

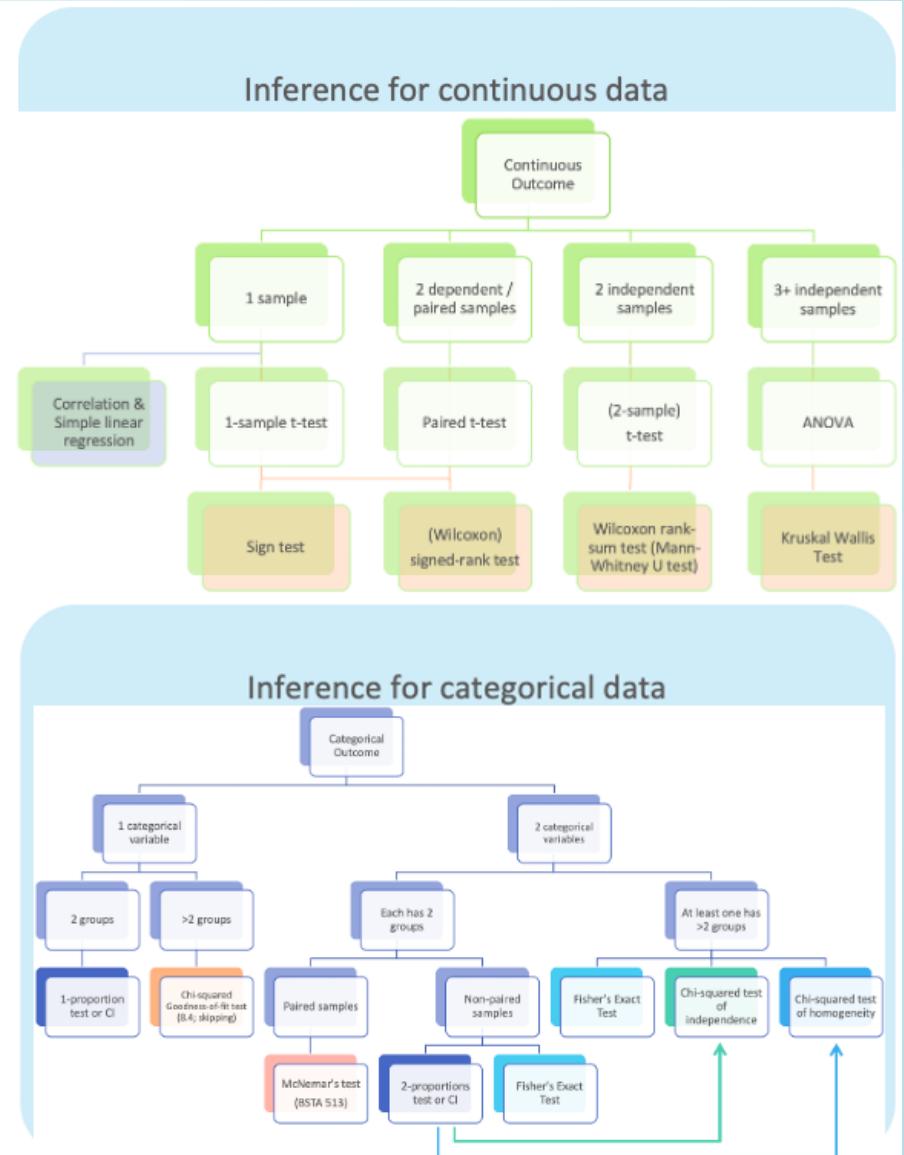
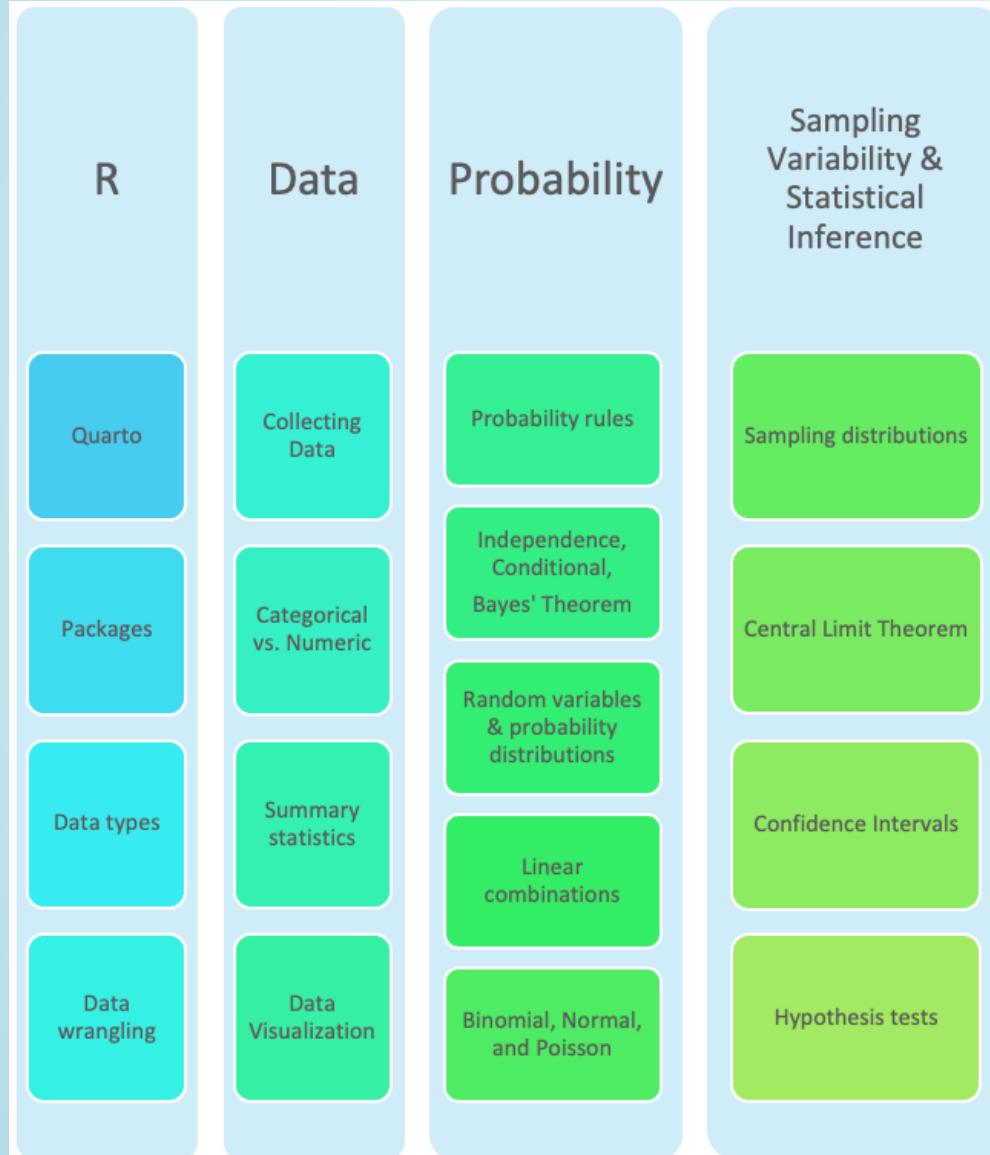
Day 14: Comparing Means with ANOVA (Section 5.5)

BSTA 511/611

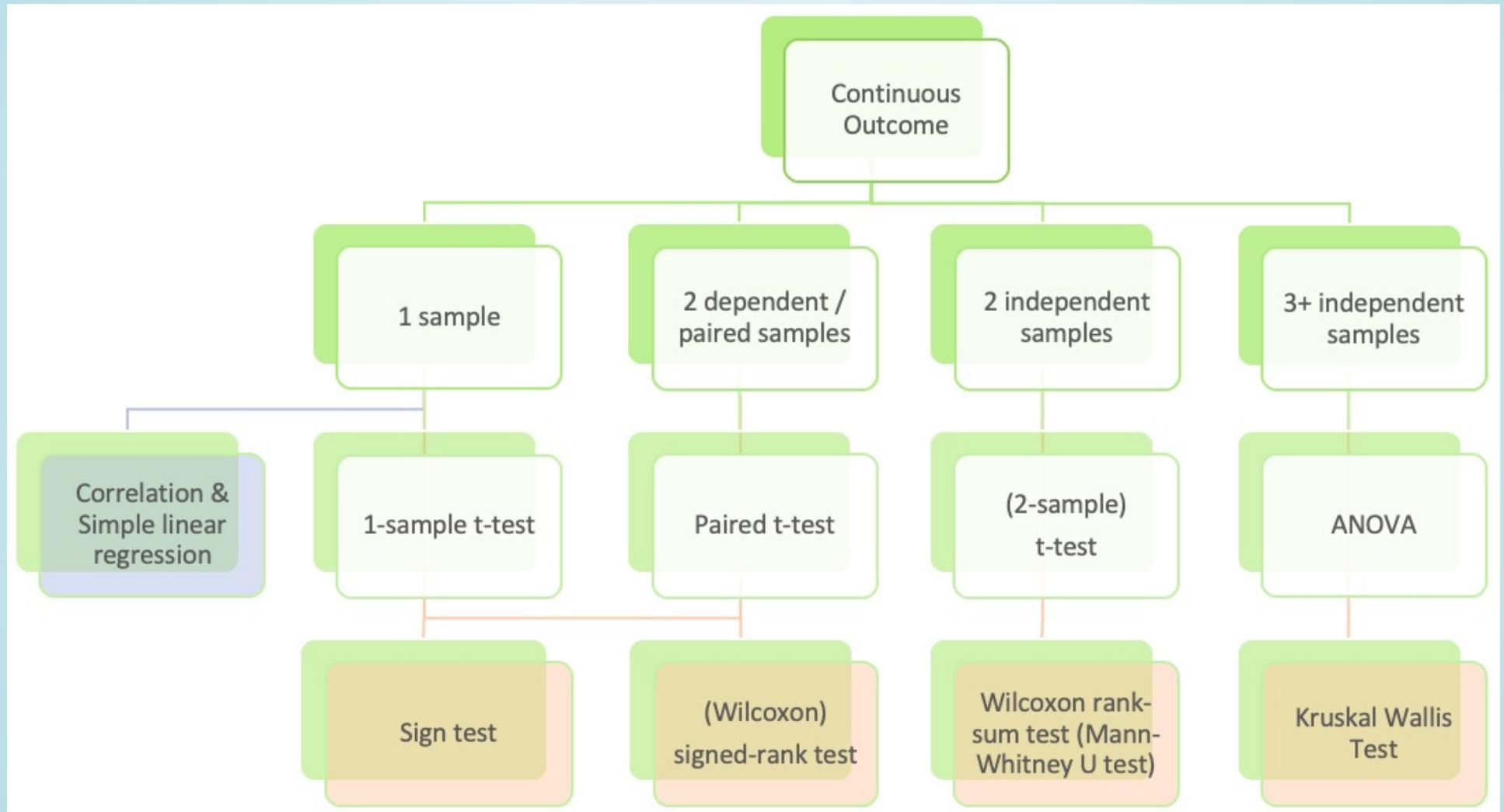
Meike Niederhausen, PhD
OHSU-PSU School of Public Health

2023-11-20

Where are we?



