Comparing Two or more Means

Stat 120

February 24 2023

Inference tools (Classical methods)

Categorical Response

1. One proportion: sample z test/CI
2. Difference in 2 props: 2 sample z test/CI OR chi-square test
3. Association between 2 categorical variables: chi-square test

Quantitative Response

1. One mean: 1 sample t test/CI
2. Difference in 2 means: 2 independent sample t test/CI OR Matched pairs
3. Compare >2 means: One-way ANOVA

Multiple Categories

So far, we've learned how to do inference for a difference in means IF the categorical variable has only two categories (i.e. compare two groups)

In this section, we'll learn how to do hypothesis tests for a difference in means **across multiple categories (i.e. compare more than two groups)**

Hypotheses

To test for a difference in true/population means across k groups:

$$
\begin{aligned}
H_0: &\quad \mu_1 = \mu_2 = \ldots = \mu_k \\
H_a: &\quad \text{At least one } \mu_i \neq \mu_j
\end{aligned}
$$

Frisbee Example

Does Frisbee grip affect the distance of a throw?

A student performed the following experiment: 3 grips, 8 throws using each grip

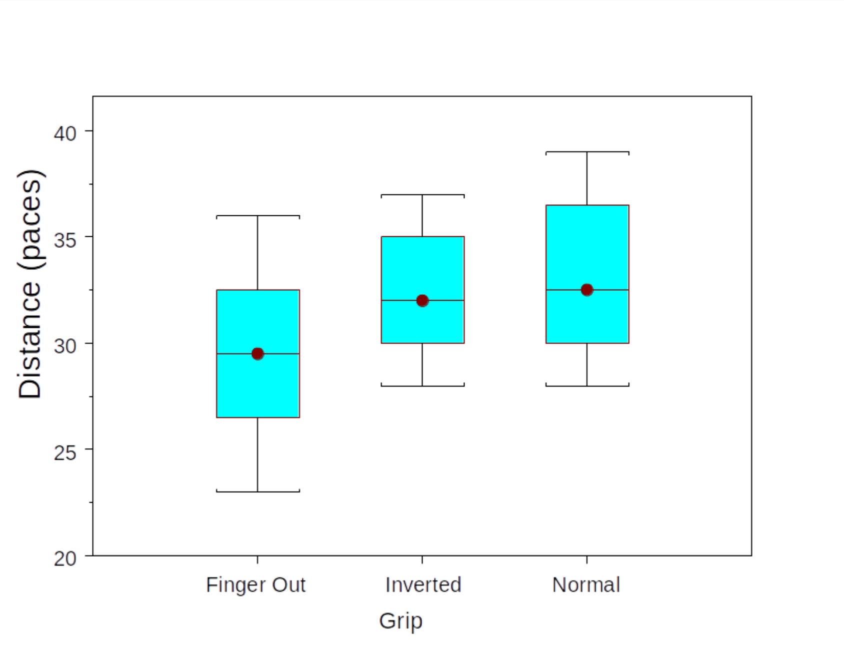
1. **Normal grip**
2. **One finger out grip**
3. **Frisbee inverted grip**

A grip type is randomly assigned to each of the 24 throws she plans on making

✨***Response: measured in paces how far her throw went***

✨***Question: How might you summarize her data?***

Frisbee Example



	Finger-out	Inverted	Normal
n	8	8	8
Mean	29.5	32.375	33.125
SD	4.175	3.159	3.944

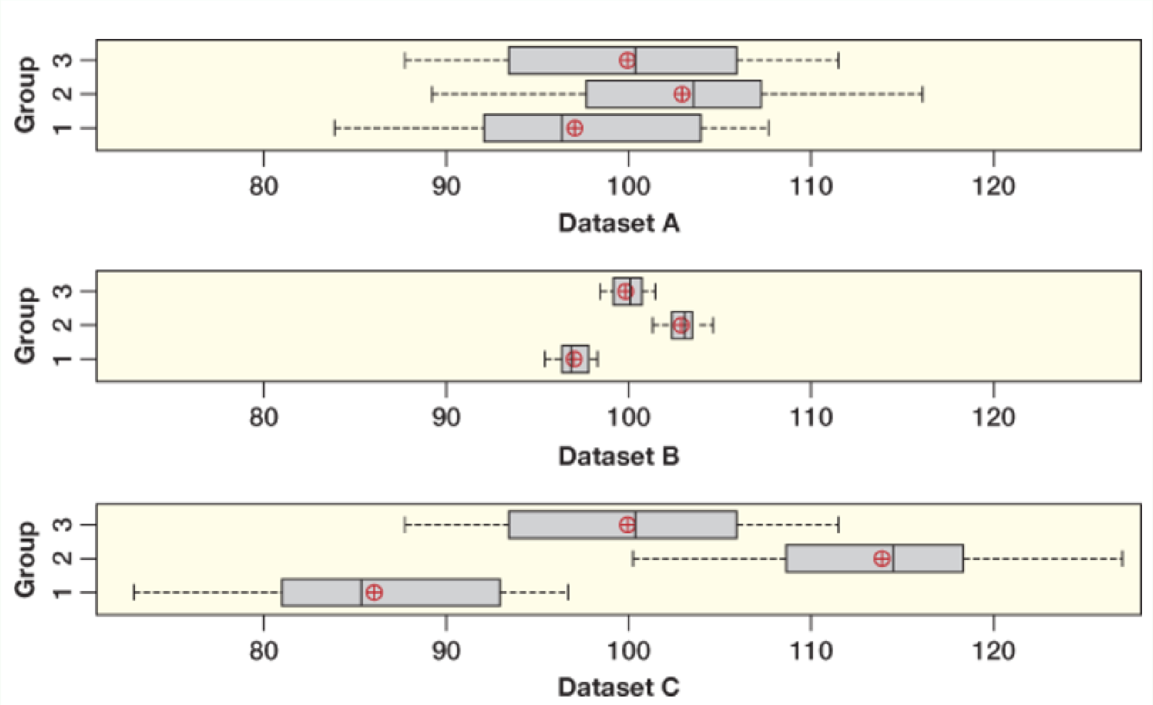
Question: Is this evidence that grip affects mean distance thrown?

$$H_0: \quad \mu_1 = \mu_2 = \mu_3$$
$$H_a: \quad \text{At least one } \mu_1, \mu_2, \mu_3 \text{ is not the same}$$

Why Analyze Variability to Test for a Difference in Means?

