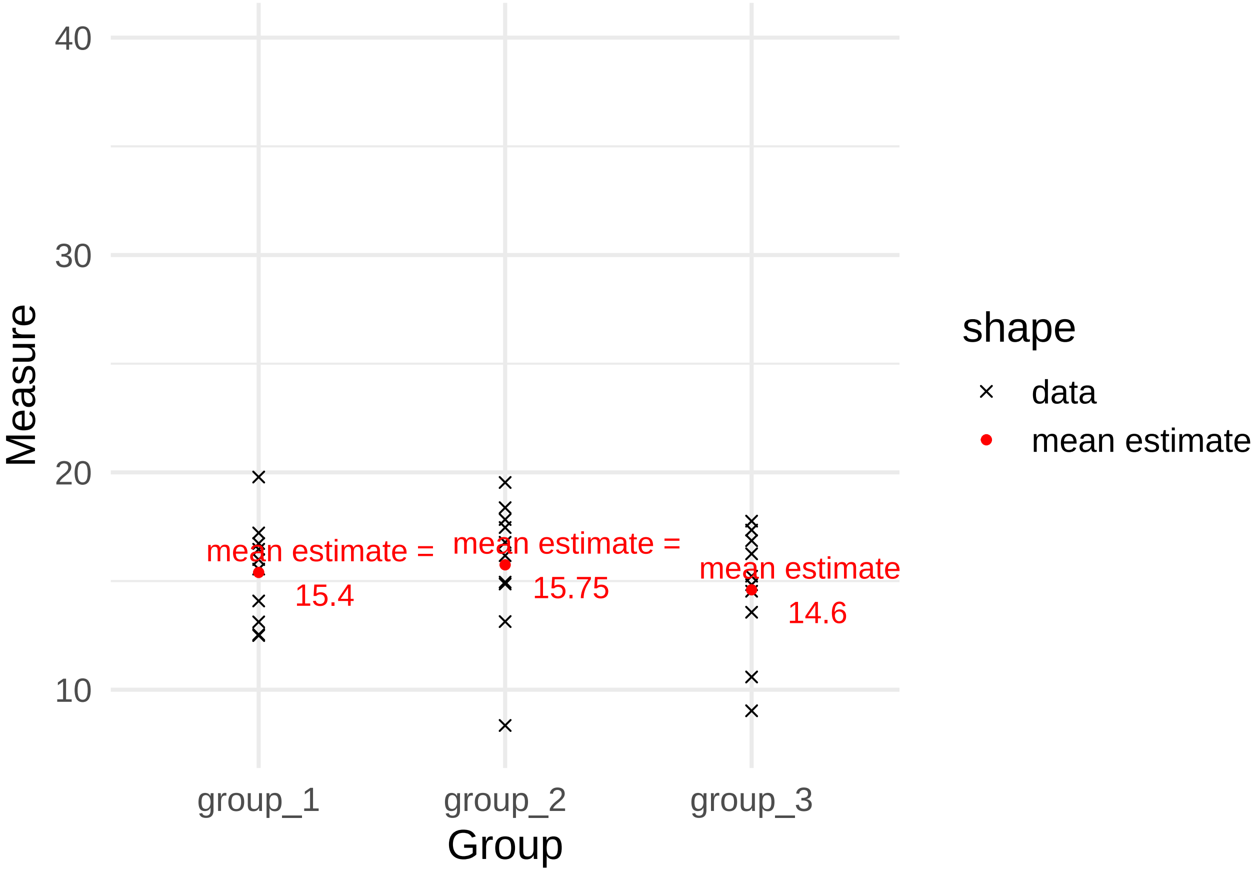
Recap

- When we have two groups, and for each group we know each individual's measure for some continuous variable, we can compare the groups means using a two-sample t-test
- The null hypothesis is that the means are the same ($\mu_1 = \mu_2$)
- What do we do if we have more than three groups that we'd like to compare on some continuous measure?