Where are we? Continuous outcome zoomed in



Goals for today (Section 5.5)

- Analysis of Variance (ANOVA)
- When to use an ANOVA
- Hypotheses
- ANOVA table
- Different sources of variation in ANOVA
- ANOVA conditions
- F-distribution
- Post-hoc testing of differences in means
- Running an ANOVA in R

Disability Discrimination Example

- The U.S. Rehabilitation Act of 1973 prohibited discrimination against people with physical disabilities.
 - The act defined a disabled person as any individual who has a physical or mental impairment that limits the person's major life activities.
- A 1980's study examined whether physical disabilities affect people's perceptions of employment qualifications
 - (Cesare, Tannenbaum, & Dalessio, 1990).
- Researchers prepared recorded job interviews, using *same actors and script each time*.
- Only difference: job applicant appeared with different disabilities.
 - *No disability*
 - *Leg amputation*
 - *Crutches*
 - *Hearing impairment*
 - *Wheelchair confinement*
- 70 undergrad students were randomly assigned to view one of the videotapes,
 - then **rated** the candidate's qualifications on a **1-10 scale**.
- The research question: **are qualifications evaluated differently depending on the applicant's presented disability?**

Load interview data from `.txt` file

- `.txt` (text) files are usually tab-delimited files
 - `.csv` files are comma-separated files
- `read_delim` is from the `readr` package, just like `read_csv`, and loads with other `tidyverse` packages

```
1 employ <- read_delim(  
2   file = here::here("data", "DisabilityEmployment.txt"),  
3   delim = "\t",    # tab delimited  
4   trim_ws = TRUE)
```

`trim_ws`: specify whether leading and trailing white space should be trimmed from each field before parsing it

```
1 glimpse(employ)
```

```
Rows: 70  
Columns: 2  
$ disability <chr> "none", "none", "none", "none", "none", "none", "no...  
$ score      <dbl> 1.9, 2.5, 3.0, 3.6, 4.1, 4.2, 4.9, 5.1, 5.4, 5.9, 6.1, 6.7,...
```

```
1 summary(employ)
```

```
disability          score  
Length:70           Min.   :1.400  
Class :character    1st Qu.:3.700  
Mode  :character    Median :5.050  
                  Mean   :4.929  
                  3rd Qu.:6.100  
                  Max.   :8.500
```

```
1 employ %>% tabyl(disability)
```

```
disability n percent  
amputee   14     0.2  
crutches  14     0.2  
hearing   14     0.2  
none      14     0.2  
wheelchair 14     0.2
```

MoRitz's tip of the day

Read [OHSU's Inclusive Language Guide](#) (below is from pgs. 22-25)

"... an evolving tool to help OHSU members learn about and use inclusive language..."

Sections on: Race and ethnicity, Immigration status, Gender and sexual orientation, and Ability (including physical, mental and chronological attributes)



Ability, physical, mental and chronological attributes

Following is a glossary promoting language around ability and physical, mental and chronological attributes. The Community of People with Disabilities is by definition inclusive and intersectional. At the request of OHSU members, we have also added a segment on body weight and age.

RESPECTFUL LANGUAGE	
TERM	DEFINITION
Person with a disability/people with disabilities	This represents person-first language; see the person, not the disability. Widely, <i>but not universally used</i> in the community for people with disabilities. For example, Deaf people and autistic (neurodiverse people) prefer the respective adjectives to proceed the word people. People with disabilities are not all the same.

TERMS TO AVOID			
ABLEIST LANGUAGE			
Amp/amputee	Handicapped	The Spectrum/on the Spectrum	
Cripple, crippled	Invalid	Wheelchair-bound, or confined to a wheelchair (wheelchairs are mobility tools, and people are not stuck in them)	
Diabetic	Lame		
Gimp	Spaz	Hearing impaired is a less favored term in the deaf/hard-of-hearing community as the word impaired can have negative connotations and focuses on what a person can't do.	
SANIST LANGUAGE			
Addict, addicted	Drug baby	Invalid	Opioid addict
Bipolar	Handicapped	Lunatic	
Crazy	Idiot	Manic	Retarded and variants including words with prefixes attached to -tard.
Deranged	Imbecile	Maniac	
Drug addict	Insane	Nuts	Weird

Factor variable: Make **disability** a factor variable

```
1 glimpse(employ)
```

```
Rows: 70  
Columns: 2  
$ disability <chr> "none", "none", "none", "none", "none", "none", "no...  
$ score      <dbl> 1.9, 2.5, 3.0, 3.6, 4.1, 4.2, 4.9, 5.1, 5.4, 5.9, 6.1, 6.7,...
```

```
1 summary(employ)
```

	disability	score
Length:	70	Min. :1.400
Class :	character	1st Qu.:3.700
Mode :	character	Median :5.050
		Mean :4.929
		3rd Qu.:6.100
		Max. :8.500

Make **disability** a factor variable:

```
1 employ <- employ %>%  
2   mutate(disability = factor(disability))
```

What's different now?

```
1 glimpse(employ)
```

```
Rows: 70  
Columns: 2  
$ disability <fct> none, none, none, none, none, none, none, none, none,...  
$ score      <dbl> 1.9, 2.5, 3.0, 3.6, 4.1, 4.2, 4.9, 5.1, 5.4, 5.9, 6.1, 6.7,...
```

```
1 summary(employ)
```

	disability	score
amputee	:14	Min. :1.400
crutches	:14	1st Qu.:3.700
hearing	:14	Median :5.050
none	:14	Mean :4.929
wheelchair	:14	3rd Qu.:6.100
		Max. :8.500

Factor variable: Change order & name of disability levels

What are the current level names and order?

```
1 levels(employ$disability)
[1] "amputee"     "crutches"    "hearing"      "none"        "wheelchair"
```

What changes are being made below?

```
1 employ <- employ %>%
2   mutate(
3     # make "none" the first level
4     # by only listing the level none, all other levels will be in original order
5     disability = fct_relevel(disability, "none"),
6     # change the level name amputee to amputation
7     disability = fct_recode(disability, amputation = "amputee")
8   )
```

- `fct_relevel()` and `fct_recode()` are from the `forcats` package:
<https://forcats.tidyverse.org/index.html>.
- `forcats` is loaded with `library(tidyverse)`.

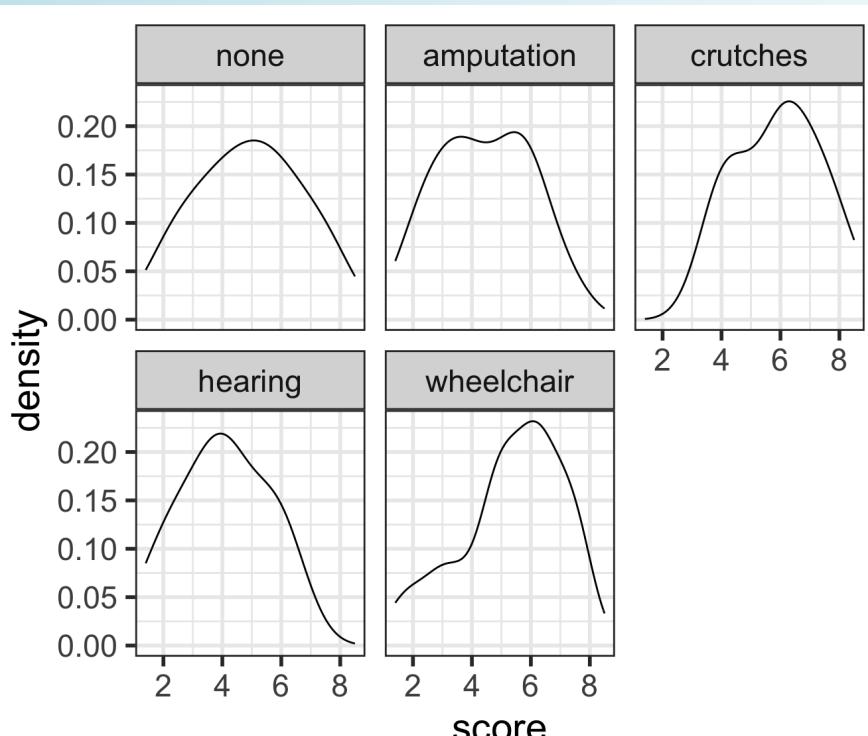
New order & names:

```
1 levels(employ$disability) # note the new order and new name
[1] "none"        "amputation"  "crutches"    "hearing"      "wheelchair"
```

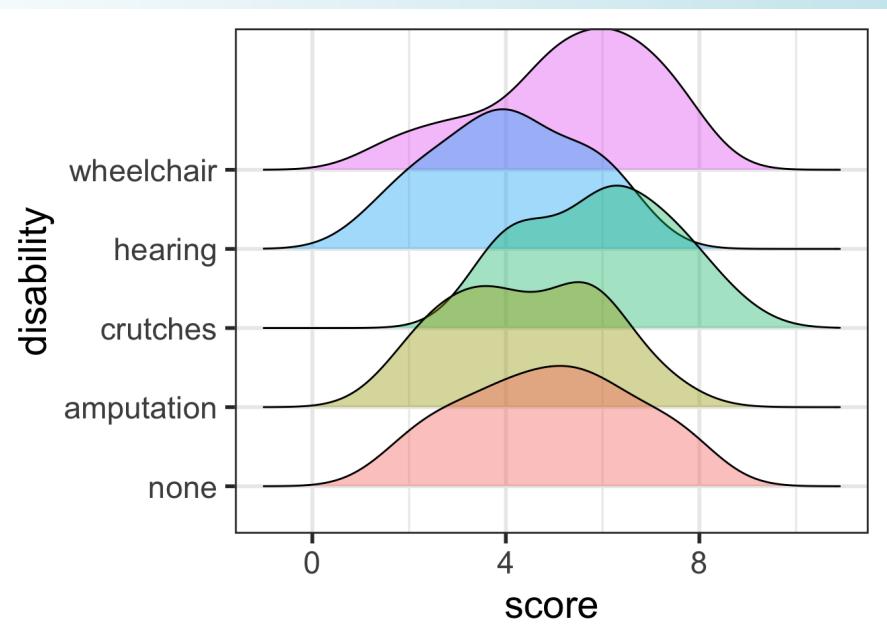
Data viz (1/2)

- What are the **score** distribution shapes within each group?
- Any unusual values?

```
1 ggplot(employ, aes(x=score)) +  
2   geom_density() +  
3   facet_wrap(~ disability)
```



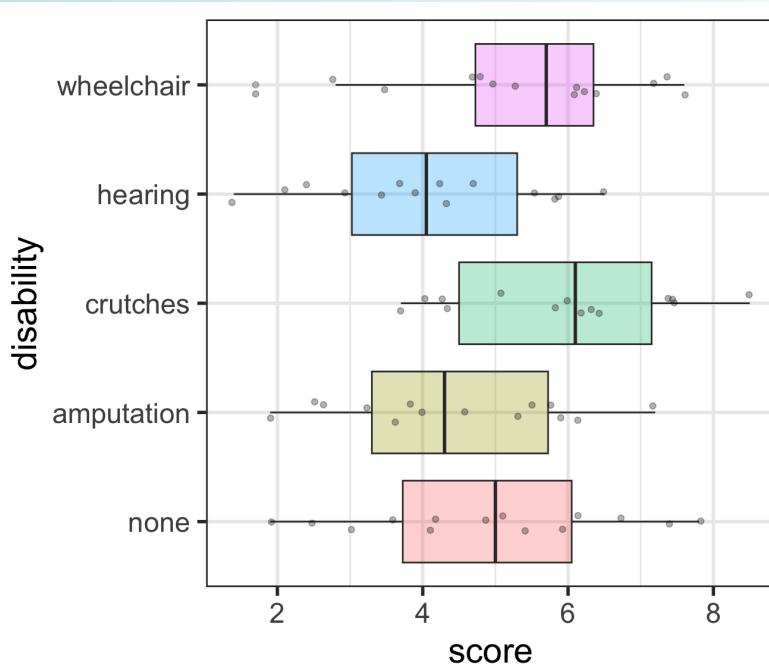
```
1 library(gggridges)  
2 ggplot(employ,  
3         aes(x=score,  
4                  y = disability,  
5                  fill = disability)) +  
6   geom_density_ridges(alpha = 0.4) +  
7   theme(legend.position="none")
```