✨The group means in Datasets A and B are the same, but the boxes show different spread.

✨Datasets A and C have the same spread for the boxes, but different group means.

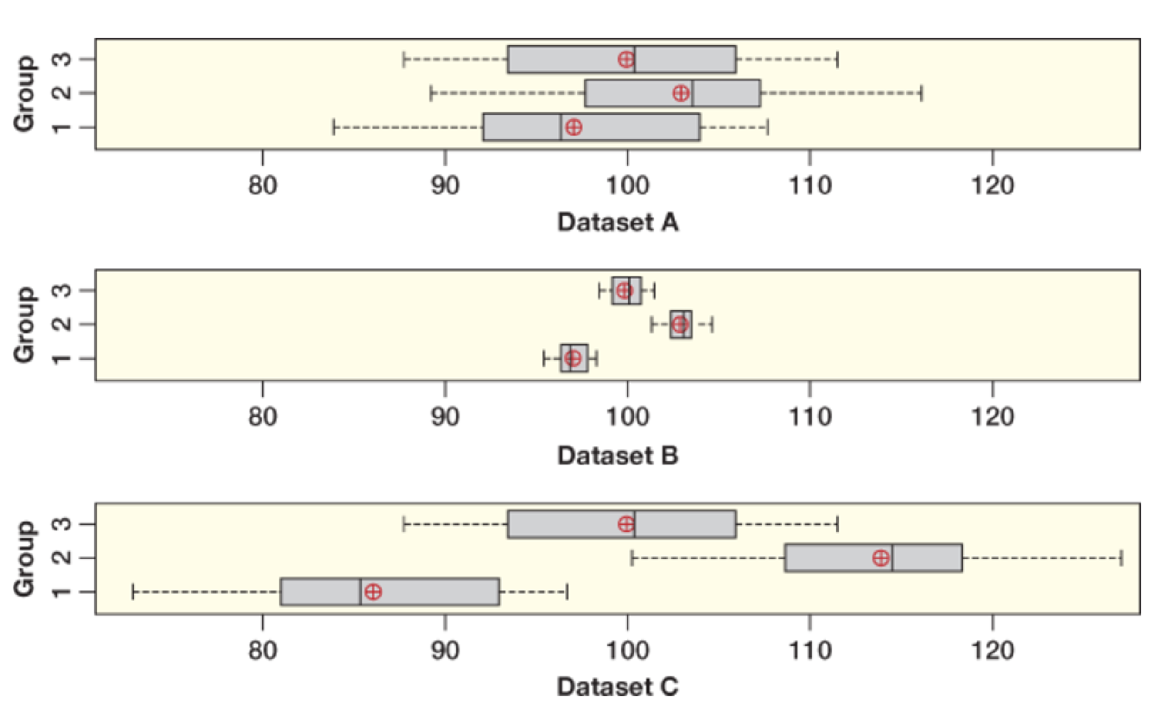


Which of these graphs appear to give strong visual evidence for a difference in the group means?

Why Analyze Variability to Test for a Difference in Means?

✨Dataset A = weakest evidence for a difference in means.

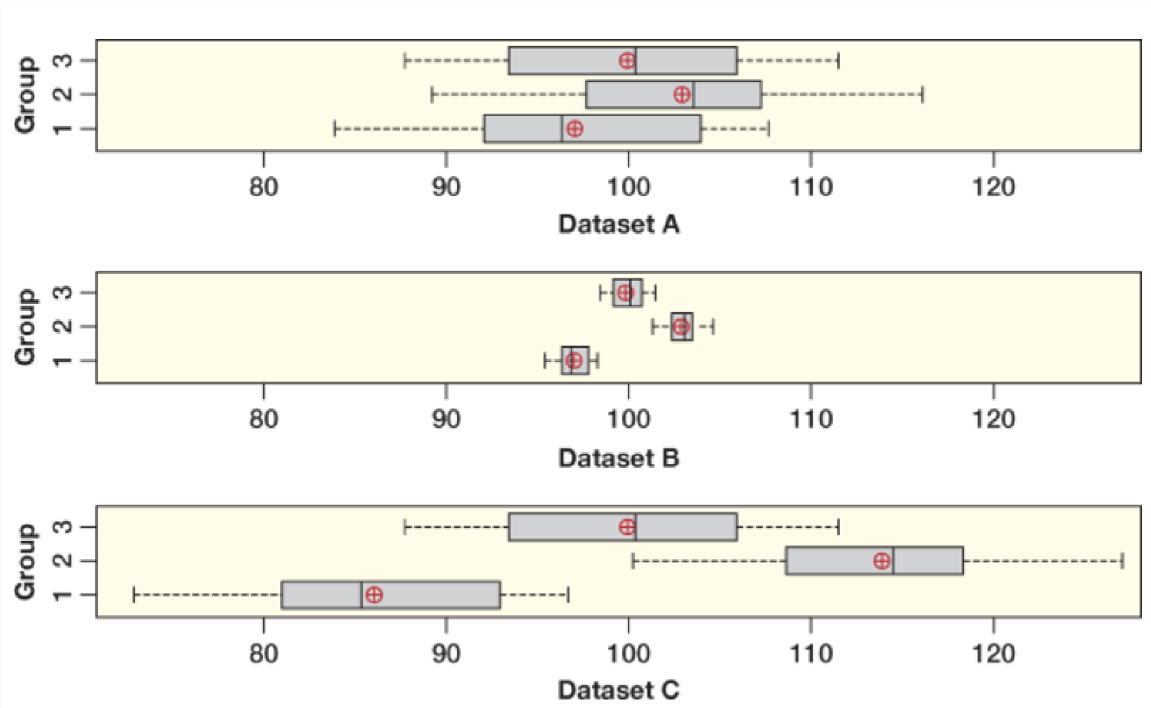
✨Datasets B and C = strong evidence for a difference in means.