Learning objective for today

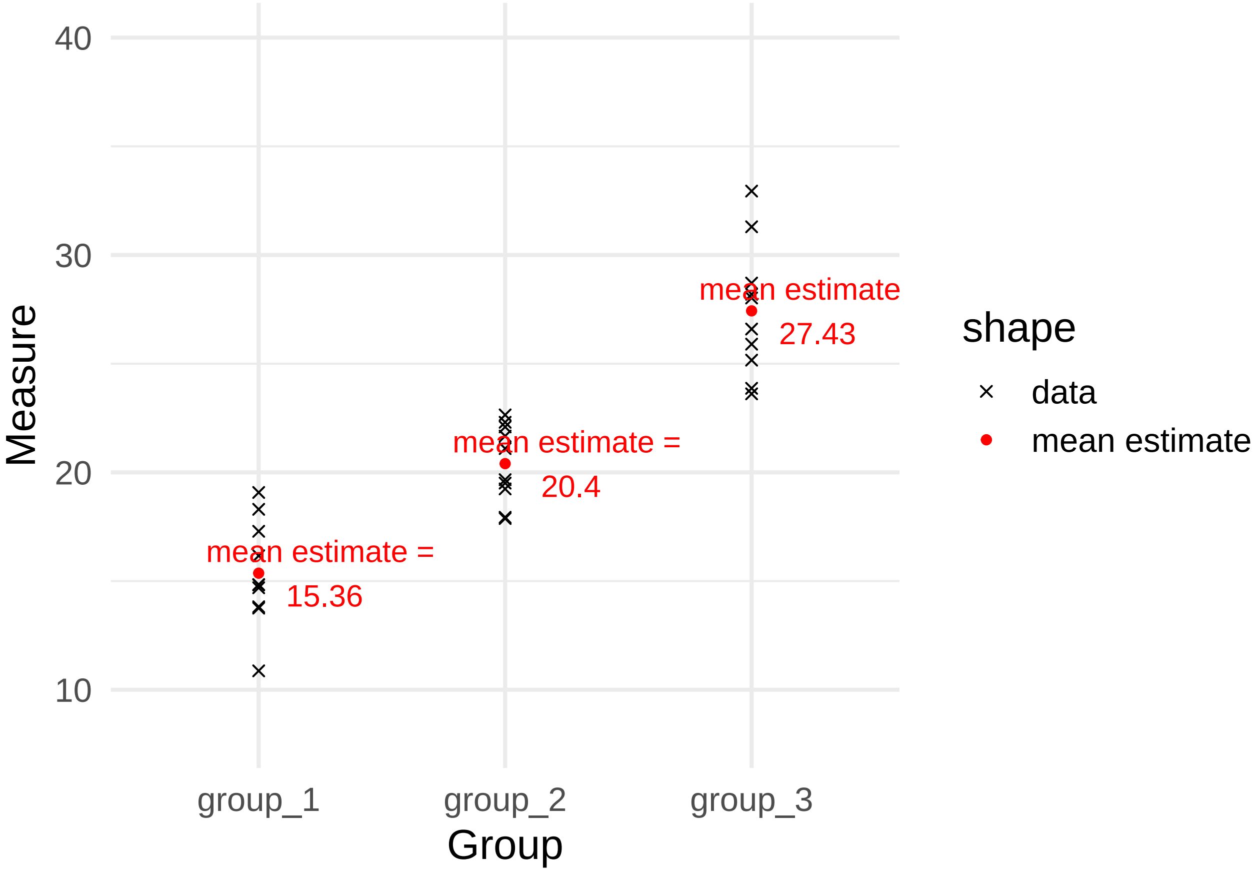
- Learn how to conduct a hypothesis test to evaluate whether there is a difference across multiple means. This test is known as the analysis of variance, or ANOVA
- Conduct this test using the `aov()` function in R
- Then, learn how to detect specifically which means are different from one another using Tukey's HSD test in R

Example A: Is there a difference between these means?



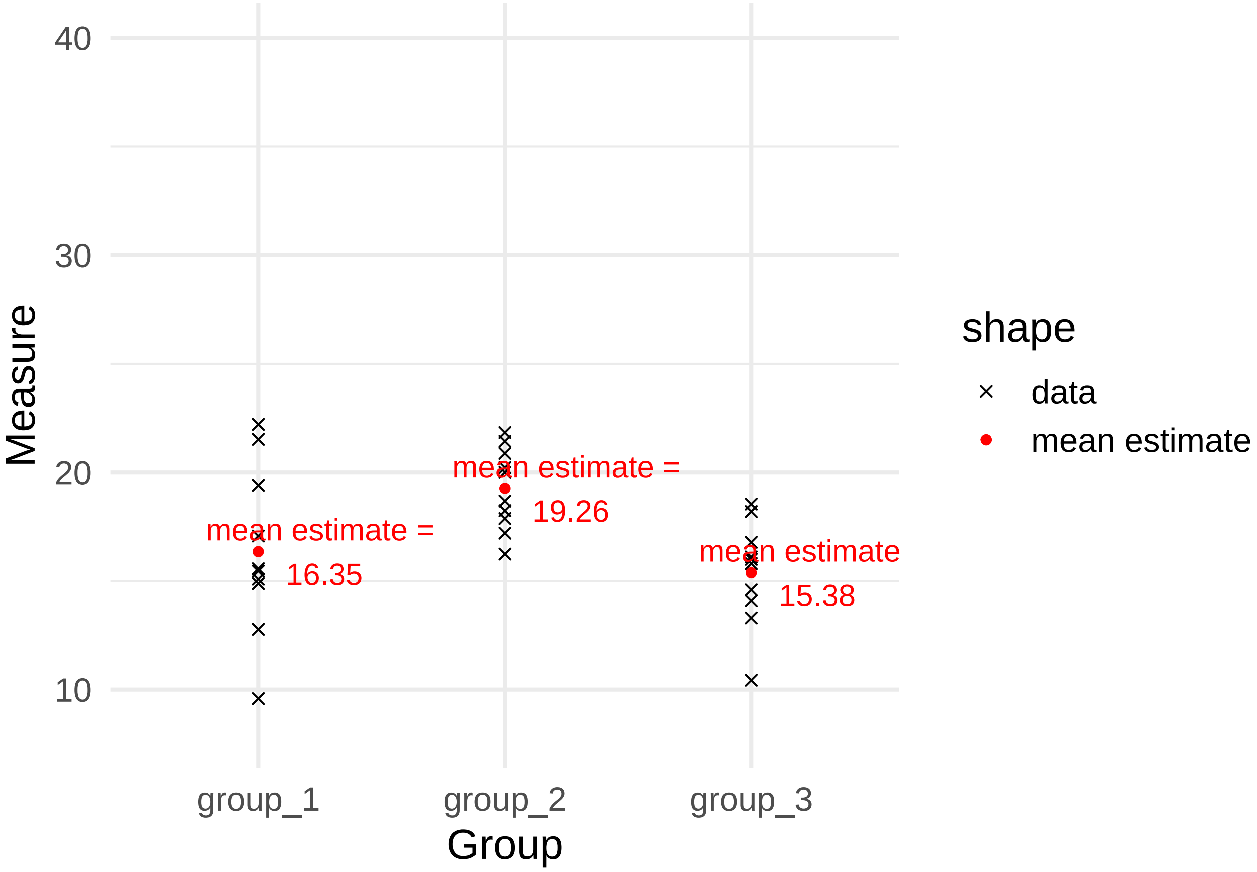
- The means (red dots) are not very different across the groups. This means the variation **between** the group means is small.
- The distribution of the data (black "X" marks) is wide enough that the distribution of points for each group overlap almost completely. This means that the variation **within** each group is relatively wide