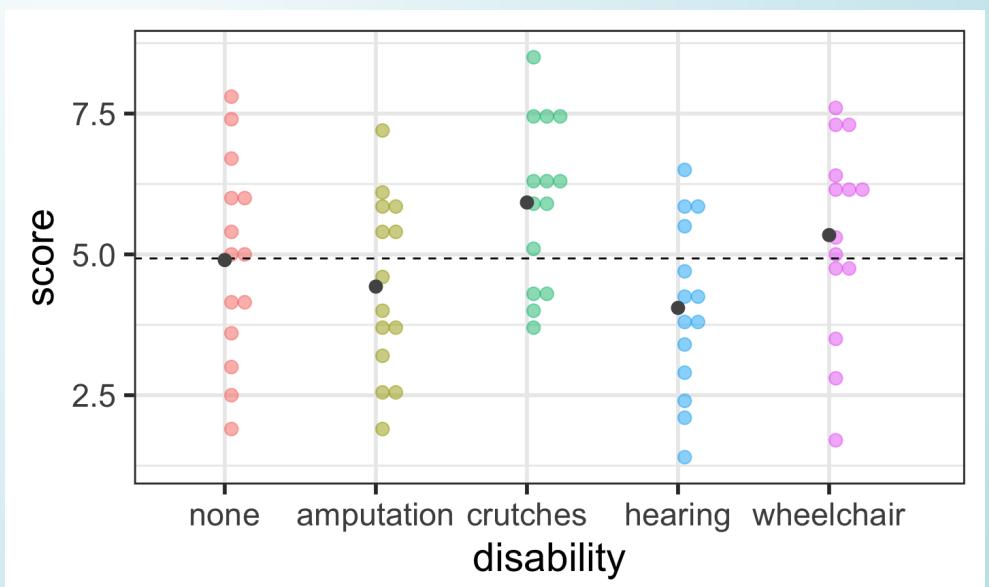
Data viz (2/2)

- Compare the **score** measures of **center** and **spread** between the groups

```
1 ggplot(employ,
2   aes(y=score,
3     x = disability,
4     fill = disability)) +
5   geom_boxplot(alpha = 0.3) +
6   coord_flip() +
7   geom_jitter(width = 0.1,
8     alpha = 0.3) +
9   theme(legend.position = "none")
```



```
1 ggplot(employ,
2   aes(x = disability,
3     y=score,
4     fill=disability,
5     color=disability)) +
6   geom_dotplot(binaxis = "y", alpha = 0.5) +
7   geom_hline(aes(yintercept = mean(score)),
8     lty = "dashed") +
9   stat_summary(fun ="mean", geom="point",
10   size = 3, color = "grey33", alpha = 1) +
11  theme(legend.position = "none")
```



Hypotheses

To test for a difference in means across k groups:

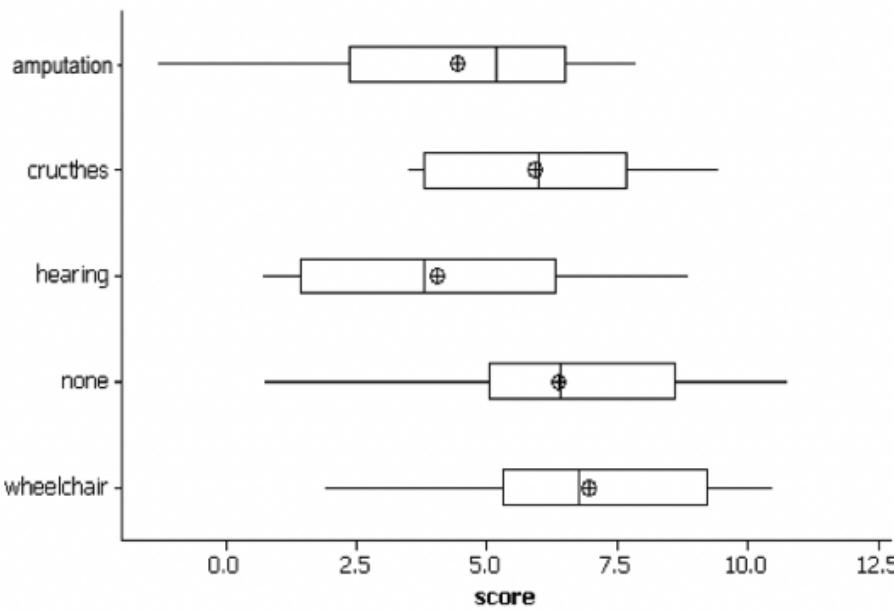
$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

vs. $H_A : \text{At least one pair } \mu_i \neq \mu_j \text{ for } i \neq j$

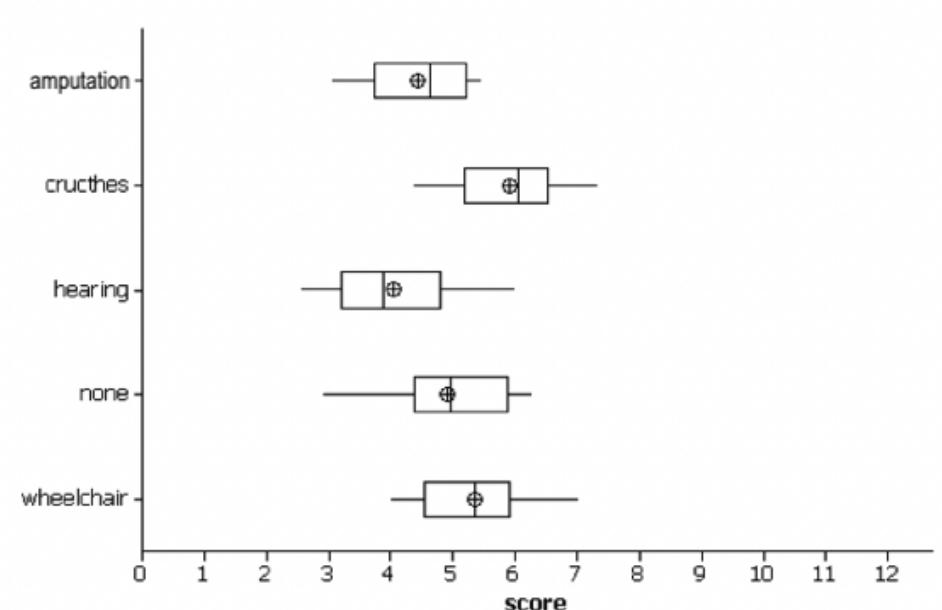
Hypothetical examples:

In which set (A or B) do you believe the evidence will be stronger that at least one population differs from the others?

A:



B:



Comparing means

Whether or not two means are significantly different depends on:

- How far apart the **means** are
- How much **variability** there is within each group

Questions:

- How to measure variability **between** groups?
- How to measure variability **within** groups?
- How to compare the two measures of variability?
- How to determine significance?

ANOVA in base R

- There are several options to run an ANOVA model in R
- Two most common are `lm` and `aov`
 - `lm` = linear model; will be using frequently in BSTA 512

```
1 lm(score ~ disability, data = employ) %>% anova()
```

Analysis of Variance Table

Response: score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
disability	4	30.521	7.6304	2.8616	0.03013 *
Residuals	65	173.321	2.6665		

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

```
1 aov(score ~ disability, data = employ) %>% summary()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
disability	4	30.52	7.630	2.862	0.0301 *
Residuals	65	173.32	2.666		

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

Hypotheses:

$$H_0 : \mu_{none} = \mu_{amputation} = \mu_{crutches} = \mu_{hearing} = \mu_{wheelchair}$$

vs. $H_A : \text{At least one pair } \mu_i \neq \mu_j \text{ for } i \neq j$

Do we reject or fail to reject H_0 ?

ANOVA tables

Disability example ANOVA table from R:

```
1 lm(score ~ disability, data = employ) %>% anova()
```

Analysis of Variance Table

Response: score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
disability	4	30.521	7.6304	2.8616	0.03013 *
Residuals	65	173.321	2.6665		

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

Generic ANOVA table:

The “mean square” is the sum of squares divided by the degrees of freedom				
Source	df	Sum of Squares	Mean Square	F-Statistic
Groups	$k-1$	SSG	$MSG = \frac{SSG}{k-1}$	$\frac{MSG}{MSE}$
Error	$N-k$	SSE	$MSE = \frac{SSE}{N-k}$	
Total	$N-1$	SST		

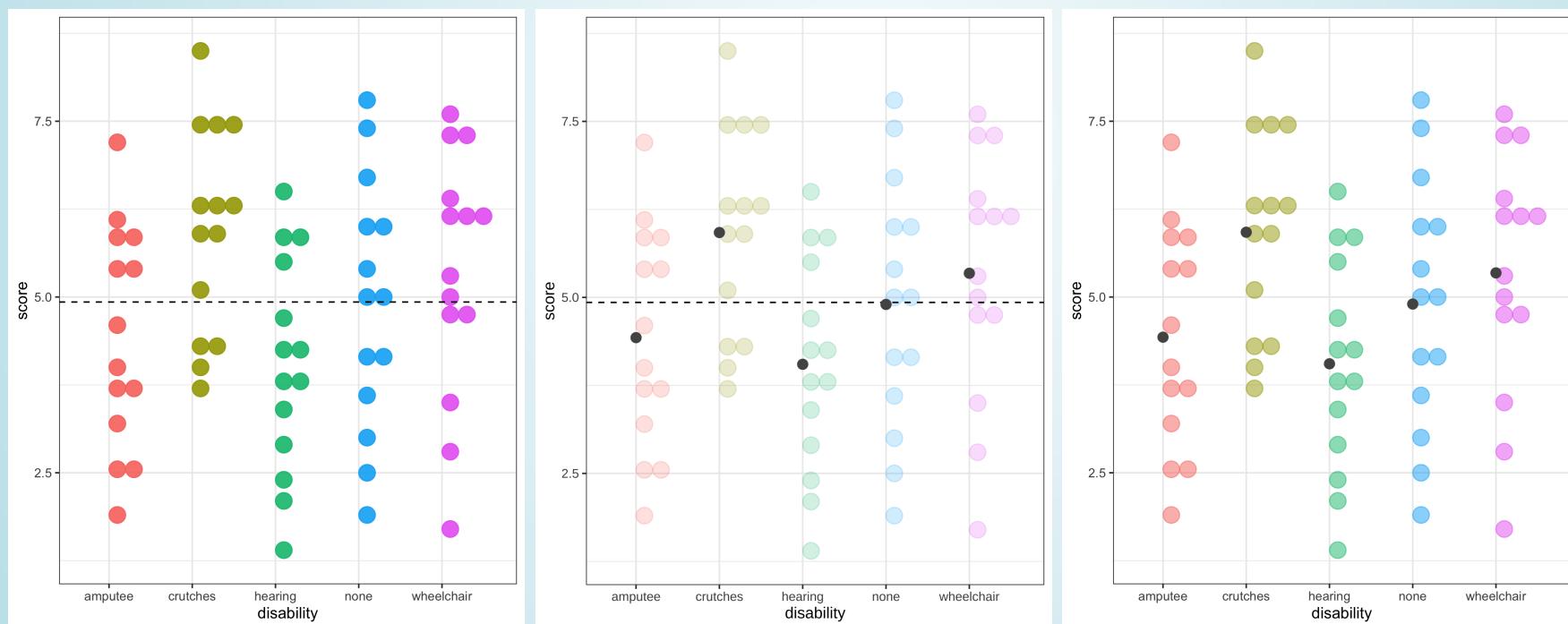
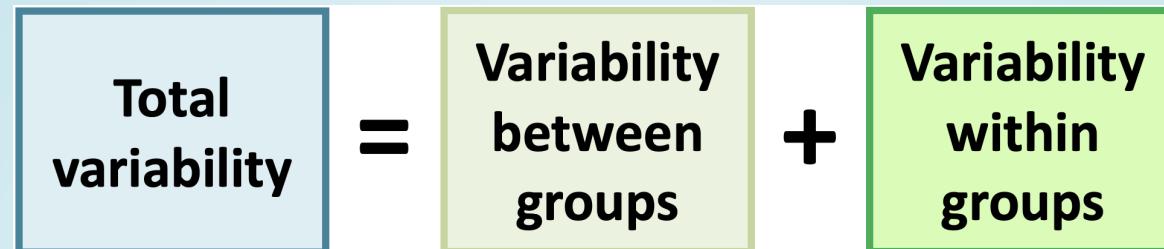
↑
variability

↑
average variability

The **F-statistic** is a ratio of the average variability **between** groups to the average variability **within** groups

ANOVA: Analysis of Variance

ANOVA compares the variability between groups to the variability within groups



ANOVA: Analysis of Variance

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

Total
variability

=

Variability
between
groups

+

Variability
within
groups

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

SST
(Total sum
of squares)

=

SSG
(sum of squares
due to groups)

+

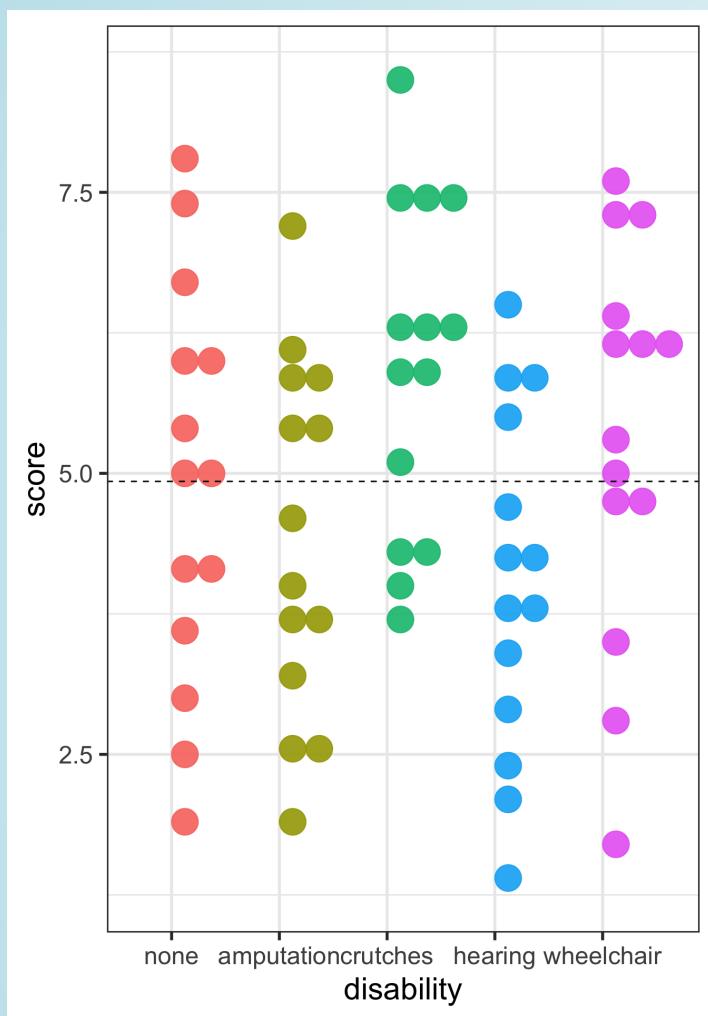
SSE
(Error sum
of squares)

Notation

- k groups
- n_i observations in each of the k groups
- Total sample size is $N = \sum_{i=1}^k n_i$
- \bar{x}_i = mean of observations in group i
- \bar{x} = mean of *all* observations
- s_i = sd of observations in group i
- s = sd of *all* observations

Observation	$i = 1$	$i = 2$	$i = 3$...	$i = k$	overall
$j = 1$	x_{11}	x_{21}	x_{31}	...	x_{k1}	
$j = 2$	x_{12}	x_{22}	x_{32}	...	x_{k2}	
$j = 3$	x_{13}	x_{23}	x_{33}	...	x_{k3}	
$j = 4$	x_{14}	x_{24}	x_{34}	...	x_{k4}	
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots	
$j = n_i$	x_{1n_1}	x_{2n_2}	x_{3n_3}	...	x_{kn_k}	
Means	\bar{x}_1	\bar{x}_2	\bar{x}_3	...	\bar{x}_k	\bar{x}
Variance	s_1^2	s_2^2	s_3^2	...	s_k^2	s^2

Total Sums of Squares Visually



Total Sums of Squares:

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = (N - 1)s^2$$

- where
 - $N = \sum_{i=1}^k n_i$ is the total sample size and
 - s^2 is the grand standard deviation of all the observations
- This is the sum of the squared differences between each observed x_{ij} value and the *grand mean*, \bar{x} .
- That is, it is the total deviation of the x_{ij} 's from the grand mean.

Calculate Total Sums of Squares

Total Sums of Squares:

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = (N - 1)s^2$$

- where

- $N = \sum_{i=1}^k n_i$ is the total sample size and
- s^2 is the grand standard deviation of all the observations

Total sample size N :

```
1 (Ns <- employ %>% group_by(disability) %>% count())
# A tibble: 5 × 2
# Groups:   disability [5]
  disability     n
  <fct>       <int>
1 none          14
2 amputation    14
3 crutches      14
4 hearing        14
5 wheelchair    14
```

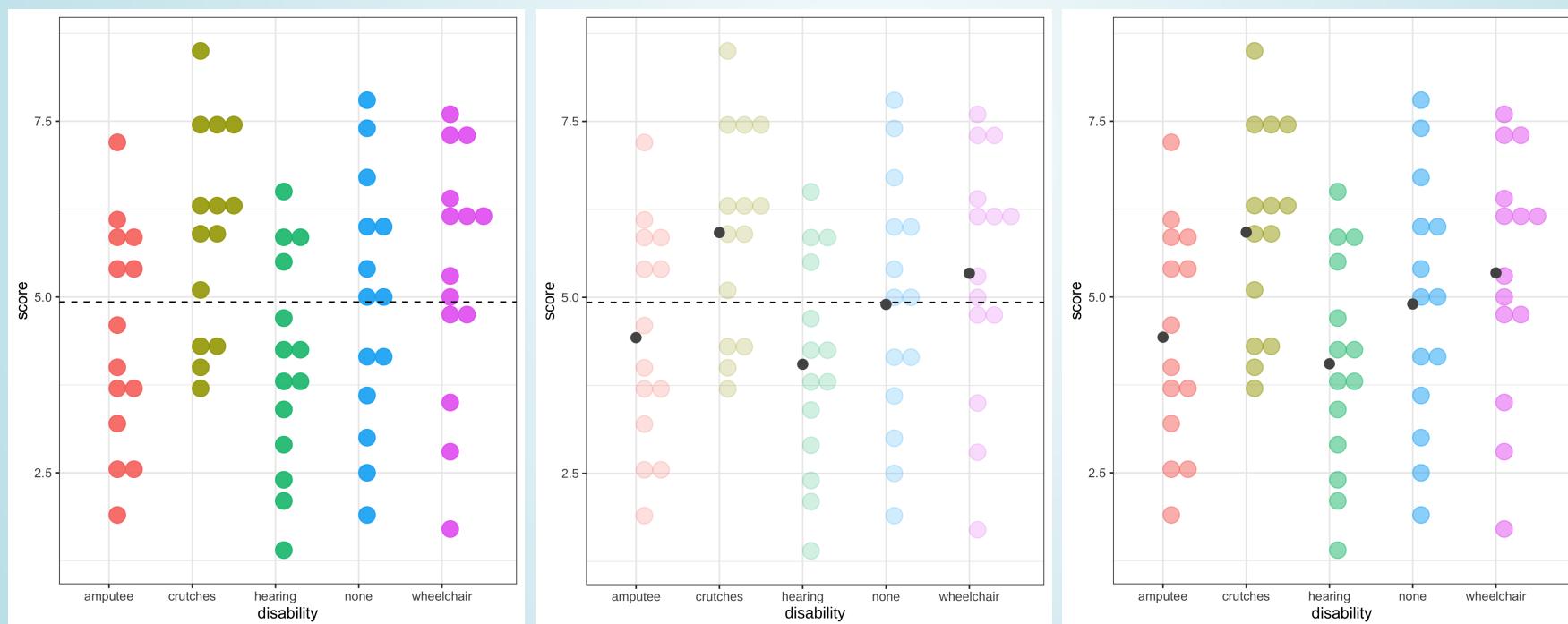
SST :