Why Analyze Variability to Test for a Difference in Means?

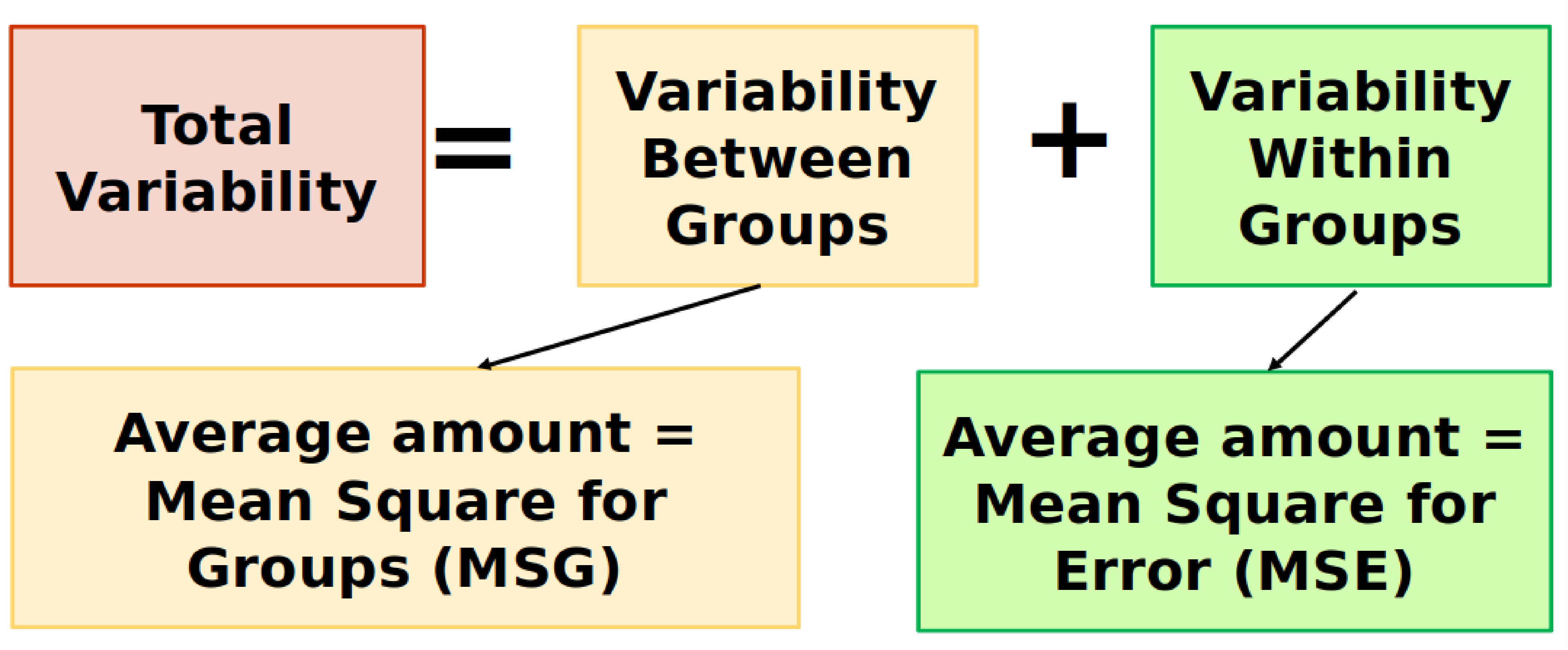
Conclusion: An assessment of the difference in means between several groups depends on two kinds of variability:

1. ***How different the means are between each groups***
2. ***The amount of variability within each groups***



Analysis of Variance

Analysis of Variance (ANOVA) compares the variability between groups to the variability within groups



F-Statistic

The F-statistic is a ratio:

$$F = \frac{MSG}{MSE} = \frac{\text{average between group variability}}{\text{average within group variability}}$$

If there really is a difference between the groups $(H_A \text{ true})$***, we would expect the F-statistic to be***

a) Large positive

b) Large negative

c) Close to 0

▶ Click for answer

Frisbee Example

```
frisbee <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Frisbee.csv")
frisbee.anova <- aov(Distance ~ Grip, data = frisbee) # fit an ANOVA model
```

```
summary(frisbee.anova)
            Df Sum Sq Mean Sq F value Pr(>F)
Grip         2  58.58   29.29   2.045  0.154
Residuals   21 300.75   14.32
```

F-test statistic: 2.045

P-value: 0.154

How to determine significance?