Example B: Is there a difference between these means?



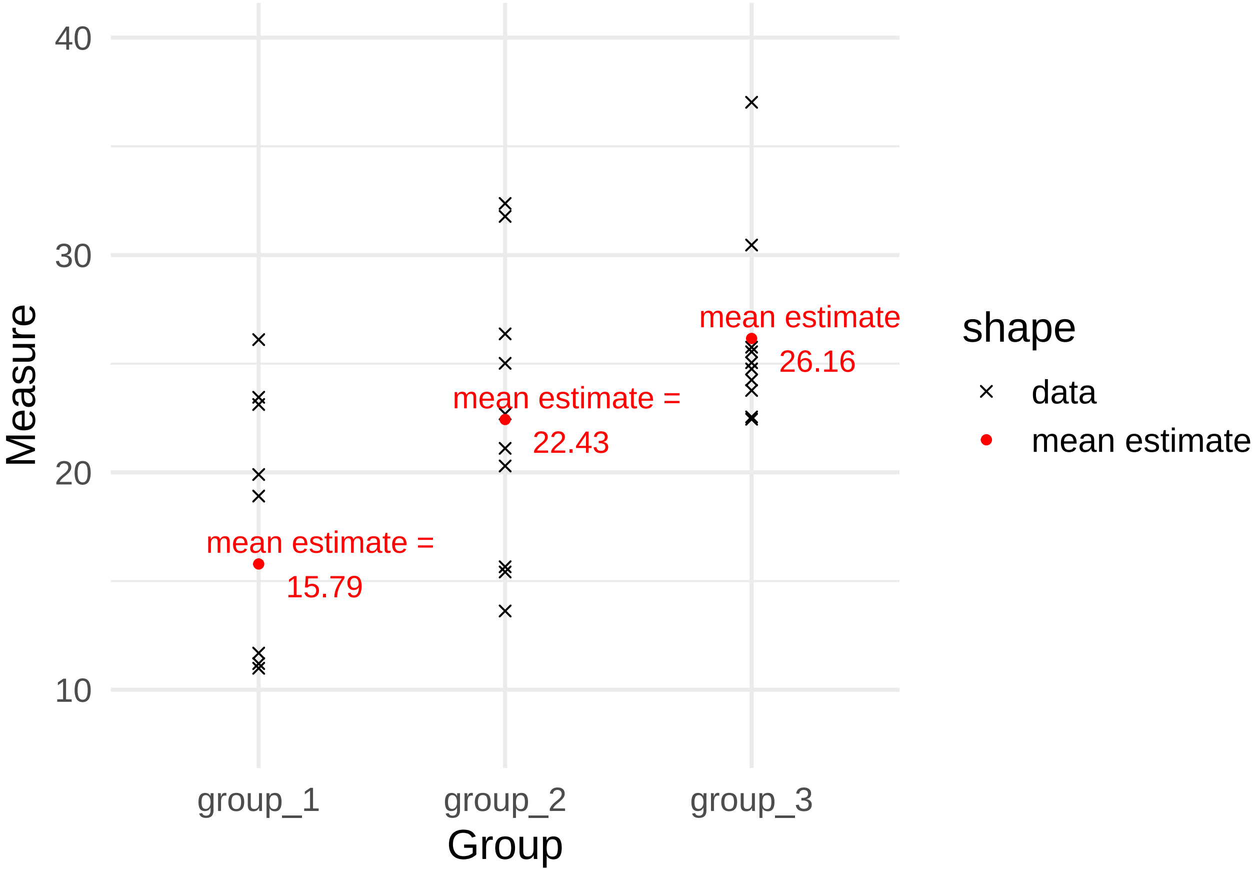
- The means are quite different across the groups. The variation **between** the group means is larger than in plot (A).
- The distribution of the data overlaps between groups 1 and 2 and 2 and 3, but not 1 and 3. The variation **within** each group is as wide as it was in Plot (A) but doesn't mask the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

Example C: Is there a difference between these means?