```
1 (SST <- (sum(Ns$n) - 1) * sd(employ$score)^2)
[1] 203.8429
```

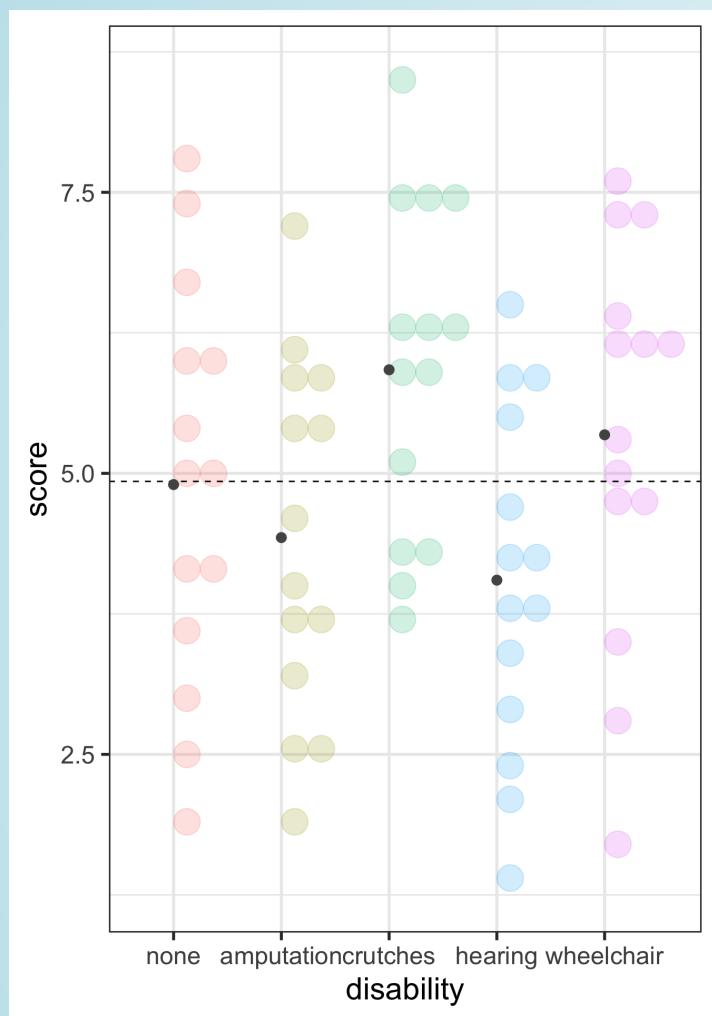
ANOVA: Analysis of Variance

ANOVA compares the variability between groups to the variability within groups

$$\text{Total variability} = \text{Variability between groups} + \text{Variability within groups}$$



Sums of Squares due to Groups Visually (“between” groups)



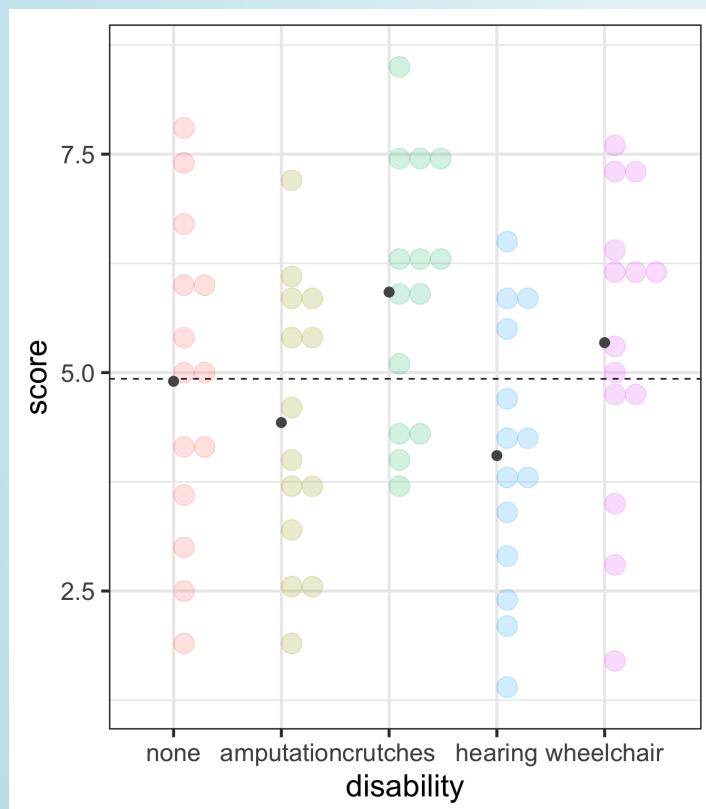
Sums of Squares due to Groups:

$$SSG = \sum_{i=1}^k n_i(\bar{x}_i - \bar{x})^2$$

- This is the sum of the squared differences between each *group* mean, \bar{x}_i , and the *grand mean*, \bar{x} .
- That is, it is the deviation of the group means from the grand mean.
- Also called the Model SS, or SS_{model} .

Calculate Sums of Squares due to Groups (“between” groups)

$$SSG = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$$



Calculate means \bar{x}_i for each group:

```
1 xbar_groups <- employ %>%
 2   group_by(disability) %>%
 3   summarise(mean = mean(score))
 4 xbar_groups

# A tibble: 5 × 2
  disability     mean
  <fct>       <dbl>
1 none         4.9 
2 amputation  4.43
3 crutches    5.92
4 hearing     4.05
5 wheelchair  5.34
```

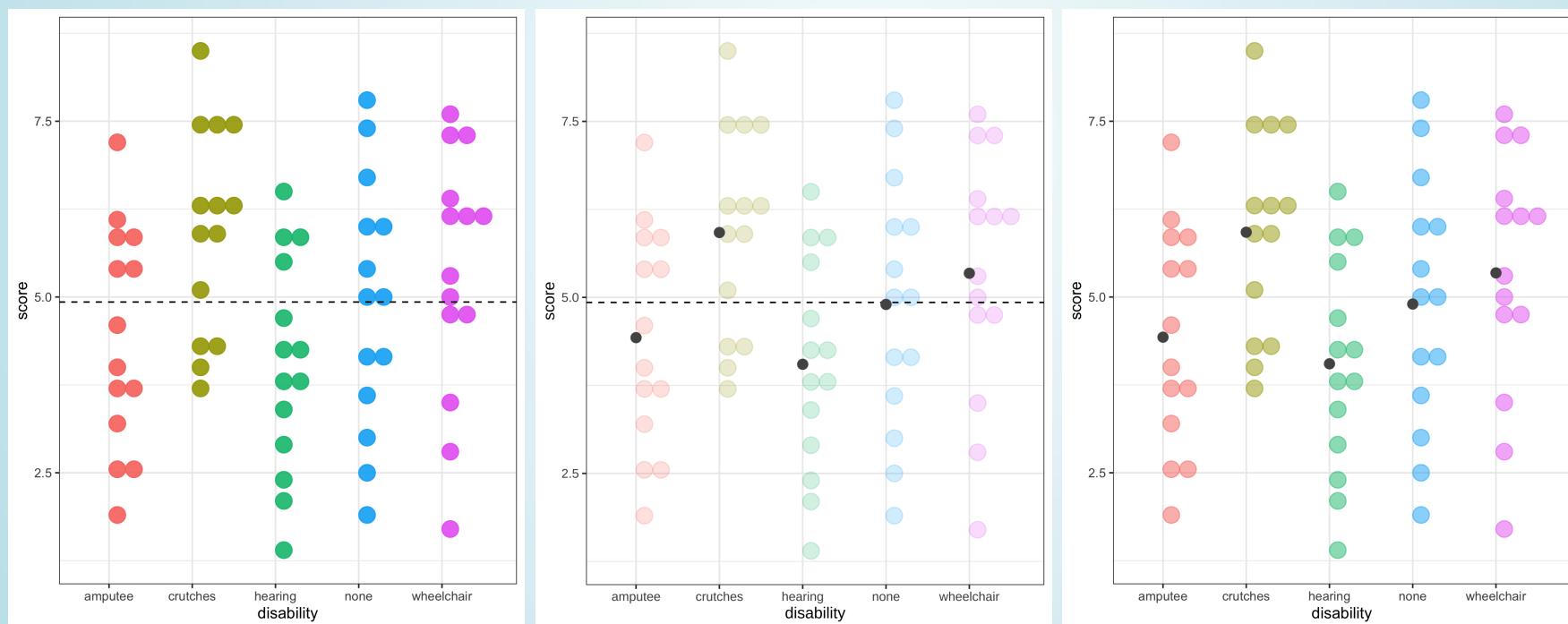
Calculate SSG :

```
1 (SSG <- sum(Ns$n *
 2   (xbar_groups$mean - mean(employ$score))^2))
[1] 30.52143
```

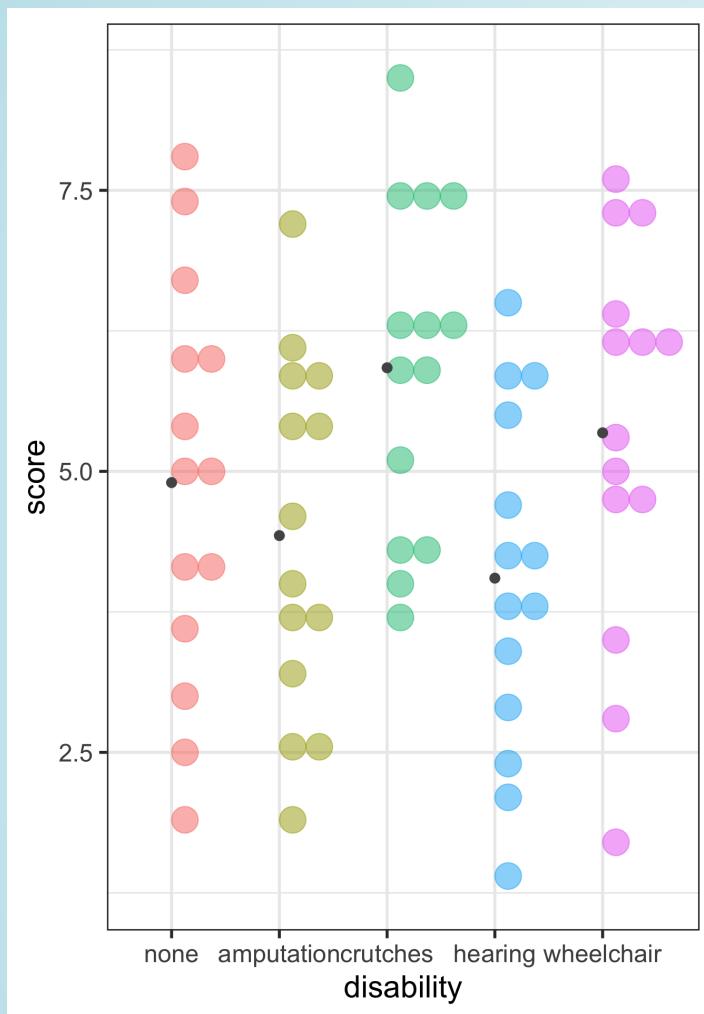
ANOVA: Analysis of Variance

ANOVA compares the variability between groups to the variability within groups

$$\text{Total variability} = \text{Variability between groups} + \text{Variability within groups}$$



Sums of Squares Error Visually (within groups)



Sums of Squares Error:

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^k (n_i - 1)s_i^2$$

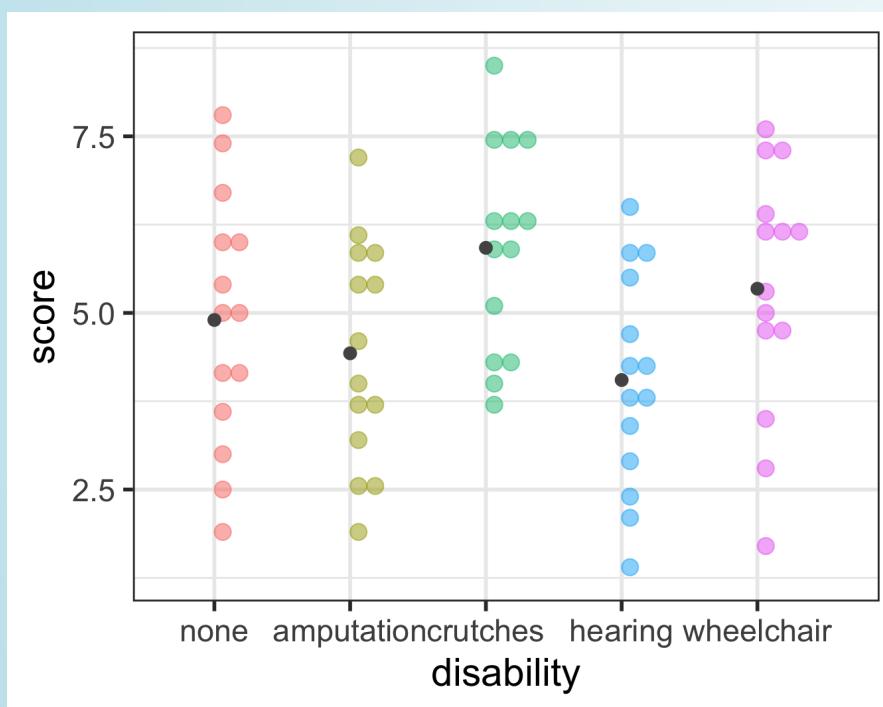
where s_i is the standard deviation of the i^{th} group

- This is the sum of the squared differences between each observed x_{ij} value and its group mean \bar{x}_i .
- That is, it is the deviation of the x_{ij} 's from the predicted score by group.
- Also called the residual sums of squares, or $SS_{residual}$.

Calculate Sums of Squares Error (within groups)

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^k (n_i - 1)s_i^2$$

where s_i is the standard deviation of the i^{th} group



Calculate sd's s_i for each group:

```
1 sd_groups <- employ %>%
2   group_by(disability) %>%
3   summarise(SD = sd(score))
4 sd_groups
```

```
# A tibble: 5 × 2
  disability     SD
  <fct>      <dbl>
1 none        1.79
2 amputation  1.59
3 crutches    1.48
4 hearing     1.53
5 wheelchair  1.75
```

Calculate SSE :

```
1 (SSE <- sum(
2 (Ns$n-1)*sd_groups$SD^2 ))
```

[1] 173.3214

Verify $SST = SSG + SSE$

ANOVA compares the variability between groups to the variability within groups

$$\begin{array}{c} \text{Total} \\ \text{variability} \end{array} = \begin{array}{c} \text{Variability} \\ \text{between} \\ \text{groups} \end{array} + \begin{array}{c} \text{Variability} \\ \text{within} \\ \text{groups} \end{array}$$

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = n_i \sum_{i=1}^k (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$$(N - 1)s^2 = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^k (n_i - 1)s_i^2$$

$$\begin{array}{c} \text{SST} \\ \text{(Total sum} \\ \text{of squares)} \end{array} = \begin{array}{c} \text{SSG} \\ \text{(sum of squares} \\ \text{due to groups)} \end{array} + \begin{array}{c} \text{SSE} \\ \text{(Error sum} \\ \text{of squares)} \end{array}$$

```
1 SST
```

```
[1] 203.8429
```

```
1 SSG + SSE
```

```
[1] 203.8429
```

ANOVA table



The “mean square” is the sum of squares divided by the degrees of freedom

Source	df	Sum of Squares	Mean Square	F-Statistic
Groups	$k-1$	SSG	$MSG = SSG/(k-1)$	$\frac{MSG}{MSE}$
Error	$N-k$	SSE	$MSE = SSE/(N-k)$	
Total	$N-1$	SST		↑ average variability

The **F-statistic** is a ratio of the average variability **between** groups to the average variability **within** groups

Thinking about the F-statistic

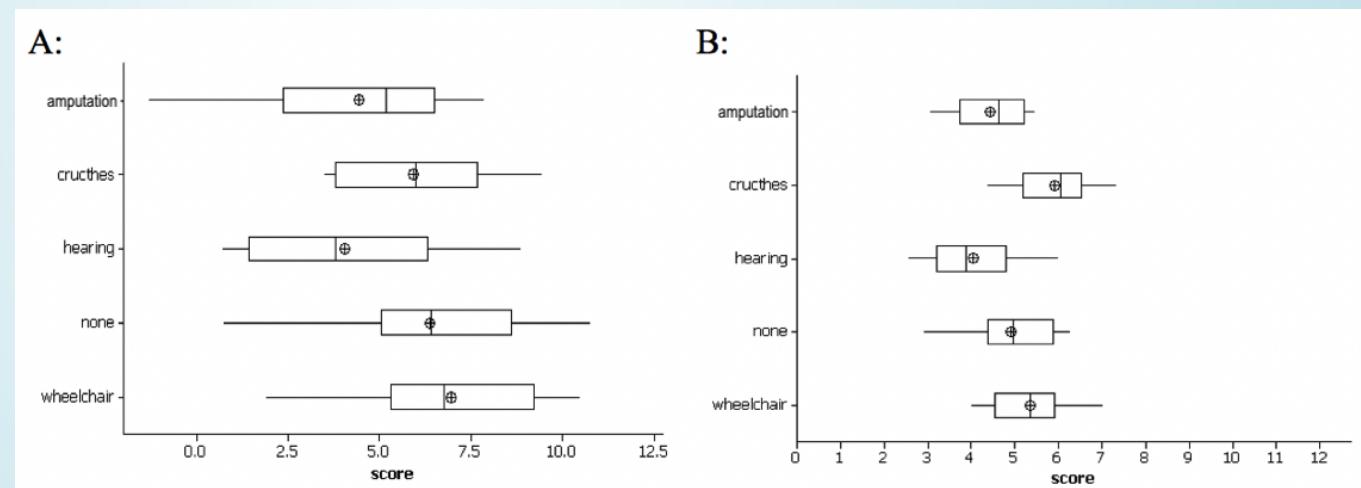
If the groups are actually different, then which of these is more accurate?

1. The variability between groups should be higher than the variability within groups
2. The variability within groups should be higher than the variability between groups

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

If there really is a difference between the groups, we would expect the F-statistic to be which of these:

1. Higher than we would observe by random chance
2. Lower than we would observe by random chance



ANOVA in base R

```
1 # Note that I'm saving the tidy anova table
2 # Will be pulling p-value from this on future slide
3
4 empl_lm <- lm(score ~ disability, data = employ) %>%
5   anova() %>%
6   tidy()
7
8 empl_lm %>% gt()
```

term	df	sumsq	meansq	statistic	p.value
disability	4	30.52143	7.630357	2.86158	0.03012686
Residuals	65	173.32143	2.666484	NA	NA

Hypotheses:

$$H_0 : \mu_{none} = \mu_{amputation} = \mu_{crutches} = \mu_{hearing} = \mu_{wheelchair}$$

vs. $H_A : \text{At least one pair } \mu_i \neq \mu_j \text{ for } i \neq j$

Do we reject or fail to reject H_0 ?

Conclusion to hypothesis test

$H_0 : \mu_{none} = \mu_{amputation} = \mu_{crutches} = \mu_{hearing} = \mu_{wheelchair}$
vs. $H_A : \text{At least one pair } \mu_i \neq \mu_j \text{ for } i \neq j$

```
1 empl_lm # tidy anova output

# A tibble: 2 × 6
  term          df    sumsq   meansq statistic p.value
  <chr>     <int> <dbl>    <dbl>     <dbl>    <dbl>
1 disability      4   30.5     7.63     2.86    0.0301
2 Residuals     65  173.     2.67     NA      NA

1 # Note that this is a vector:
2 empl_lm$p.value

[1] 0.03012686      NA
```

Pull the p-value using base R:

```
1 round(empl_lm$p.value[1],2)

[1] 0.03
```

Pull the p-value using tidyverse:

```
1 empl_lm %>%
  2 filter(term == "disability") %>%
  3 pull(p.value) %>%
  4 round(2)

[1] 0.03
```

- Use $\alpha = 0.05$.
- Do we reject or fail to reject H_0 ?

Conclusion statement:

- There is sufficient evidence that at least one of the disability groups has a mean employment score statistically different from the other groups. ($p\text{-value} = 0.03$).

Conditions for ANOVA

IF ALL of the following conditions hold:

1. The null hypothesis is true
2. Sample sizes in each group group are large (each $n \geq 30$)
 - OR the data are relatively normally distributed in each group
3. Variability is “similar” in all group groups:
 - Is the within group variability about the same for each group?
 - As a rough rule of thumb, this condition is *violated if the standard deviation of one group is more than double the standard deviation of another group*

Checking the **equal variance** condition:

```
1 sd_groups # previously defined  
# A tibble: 5 × 2  
  disability     SD  
  <fct>       <dbl>  
1 none         1.79  
2 amputation   1.59  
3 crutches     1.48  
4 hearing      1.53  
5 wheelchair   1.75  
1 max(sd_groups$SD) / min(sd_groups$SD)  
[1] 1.210425
```

THEN the sampling distribution of the **F-statistic** is an **F-distribution**

Testing variances (Condition 3)

Bartlett's test for equal variances

- H_0 : population variances of group levels are equal
- H_A : population variances of group levels are NOT equal

Note: H_A is same as saying that at least one of the group levels has a different variance

Caution

- Bartlett's test assumes the data in each group are normally distributed.
- Do not use if data do not satisfy the normality condition.

```
1 bartlett.test(score ~ disability, data = employ)
```

```
Bartlett test of homogeneity of variances

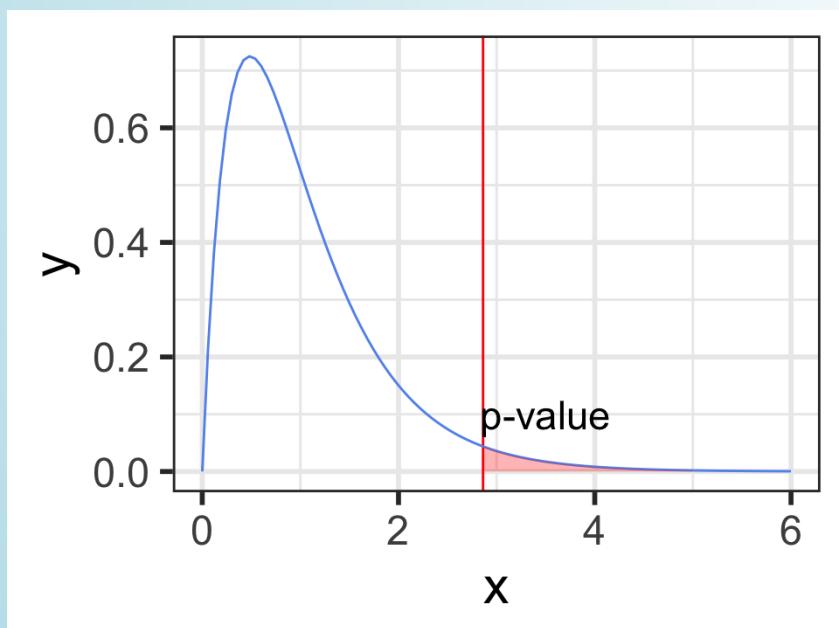
data: score by disability
Bartlett's K-squared = 0.7016, df = 4, p-value = 0.9511
```

Tip

Levene's test for equality of variances is not as restrictive: see <https://www.statology.org/levenes-test-r/>

The F-distribution

- The F-distribution is skewed right.
- The F-distribution has **two different degrees of freedom**:
 - one for the **numerator** of the ratio ($k - 1$) and
 - one for the **denominator** ($N - k$)
- ***p-value***
 - is always the **upper tail**
 - (the area as extreme or more extreme)