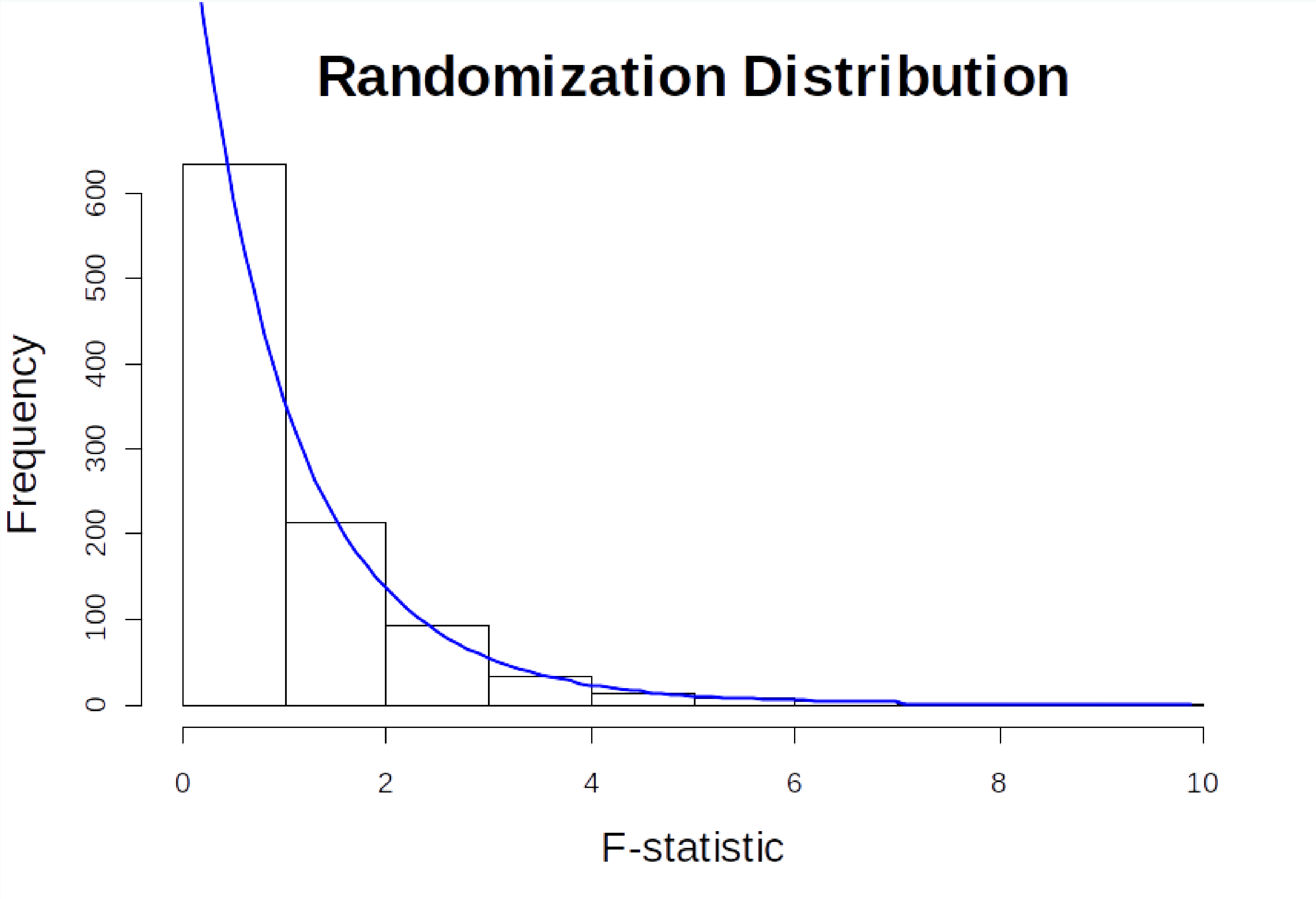
We have a test statistic. What else do we need to perform the hypothesis test?

A distribution of the test statistic assuming H_0 is true.

How do we get this? Two options:

1. ***Simulation***
2. ***Theory***

F-distribution



F-Distribution

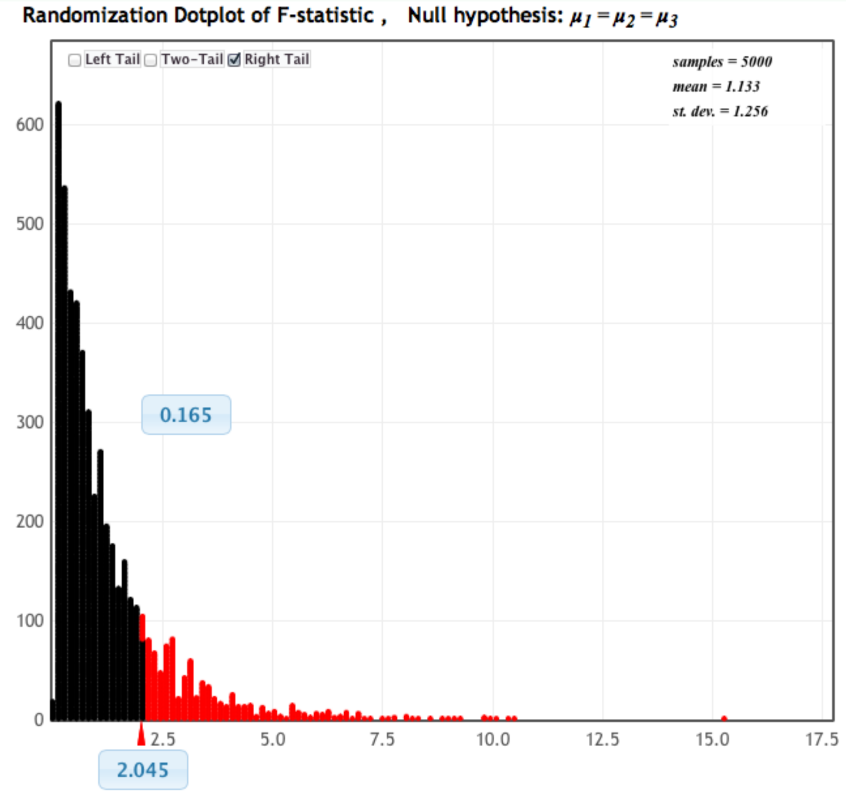
We can use the F-distribution to generate a p-value if:

1. ***Sample sizes in each group are large (each $n_i \geq 30$) OR the data within each group are relatively normally distributed***
2. ***Variability is similar in all groups***

✨*The F-distribution has two degrees of freedom, one for the numerator of the ratio $(k-1)$ and one for the denominator $(n-k)$*

✨*For F-statistics, the p-value (the area as extreme or more extreme) is always the right tail*

Simulation – how is this done?