- 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?

Example D: Is there a difference between these means?



- Plot (D) looks 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 **AN**alysis **O**f **VA**riance
- 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 can't detect the difference.
- This ratio is our test statistic, denoted by F

Shiny App

Try out this Shiny App.

- Try changing the population standard deviation, σ
- Try decreasing the group sample sizes
- Try moving the means around (to increase or decrease the SD **between** groups)

After each change, notice how the F statistic changes. A higher F implies that there is much more variation between vs. within the groups. Notice also how the p-value for the test changes.

Analysis of Variance (ANOVA)

- ANOVA is used to compare the means of more than two groups when the comparison variable is continuous
- ANOVA asks either of the following:
  - "Are the means different from each other?"
  - "Are one or more of the means different from the others?"

Data

What would the data look like in a data frame?

Data

What would the data look like in a data frame?

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

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

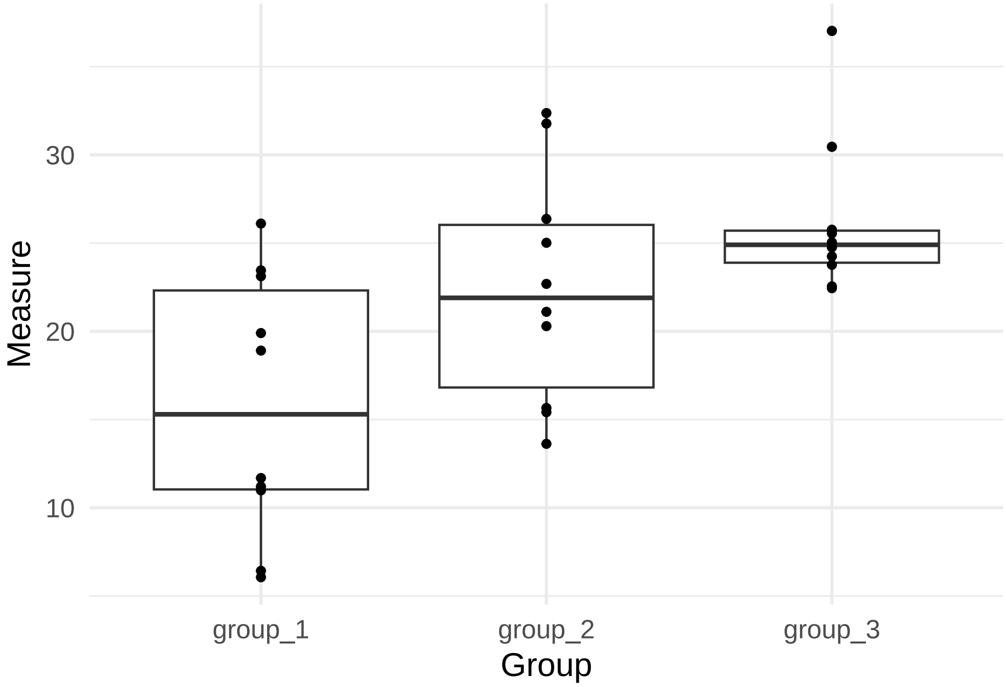
What test asks if there is an association between two categorical variables?

Check your understanding!

Descriptive plots: Box 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)

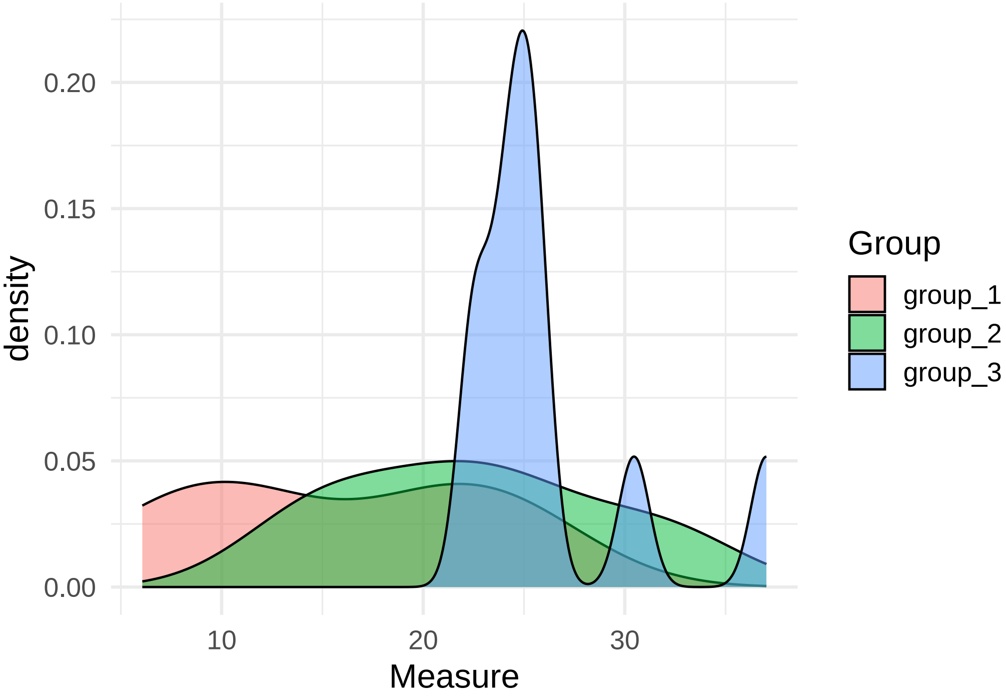
```
ggplot(diff_3_narrow, aes(x = Group, y = Measure)) +
  geom_boxplot() +
    geom_point() +
  theme_minimal(base_size = 15)
```