```
1 empl_lm %>% gt()
```

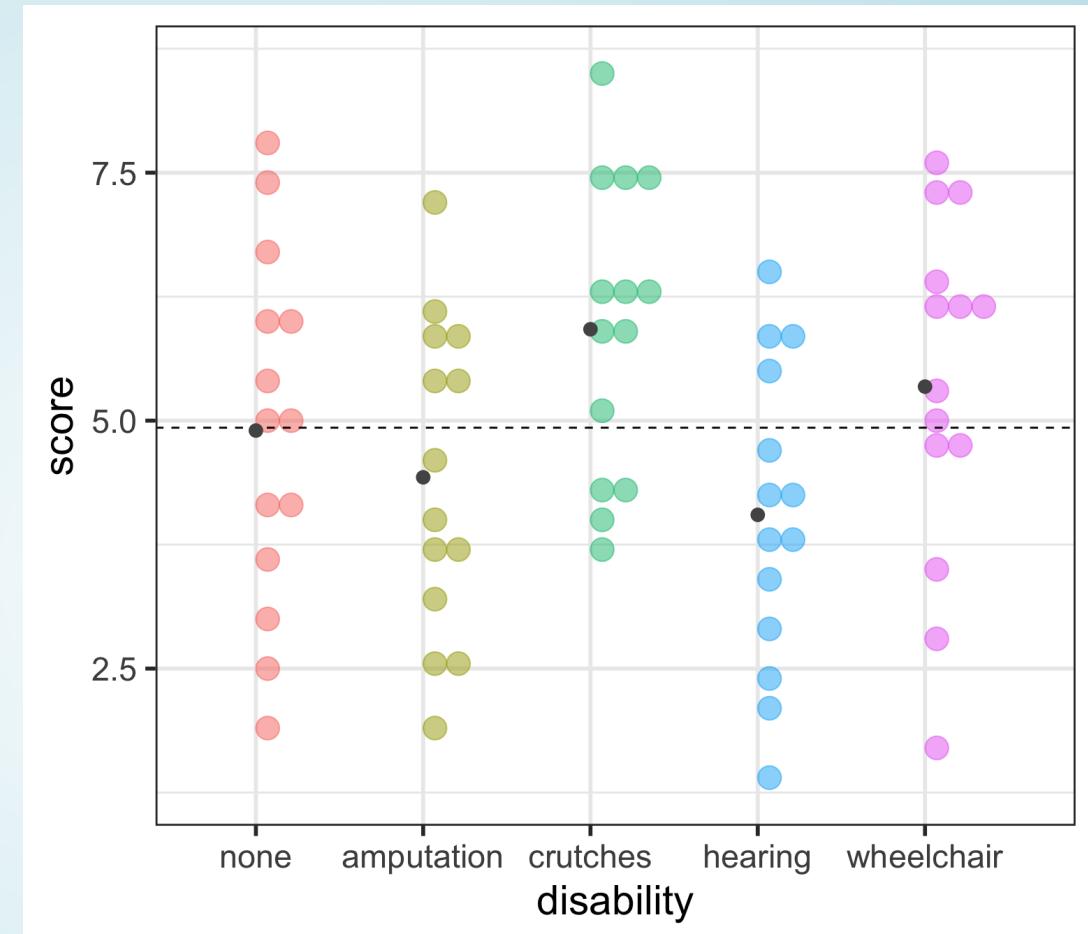
term	df	sumsq	meansq	statistic	p.value
disability	4	30.52143	7.630357	2.86158	0.03012686
Residuals	65	173.32143	2.666484	NA	NA

```
1 # p-value using F-distribution
2
3 pf(2.86158, df1=5-1, df2=70-5,
4 lower.tail = FALSE)
```

```
[1] 0.03012688
```

Which groups are statistically different?

- So far we've only determined that *at least one of the groups is different* from the others,
 - but we don't know which.
- What's your guess?



Post-hoc testing for ANOVA

determining which groups are statistically different

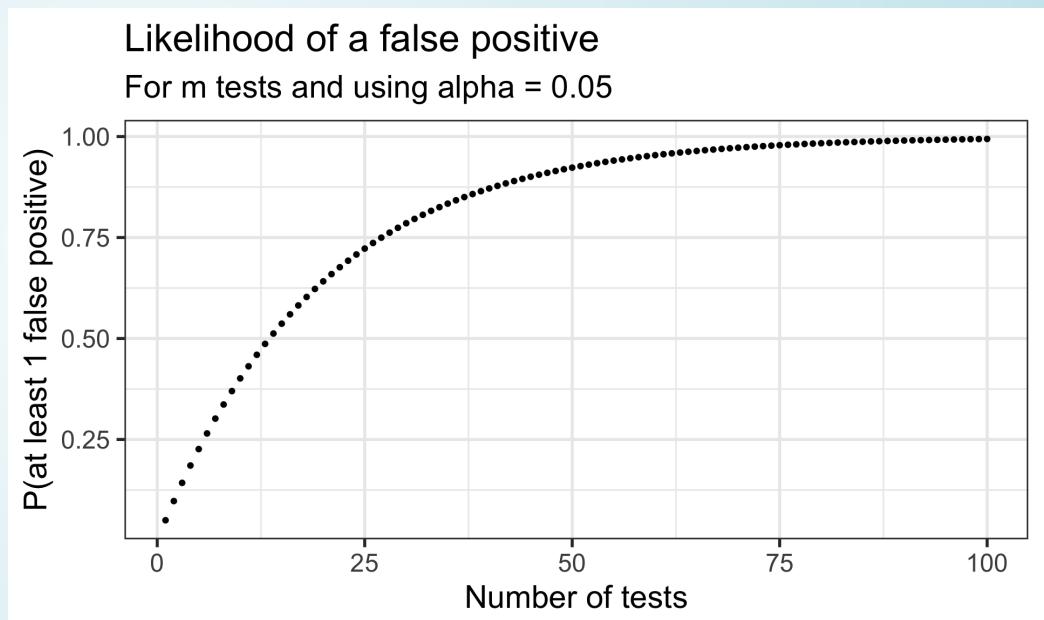
Post-hoc testing: pairwise t-tests

- In post-hoc testing we **run all the pairwise t-tests** comparing the means from each pair of groups.
- With 5 groups, this involves doing $\binom{5}{2} = \frac{5!}{2!3!} = \frac{5 \cdot 4}{2} = 10$ different pairwise tests.

Problem:

- Although the ANOVA test has an α chance of a Type I error (finding a difference between a pair that aren't different),
- the overall Type I error rate will be much higher when running many tests simultaneously.

$$P(\text{making an error}) = \alpha$$
$$P(\text{not making an error}) = 1 - \alpha$$
$$P(\text{not making an error in } m \text{ tests}) = (1 - \alpha)^m$$
$$P(\text{making at least 1 error in } m \text{ tests}) = 1 - (1 - \alpha)^m$$



The Bonferroni Correction (1/2)

A very conservative (but very popular) approach is to divide the α level by how many tests m are being done:

$$\alpha_{Bonf} = \frac{\alpha}{m}$$

- This is equivalent to multiplying the p -values by m :

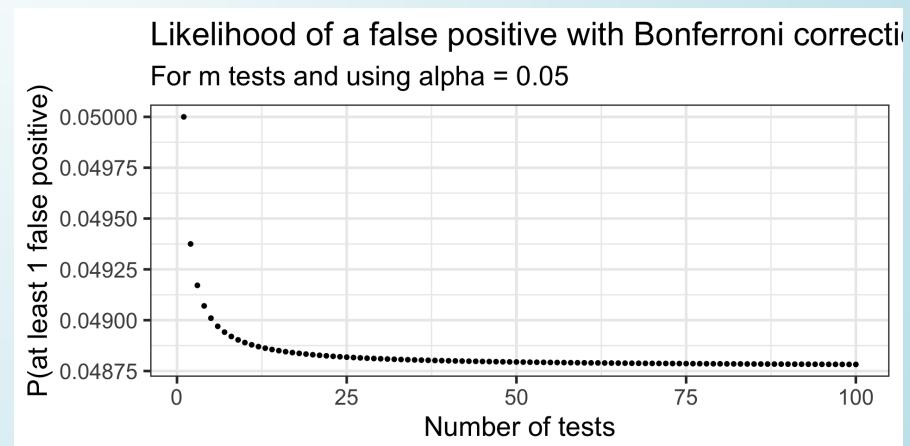
$$p\text{-value} < \alpha_{Bonf} = \frac{\alpha}{m}$$

is the same as

$$m \cdot (p\text{-value}) < \alpha$$

The Bonferroni correction is popular since it's very easy to implement.

- The **plot below** shows the **likelihood of making at least one Type I error** depending on how many tests are done.
- Notice the likelihood decreases very quickly
 - Unfortunately the likelihood of a Type II error is increasing as well
 - It becomes "harder" and harder to reject H_0 if doing many tests.



The Bonferroni Correction (2/2)

Pairwise t-tests without any *p*-value adjustments:

```
1 pairwise.t.test(employ$score,
2                      employ$disability,
3                      p.adj="none")
```

```
Pairwise comparisons using t tests with pooled SD

data: employ$score and employ$disability

      none amputation crutches hearing
amputation 0.4477 -          -         -
crutches    0.1028 0.0184   -          -
hearing     0.1732 0.5418   0.0035    -
wheelchair  0.4756 0.1433   0.3520   0.0401

P value adjustment method: none
```

Pairwise t-tests **with Bonferroni *p*-value adjustments:**

```
1 pairwise.t.test(employ$score,
2                      employ$disability,
3                      p.adj="bonferroni")
```

```
Pairwise comparisons using t tests with pooled SD

data: employ$score and employ$disability

      none amputation crutches hearing
amputation 1.000 -          -         -
crutches    1.000 0.184    -          -
hearing     1.000 1.000    0.035    -
wheelchair  1.000 1.000    1.000   0.401

P value adjustment method: bonferroni
```

- Since there were 10 tests, all the *p*-values were multiplied by 10.
- Are there any significant pairwise differences?

Tukey's Honest Significance Test (HSD)

- Tukey's Honest Significance Test (HSD) controls the “family-wise probability” of making a Type I error using a much less conservative method than Bonferroni
 - **It is specific to ANOVA**
- In addition to adjusted p -values, it also calculates Tukey adjusted CI's for all pairwise differences
- The function **TukeyHSD()** creates a **set of confidence intervals** of the differences between means with the specified **family-wise probability of coverage**.