Original Sample ANOVA Table

$n = 24$, $F = 2.045$

Statistics	Finger Out	Normal	Inverted	Overall
Sample Size	8	8	8	24
Mean	29.5	33.1	32.4	31.7
Standard Deviation	4.2	3.9	3.2	4.0

Randomization Sample ANOVA Table

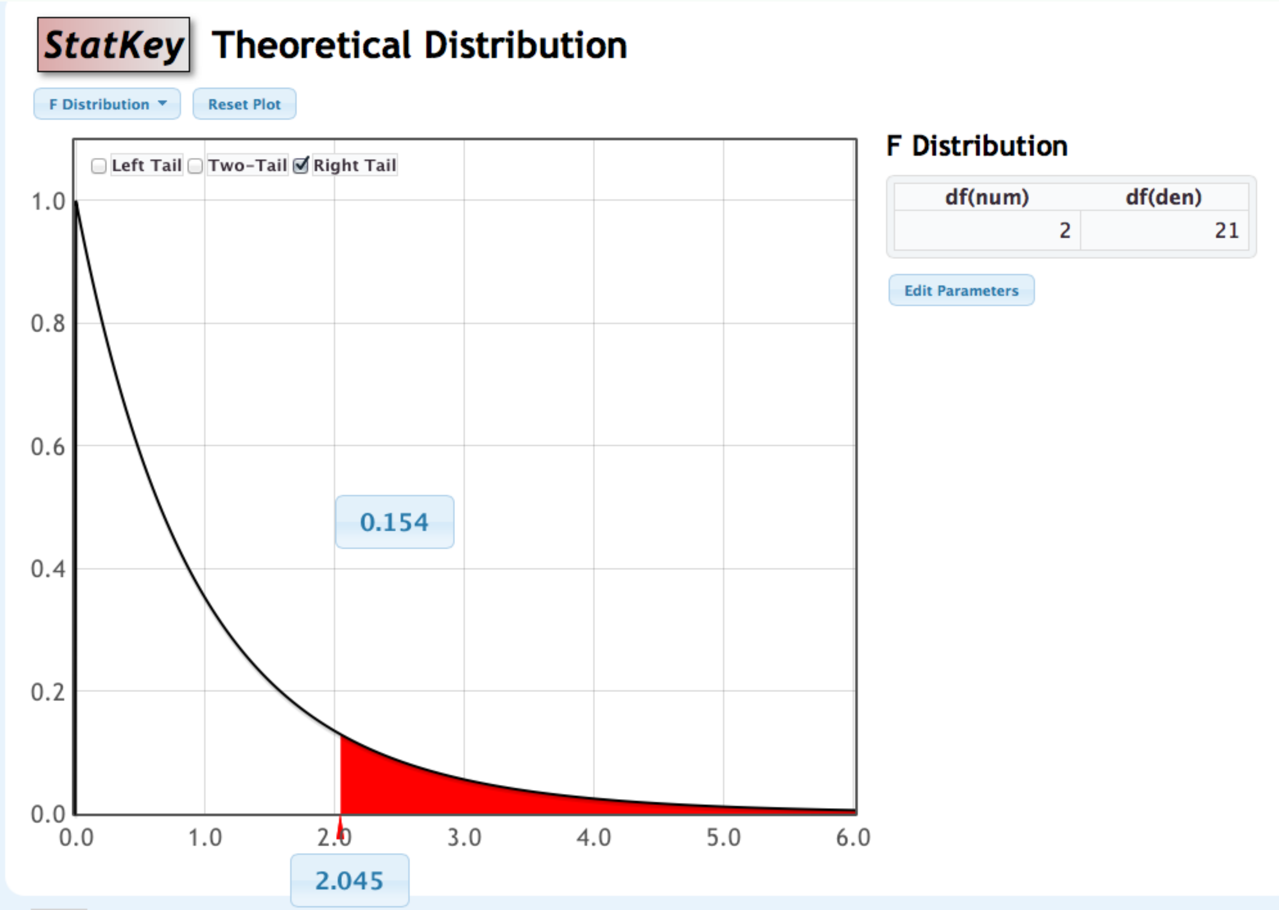
$n = 24$, $F = 2.544$

Statistics	Finger Out	Normal	Inverted	Overall
Sample Size	8	8	8	24
Mean	29.3	32.9	32.9	31.7
Standard Deviation	4.3	3.2	3.5	4.0

- ✨An F-statistic as large as 2.045 would occur by chance about 16% of the time if the means were all equal.
- ✨Our results are inconclusive and do not support the claim that grips affects average distance.