Descriptive plots: Density 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

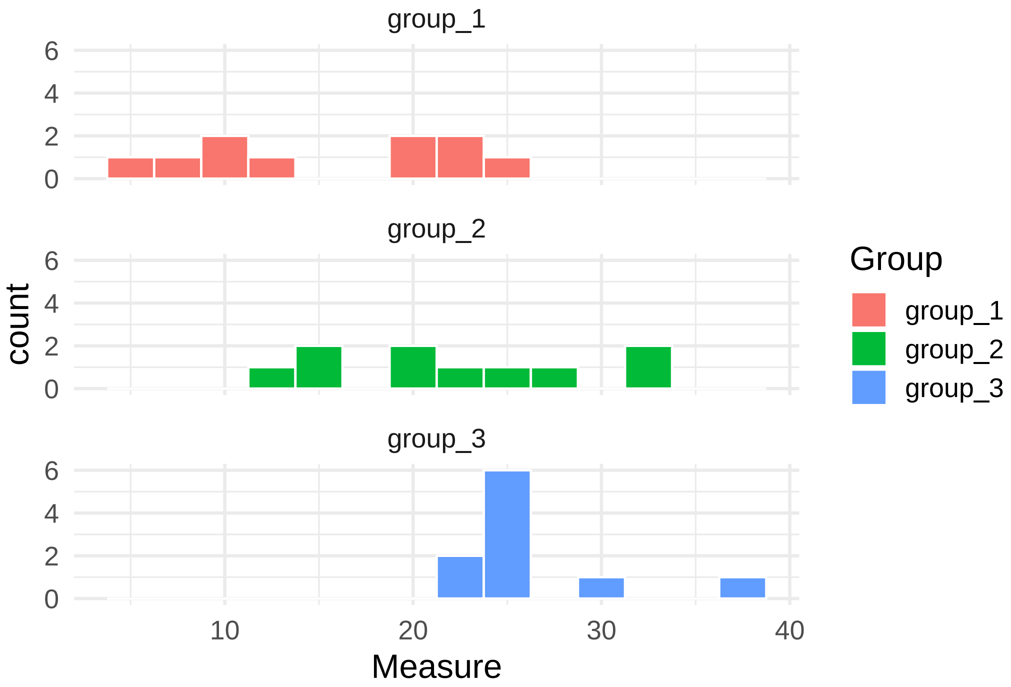
```
ggplot(diff_3_narrow, aes(x = Measure)) +
  geom_density(aes(fill = Group), alpha = 0.5) +
  theme_minimal(base_size = 15)
```



Descriptive plots: Histograms

- 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)
```



Hypothesis testing

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?

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.

Example: Cannabis to treat brain cancer in mice

- High-grade glioma is an aggressive type of brain cancer with a low long-term survival rate
- Cannabinoids, chemical compounds found in cannabis, are thought to inhibit glioma cell growth
- Researchers transplanted glioma cells into otherwise-healthy mice, and then randomly assigned these mice to 4 cancer treatments: irradiation alone, cannabinoids alone, irradiation combined with cannabinoids, or no treatment
- The treatments were administered for 21 days, after which the glioma tumor volume (in cubic millimeters) was assessed in each mouse using brain imaging

The data

```
treatment <- c(rep("Irradiation", 4),
               rep("Cannabinoids", 5),
               rep("Both", 6),
               rep("Neither", 7))

tumor_volume <- c(30, 46, 46, 95, # Irradiation
                  12, 14, 16, 41, 47, # Cannabinoids
                  5, 4, 4, 4, 10, 9, # Both
                  24, 30, 43, 51, 62, 32, 96) # Neither

cancer_data <- data.frame(treatment, tumor_volume)
```

```
head(cancer_data, 15)
```

```
##       treatment tumor_volume
## 1   Irradiation           30
## 2   Irradiation           46
## 3   Irradiation           46
## 4   Irradiation           95
## 5  Cannabinoids           12
## 6  Cannabinoids           14
## 7  Cannabinoids           16
## 8  Cannabinoids           41
## 9  Cannabinoids           47
## 10         Both            5
## 11         Both            4
## 12         Both            4
## 13         Both            4
## 14         Both           10
## 15         Both            9
```

Organize the data

- Think about how you want the data to look
- I want to plot the raw data points and display the mean for each treatment group
- I also want to specify the order that the treatment groups show up in the plot

```
# specify the order of the treatment groups for plotting
library(forcats)
cancer_data <- cancer_data %>%
  mutate(trt_order = fct_relevel(treatment,
                                 c("Neither", "Irradiation",
                                   "Cannabinoids", "Both")))