```
1 # need to run the model using `aov` instead of `lm`  
2 empl_aov <- aov(score ~ disability, data = employ)  
3  
4 TukeyHSD(x=empl_aov, conf.level = 0.95)
```

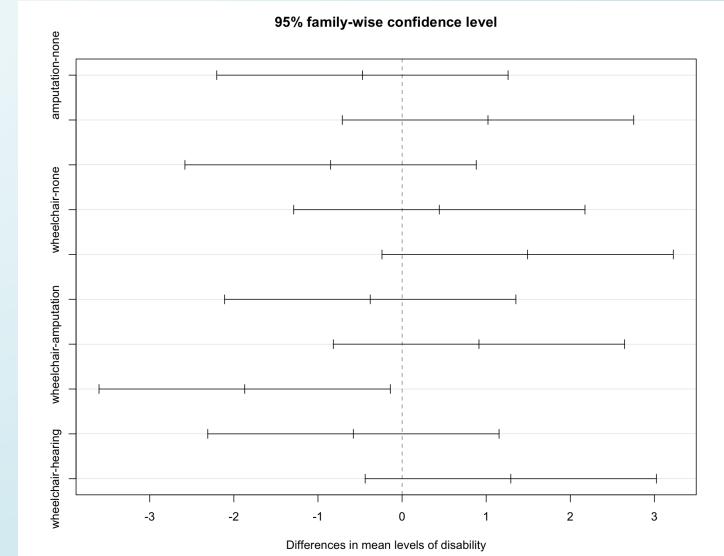
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = score ~ disability, data = employ)

\$disability

	diff	lwr	upr	p adj
amputation-none	-0.4714286	-2.2031613	1.2603042	0.9399911
crutches-none	1.0214286	-0.7103042	2.7531613	0.4686233
hearing-none	-0.8500000	-2.5817328	0.8817328	0.6442517
wheelchair-none	0.4428571	-1.2888756	2.1745899	0.9517374
crutches-amputation	1.4928571	-0.2388756	3.2245899	0.1232819
hearing-amputation	-0.3785714	-2.1103042	1.3531613	0.9724743
wheelchair-amputation	0.9142857	-0.8174470	2.6460185	0.5781165
hearing-crutches	-1.8714286	-3.6031613	-0.1396958	0.0277842
wheelchair-crutches	-0.5785714	-2.3103042	1.1531613	0.8812293
wheelchair-hearing	1.2928571	-0.4388756	3.0245899	0.2348141

```
1 plot(TukeyHSD(x=empl_aov,  
2 conf.level = 0.95))
```



There are many more multiple testing adjustment procedures

- Bonferroni is popular because it's so easy to apply
- Tukey's HSD is usually used for ANOVA
- Code below used Holm's adjustment

```
1 # default is Holm's adjustments
2 pairwise.t.test(employ$score,
3                   employ$disability)
```

```
Pairwise comparisons using t tests with pooled SD

data: employ$score and employ$disability

none    amputation   crutches   hearing
amputation 1.000      -          -          -
crutches    0.719  0.165      -          -
hearing     0.866  1.000  0.035      -
wheelchair  1.000  0.860  1.000  0.321

P value adjustment method: holm
```

- **False discovery rate (fdr)** *p*-value adjustments are popular in omics, or whenever there are *many* tests being run:

```
1 pairwise.t.test(employ$score,
2                   employ$disability,
3                   p.adj="fdr")
```

```
Pairwise comparisons using t tests with pooled SD

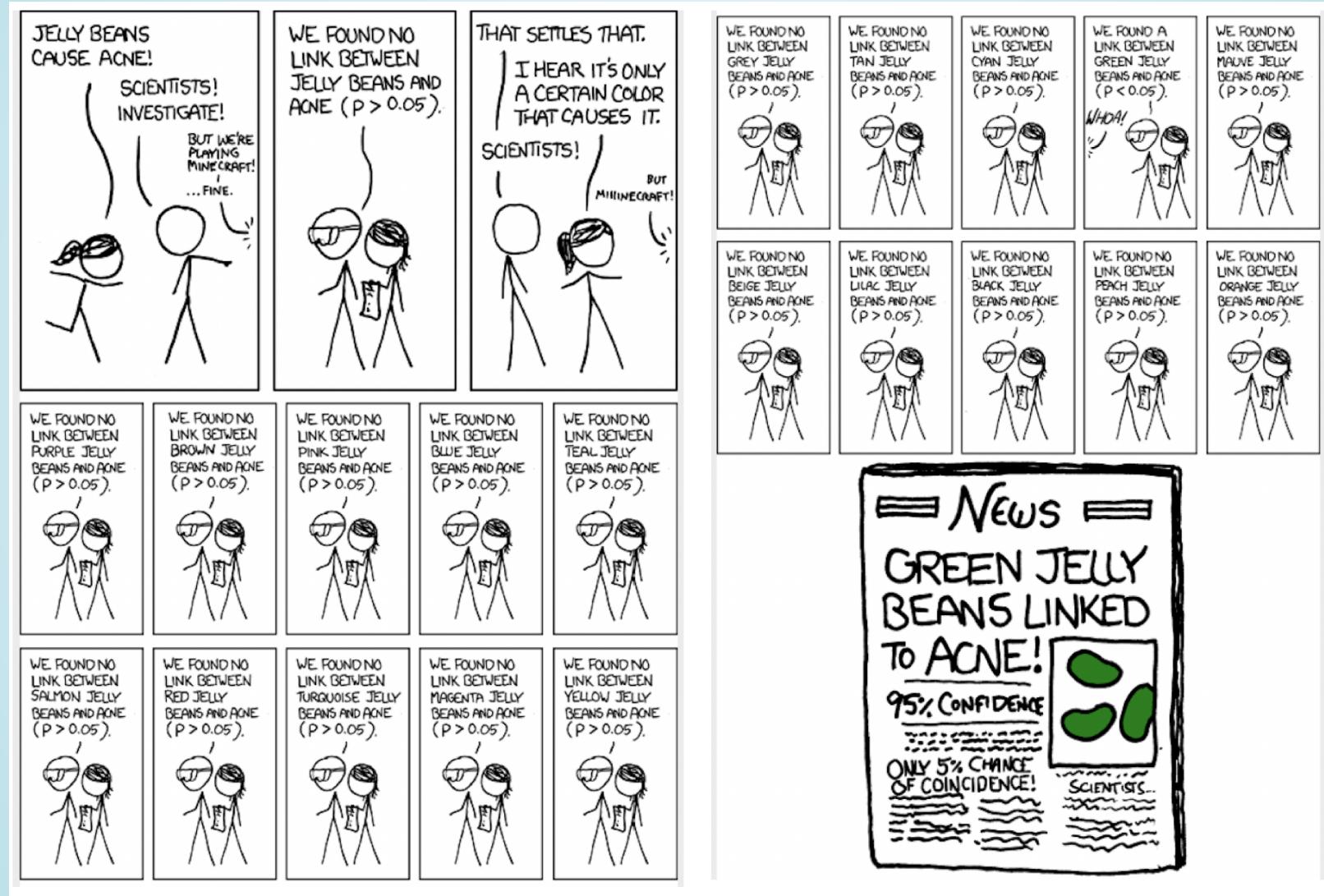
data: employ$score and employ$disability

            none    amputation   crutches   hearing
amputation 0.528      -          -          -
crutches    0.257  0.092      -          -
hearing     0.289  0.542  0.035      -
wheelchair  0.528  0.287  0.503  0.134

P value adjustment method: fdr
```

Multiple testing

post-hoc testing vs. testing many outcomes



<https://xkcd.com/882/>

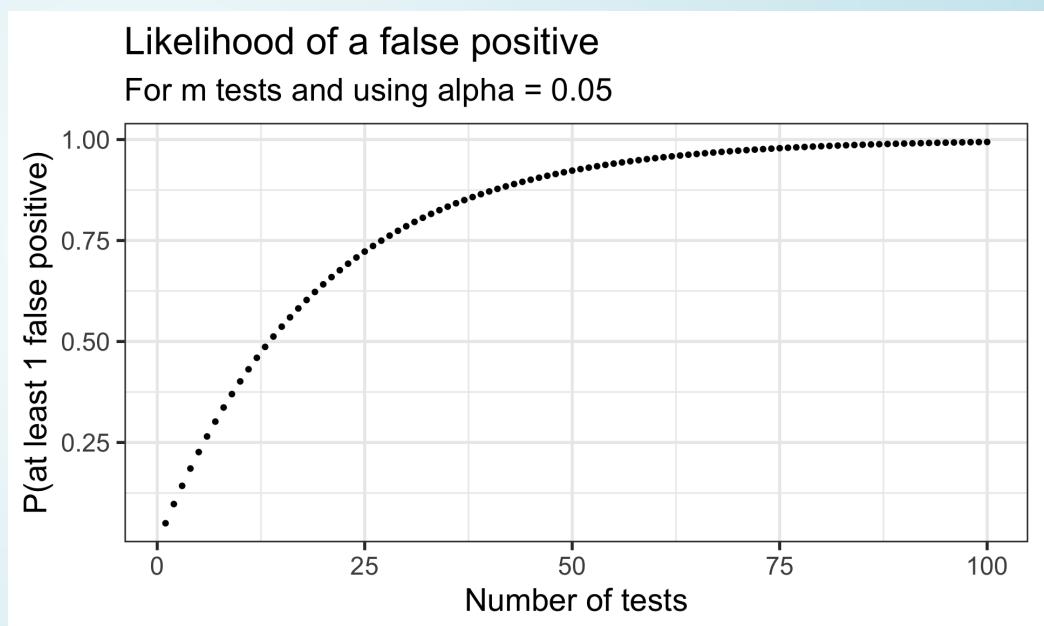
Multiple testing: controlling the Type I error rate

- The multiple testing issue is not unique to ANOVA post-hoc testing.
- It is also a concern when running separate tests for many related outcomes.
- **Beware of *p*-hacking!**

Problem:

- Although one test has an α chance of a Type I error (finding a difference between a pair that aren't different),
- the **overall Type I error rate will be much higher when running many tests simultaneously.**

$$P(\text{making an error}) = \alpha$$
$$P(\text{not making an error}) = 1 - \alpha$$
$$P(\text{not making an error in } m \text{ tests}) = (1 - \alpha)^m$$
$$P(\text{making at least 1 error in } m \text{ tests}) = 1 - (1 - \alpha)^m$$



ANOVA Summary

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

vs. $H_A : \text{At least one pair } \mu_i \neq \mu_j \text{ for } i \neq j$

ANOVA table in R:

```
1 lm(score ~ disability, data = employ) %>% anova()
```

Analysis of Variance Table

Response: score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
disability	4	30.521	7.6304	2.8616	0.03013 *
Residuals	65	173.321	2.6665		

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

ANOVA table

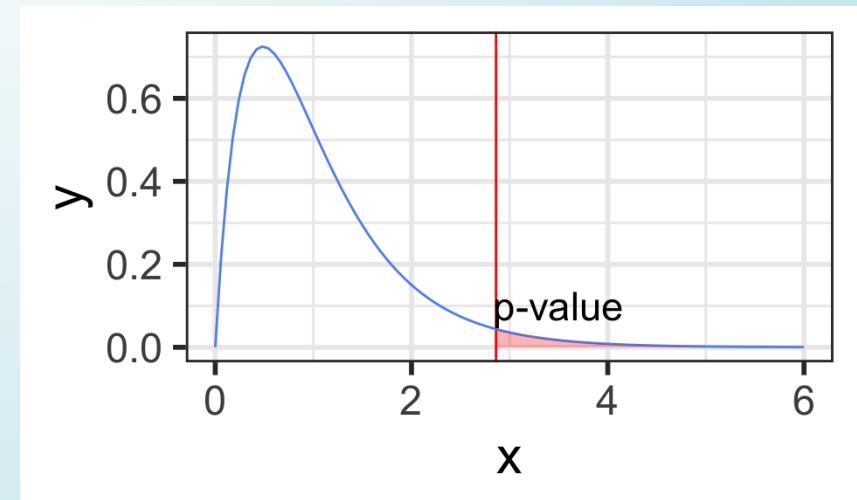
The “mean square” is the sum of squares divided by the degrees of freedom

↓

Source	df	Sum of Squares	Mean Square	F-Statistic
Groups	$k-1$	SSG	$\text{MSG} = \frac{\text{SSG}}{k-1}$	$\frac{\text{MSG}}{\text{MSE}}$
Error	$N-k$	SSE	$\text{MSE} = \frac{\text{SSE}}{N-k}$	
Total	$N-1$	SST		

↑ variability ↑ average variability

F-distribution & p-value



Post-hoc testing

What's next?