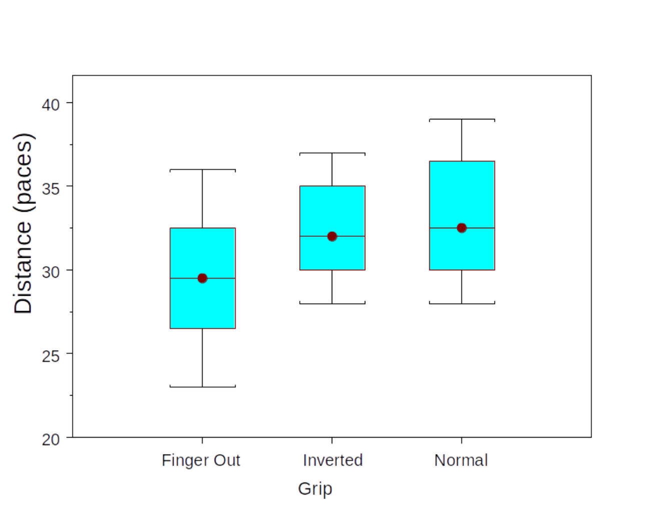
https://www.lock5stat.com/StatKey/

F-distribution




```
1 - pf(2.045,2,21)  # p-value
[1] 0.1543639
```

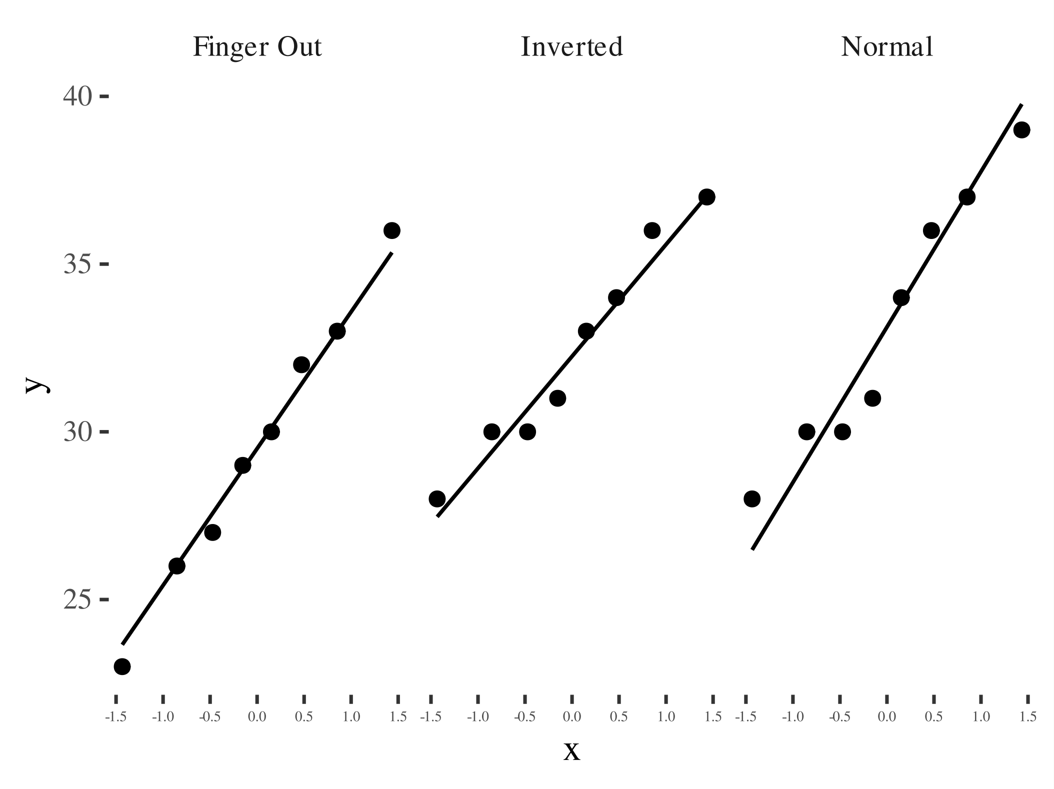
Check assumptions: large sample size (or normality)




```
# checking normality with qq-plots
ggplot(frisbee, aes(sample = Distance)) +
  geom_qq() + geom_qq_line() +  facet_wrap(~Grip) +
  theme(axis.text.x = element_text(size = 4))
```

```
table(frisbee$Grip)  # check n's

Finger Out   Inverted     Normal
         8          8          8
```



Small n_i but all groups are roughly normal

Check Assumptions: Equal Variance

The F-distribution assumes equal within group variability for each group. This is also an assumption when using the randomization distribution.

✨***As a rough rule of thumb, this assumption is violated if the largest group standard deviation is more than double the smallest group standard deviation***

```r
tapply(frisbee$Distance, frisbee$Grip, sd)
```

```
Finger Out   Inverted     Normal
  4.174754   3.159453   3.943802
```

Check if:

$$\frac{\text{largest } s}{\text{smallest } s} > 2$$

Frisbee Example: Inference

Question: Is this evidence that grip affects mean distance thrown?

$$
\begin{aligned}
H_0 &: \quad \mu_1 = \mu_2 = \mu_3 \\
H_a &: \quad \text{At least one } \mu_1, \mu_2, \mu_3 \text{ is not the same}
\end{aligned}
$$

μ_{i} is the true mean distance thrown using grip i.