# calculate the means and SD for each group
summary_stats <- cancer_data %>%
  group_by(trt_order) %>%
  summarise(mean_vol = mean(tumor_volume),
            sd_vol = sd(tumor_volume),
            samp_size = n())

summary_stats
```

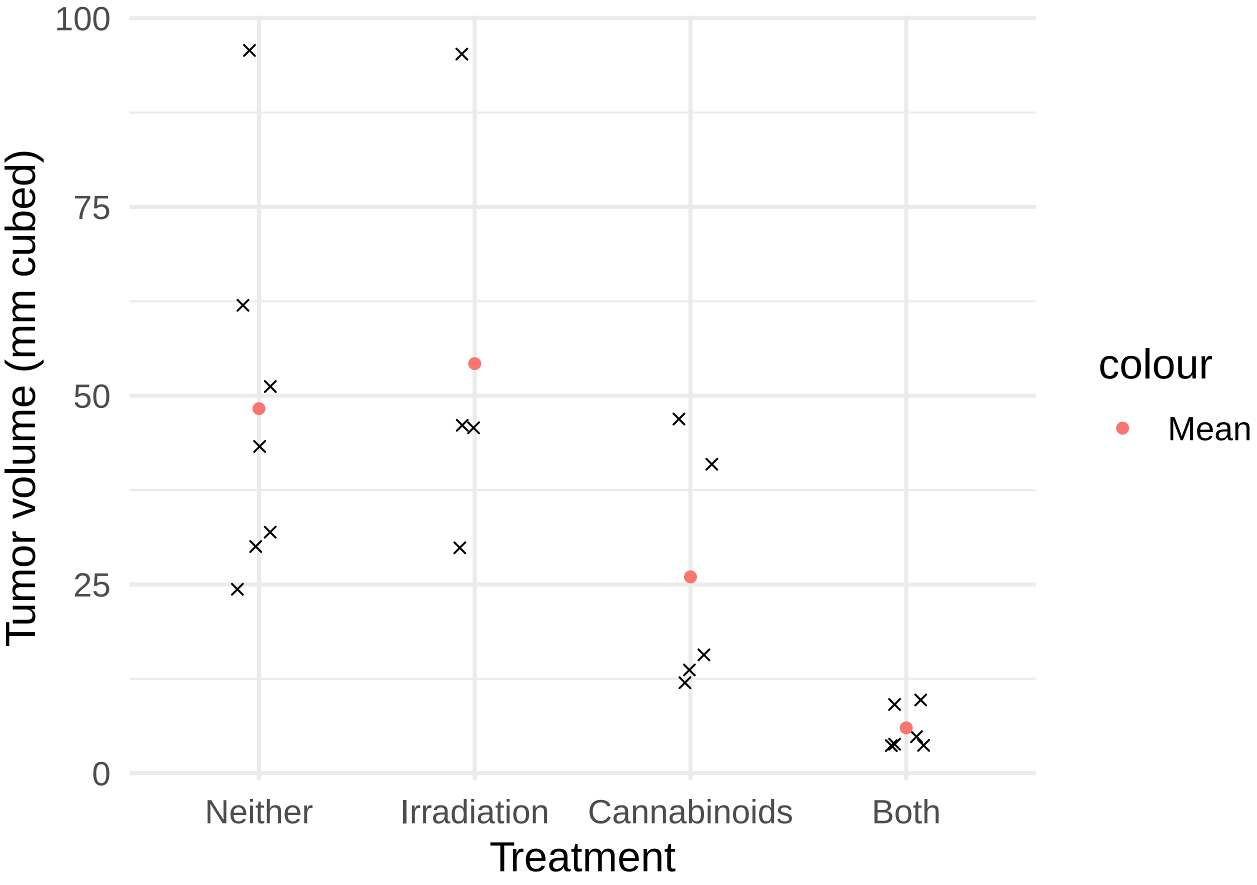
```
## # A tibble: 4 x 4
##   trt_order    mean_vol sd_vol samp_size
##   <fct>           <dbl>  <dbl>     <int>
## 1 Neither          48.3  24.8          7
## 2 Irradiation      54.2  28.2          4
## 3 Cannabinoids     26    16.6          5
## 4 Both              6     2.76         6
```

Plot the data

```
jitter.plot <- ggplot(cancer_data, aes(x = trt_order,
                                       y = tumor_volume)) +
  geom_jitter(pch = 4, width = 0.1) + # to prevent over-plotting
  geom_point(data = summary_stats,
             aes(y = mean_vol, col = "Mean"),
             pch = 19) +
  labs(y = "Tumor volume (mm cubed)", x = "Treatment") +
  theme_minimal(base_size = 15)
```

The `geom_jitter()` function with width $= 0.1$ randomly "jitters" the location of the points along the x axis so that we can see each of them since some have the exact same values.

```
jitter.plot
```



The ANOVA F test statistic

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

- Numerator is the variance of the sample means
- Denominator is an average of the group variances
- Under assumptions and the null hypothesis of no mean differences, the F statistic follows an F distribution (much like when the χ^2 statistic follow the χ^2 distribution)

The ANOVA F test statistic

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

$$F = \frac{\text{mean squares for groups}}{\text{mean squares for error}} = \frac{MSG}{MSE}$$

Numerator: Mean squares for groups (MSG)

- Let $\bar{x}$ represent the overall sample mean (across all the groups)
- The MSG is like an average of the k squared deviations, where groups with large samples are up-weighted

$MSG = \frac{n_1(\bar{x}_1-\bar{x})^2+n_2(\bar{x}_2-\bar{x})^2+....+n_k(\bar{x}_k-\bar{x})^2}{k-1}$

- Each $(\bar{x}_i - \bar{x})^2$ takes a squared difference between group i's mean and the overall mean. Thus, the larger the MSG, the further away the group means are from the overall mean, and the further away they are from each other in a global sense.

The numerator of the MSG is also called the **sum of squares for groups**:

$MSG = \frac{\text{sum of squares for groups}}{k-1}$

Denominator: Mean squares for error (MSE)

- Let the variance for each group be represented by s_i^2. The variance is our best measure of variation among individuals in the same group.
- The MSE is like a weighted average of the variation among individuals with the same group:

$MSE = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+...+(n_k-1)s_k^2}{N_{Total}-k}$

- A higher MSE means there is more variation among individuals within groups.
- The numerator of the MSE is also called the **sum of squares of error**:

$MSE = \frac{\text{sum of squares of error}}{N_{Total}-k}$

The ANOVA F test statistic

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

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

- We are comparing the variation across the groups to the variation among individuals in the same group
- If the F statistic is high, then there is relatively more variation across groups than there is within groups
- If the F statistic is less than one, then there is more variation across individuals in the same group, then there is between group means
- Go back to the Shiny App and move things around. See how the F-statistic changes. When is the F-statistic very high vs. when is it less than one?

The F distribution

- Skewed right
- Takes 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

```r
#reference: https://broom.tidyverse.org/reference/anova_tidiers.html
library(broom)
cancer_anova <- aov(formula = tumor_volume ~ treatment, data = cancer_data)
tidy(cancer_anova)
```

```
## # A tibble: 2 x 6
##   term         df sumsq meansq statistic  p.value
##   <chr>     <dbl> <dbl>  <dbl>     <dbl>    <dbl>
## 1 treatment     3 8060.  2687.      6.70  0.00313
## 2 Residuals    18 7218.   401.     NA    NA
```

- `df` displays the numerator and denominator degrees of freedom for this data-set
- `sumsq` displays the **sum of squares for groups** and **sum of squares for error**, and `meansq` displays the MSG and MSE, respectively
- You can calculate the `meansq` column by taking `sumsq/df`
- `statistic` is the F test statistic, the ratio of the MSG and MSE. This F says that the variation between the means is nearly 7 times as large as the variation within the groups.
- `p.value` is the p-value for the test. This p-value is equal to 0.003. There is a 0.3% chance of observing the F statistic we observed (or more extreme) under the null hypothesis that all the means are the same. This chance is very low so we reject the null hypothesis in favor of the alternative hypothesis that at least one of the means differs from the others.

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

```r
tidy(cancer_anova)
```

```
## # A tibble: 2 x 6
##   term         df sumsq meansq statistic  p.value
##   <chr>     <dbl> <dbl>  <dbl>     <dbl>    <dbl>
## 1 treatment     3 8060.  2687.      6.70  0.00313
## 2 Residuals    18 7218.   401.     NA    NA
```

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:

```r
pf(6.699489, df1 = 4 - 1, df2 = 22 - 4, lower.tail = F)
```

```
## [1] 0.003131703
```

The p-value equals 0.003. Under the null hypothesis of no difference between the group means, there is a 0.3% chance of observing the F-statistic that we calculated or a more extreme one. This is a very small probability, and provides evidence against the null in favor of the alternative hypothesis that at least one mean is different from the others.

Next steps

- The interpretation of the p-value leaves something to be desired: **Which** group or groups is different from the others?

- You could look at all pairwise differences (i.e., comparing each combination of two treatments to each other using two-sample test), but we have to be careful because we will find differences "just by chance" if we compare enough groups

Tukey's honestly significant differences (Tukey's HSD)

- Tukey's test maintains a 5% **experiment-wise** or **"family"** error rate
- Even if you make every pairwise comparison, the overall error rate is fixed at $\leq 5\%$
  - The probability of any false null among all tests performed α, typically 0.05
- Using Tukey's HSD overcomes the issue of *multiple testing*, which is a huge issue in the era of Big Data
  - Recall: If you conducted 100 tests with a 5% error rate (i.e., $\alpha = 0.05$) AND the H_0 was always true, how many p-values would you expect to be < 0.05?

`TukeyHSD()` to calculate the differences in R

Here is the R code and output:

```
diffs <- TukeyHSD(cancer_anova, conf.level = 0.95) %>% tidy()
diffs
```

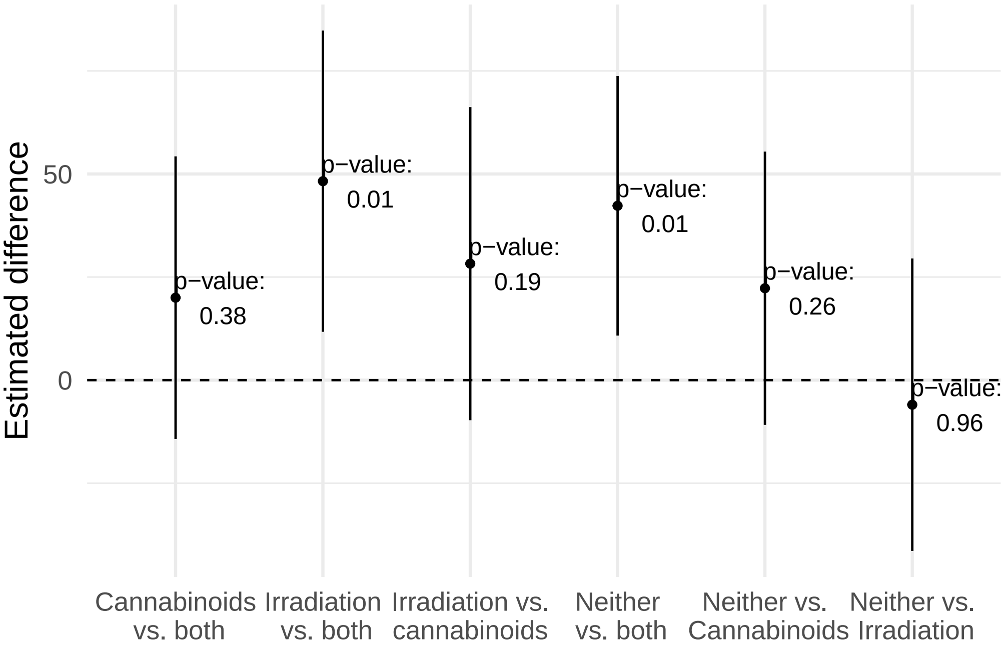
```
## # A tibble: 6 x 7
##   term      contrast          null.value estimate conf.low conf.high adj.p.value
##   <chr>     <chr>                  <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
## 1 treatment Cannabinoids-Both          0    20.0    -14.3       54.3     0.378
## 2 treatment Irradiation-Both           0    48.2     11.7       84.8     0.00756
## 3 treatment Neither-Both               0    42.3     10.8       73.8     0.00661
## 4 treatment Irradiation-Cann~          0    28.2     -9.72      66.2     0.190
## 5 treatment Neither-Cannabin~          0    22.3    -10.9       55.4     0.263
## 6 treatment Neither-Irradiat~          0    -5.96   -41.4       29.5     0.964
```

Each row in the table corresponds to a pairwise test. So the first row is looking at the difference between Cannabinoids vs. Both treatments. The estimated difference in means is 20 and the 95% CI is -14.3 to 54.3. The adjusted p-value is 0.38.

- "Adjusted" means that it is adjusted for conducting multiple tests. The unadjusted p-value would be smaller. The unadjusted p-value would be < 0.05 if the 95% CI didn't include 0.
- **Thus, when you have an adjusted test you can't use the CI to infer the value of the p-value!**

Visualize the pairwise differences

```
ggplot(diffs, aes(x = contrast, y = estimate)) + geom_point() +
  geom_segment(aes(y = conf.low, yend = conf.high, xend = contrast)) +
  theme_minimal(base_size = 15) +
  geom_hline(aes(yintercept = 0), lty =2) +
  geom_text(aes(label = paste0("p-value:\n ", round(adj.p.value, 2))), nudge_x = 0.3) +
  labs(y = "Estimated difference", x = "") +
  scale_x_discrete(labels = c("Cannabinoids\n vs. both", "Irradiation\n vs. both",
                              "Irradiation vs.\ncannabinoids", "Neither\n vs. both",
                              "Neither vs.\n Cannabinoids", "Neither vs.\n Irradiation"))
```



Using Tukey's HSD, we would conclude that the mean for treatment "Irradiation" is different from the mean for treatment "Both", and the mean for treatment "Neither" is different from the mean for treatment "Both".

Even though these two CIs don't overlap with the null value, for the other four comparisons, their adjusted p-values are $> 5\%$, so we cannot reject the null hypothesis.

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 not as crucial, but still has an impact
- That is, if the sample size is sufficiently large, the ANOVA test is **robust** to non-Normality
- What matters more is Normality of the sample means (guaranteed as n increases because of the CLT)
- If the sample size is small (say 4-5 individuals per group) then need data that is symmetric with no outliers (however, hard to tell if only have 4-5 individuals - should consider other methods like permutation)

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 standard deviations to see if they're similar and indicative that the population parameters are too
- Rule of thumb: want the largest sample standard deviation to be less than twice as large as the smallest one. I.e., $\frac{s_{max}}{s_{min}} < 2$

Good video on running ANOVA in R

ANOVA Multiple Comparisons & Kruskal Wallis in R by Mike Marin and Ladan Hamadani

https://youtu.be/lpdFr5SZR0Q

- 5 minutes long
- The video also talks about the Kruskal-Wallis test, which is a non-parametric alternative to ANOVA that we will go over next time

Check your understanding!