$$F = 2.05(\mathrm{df} = 2, 21),\ \text{P-value } = 0.1543$$

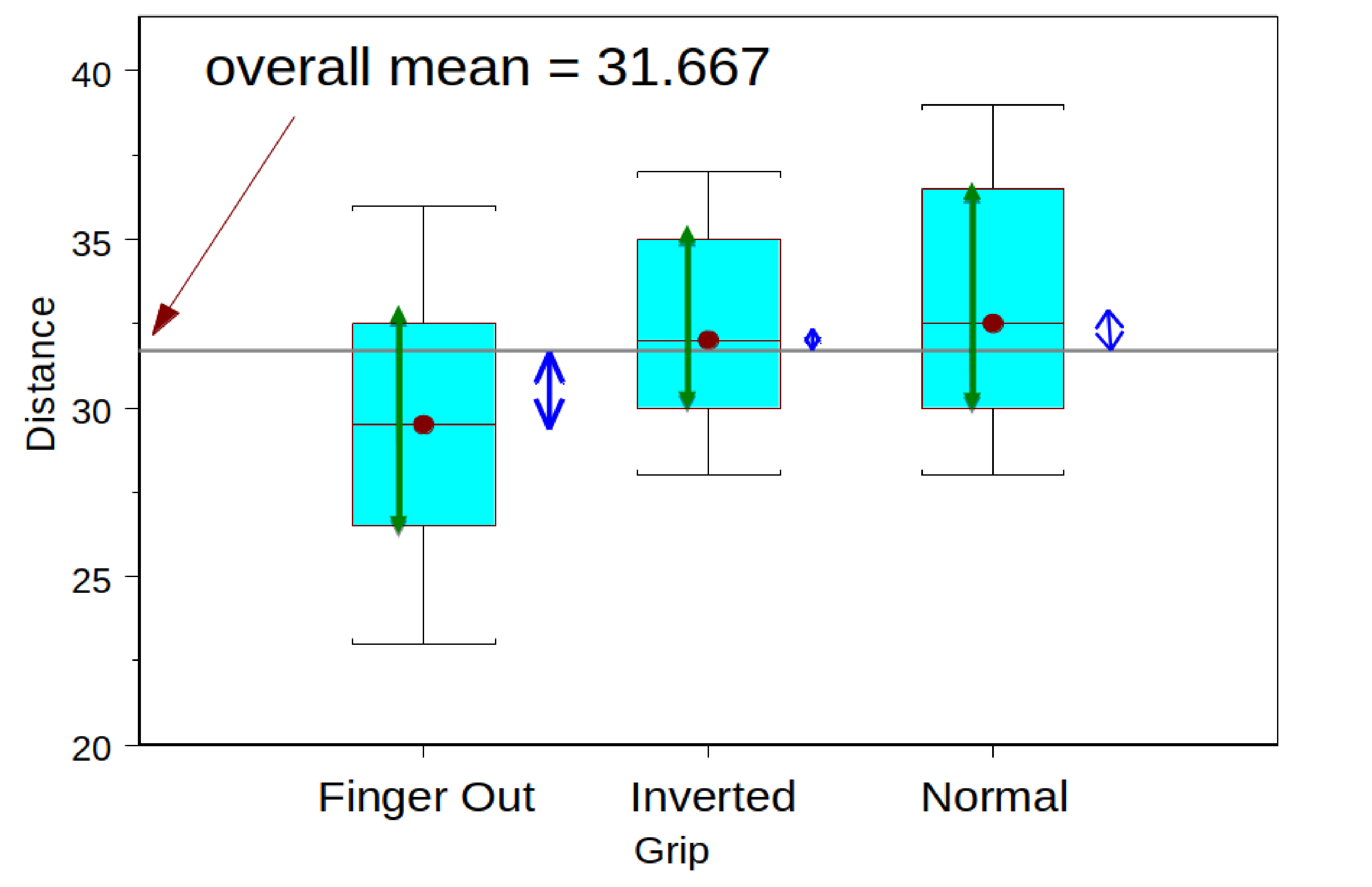
Conclusion: Do not reject the Null hypothesis. The difference in observed means is not statistically significant.

About 15% of the time we would see the grip differences like those observed, or even bigger, when there is actually no difference between the true mean distances thrown with different grips.

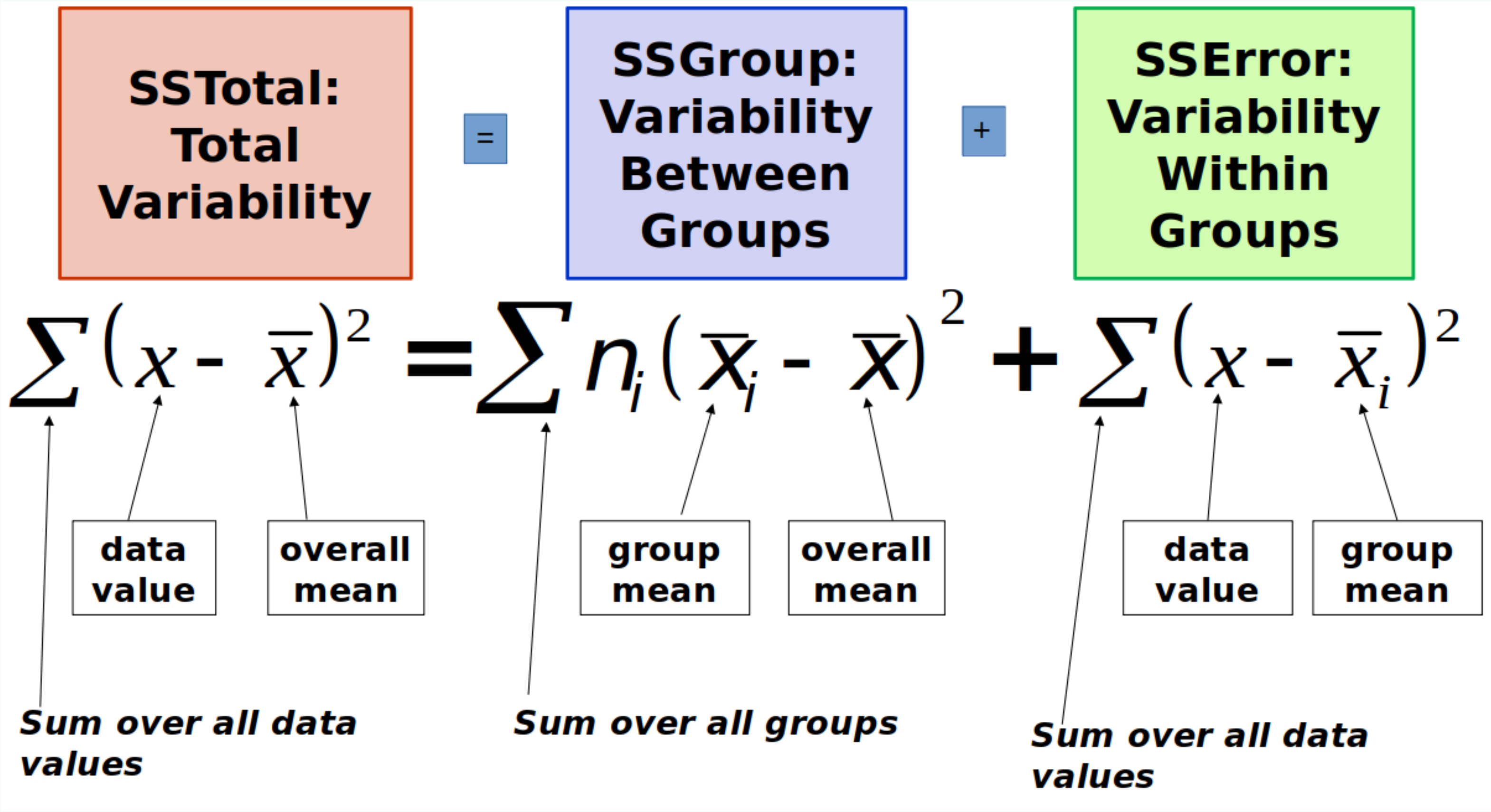
Picturing the variation

Green: Variation within groups

Blue: Variation between groups



Sums of Squares



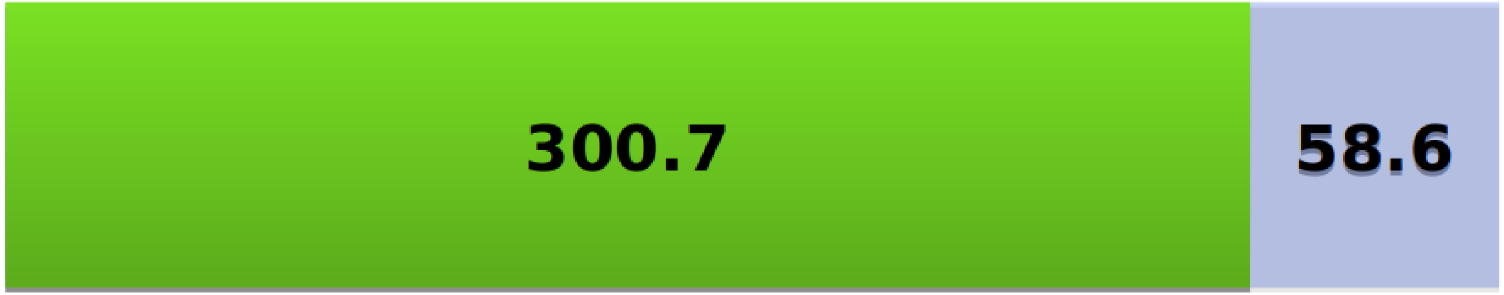
Sums of Squares: Frisbee Data

$$SSG = \sum n_i (\bar{x}_i - \bar{x})^2 = 58.58333$$

$$SSE = \sum (x - \bar{x}_i)^2 = 300.7500$$

$$SSTotal = \sum (x - \bar{x})^2 = 359.3333$$

$$= s^2 (n - 1) = 3.952617^2 (24 - 1)$$



ANOVA Table for Frisbee data

$$\text{F test stat} = 29.29/14.32 = 2.045$$

Source	df	Sum of Squares	Mean Square
Groups	#groups -1 **3-1 = 2**	SSG **58.583**	SSG/df **58.583/2 = 29.29**
Error (residual)	n - #groups **24-3 = 21**	SSE **300.750**	SSE/df **300.75/21= 14.32**
Total	n-1 **24-1 = 23**	SSTotal **359.333**	

Frisbee Example: ANOVA table in R

```
frisbee.anova <- aov(Distance ~ Grip, data = frisbee)
```

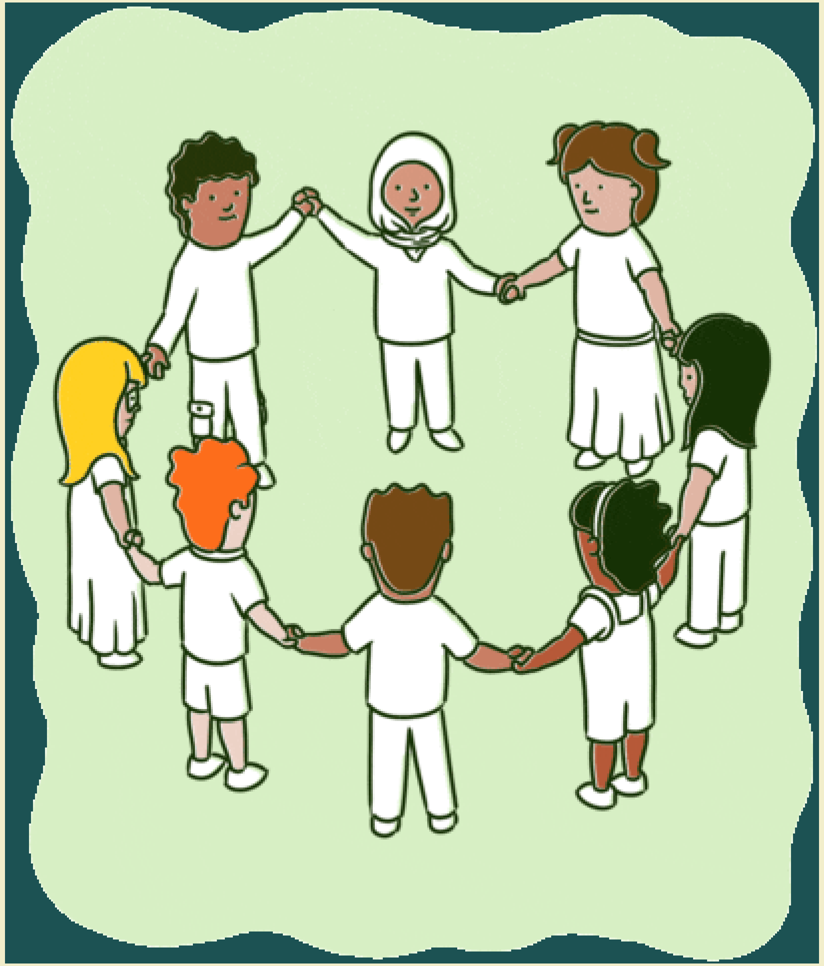
```
summary(frisbee.anova)
            Df Sum Sq Mean Sq F value Pr(>F)
Grip         2  58.58   29.29   2.045  0.154
Residuals   21 300.75   14.32
```

```
library(broom)
knitr::kable(tidy(frisbee.anova)) # nicer summary tables
```

term	df	sumsq	meansq	statistic	p.value
Grip	2	58.58333	29.29167	2.045303	0.1543247
Residuals	21	300.75000	14.32143	NA	NA

Your Turn 1

05:00



✨Go over to the in class activity file

✨Complete the activity as much as possible