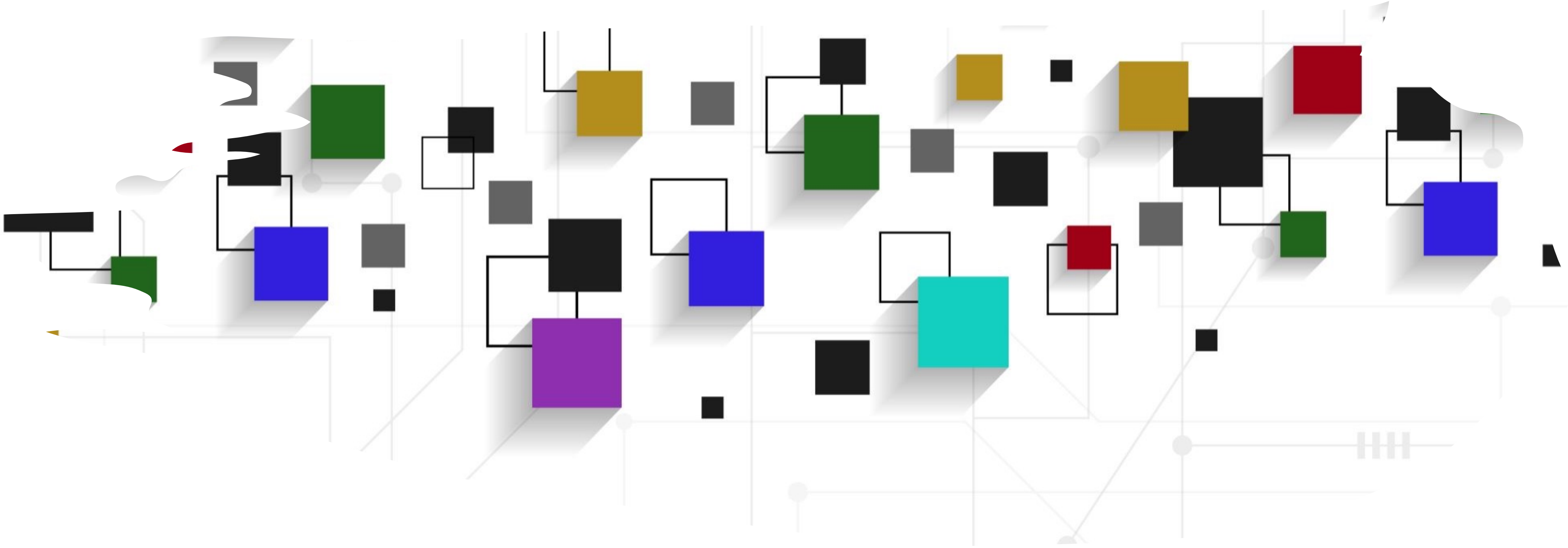


# CogLab: Making Inferences

WEEK 11

# recap: Oct 24/26, 2023

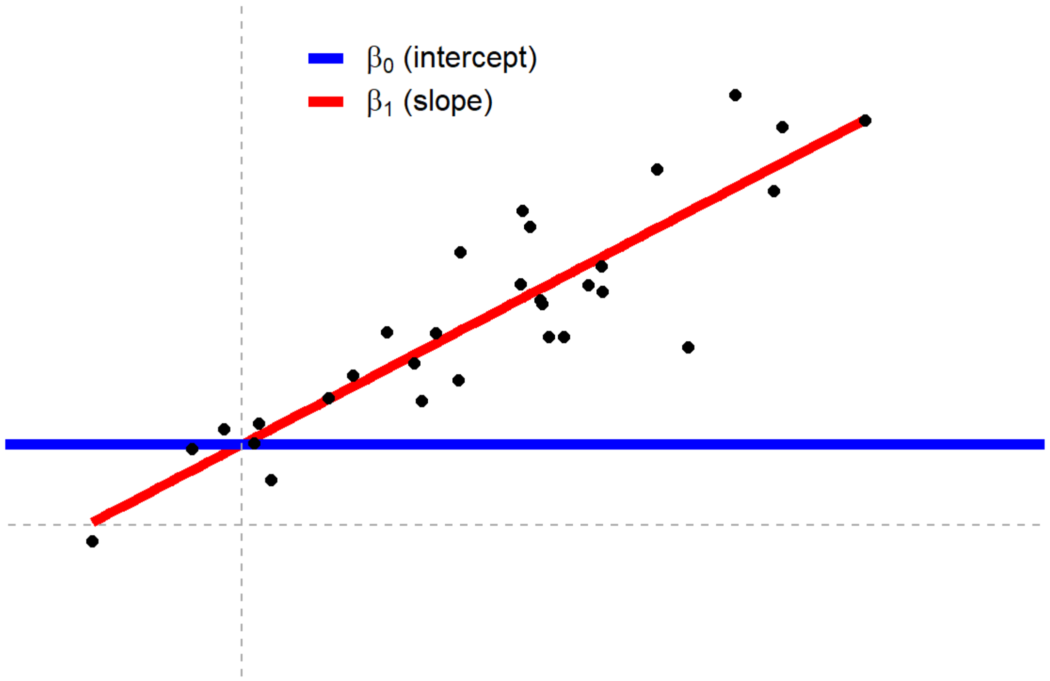
- what we covered:
  - manipulating data using tidyverse verbs
  - linear regression
- your to-do's were:
  - *prep*: complete all primers
  - *prep*: read about hypothesis testing
  - *schedule*: group meeting

# today’s agenda

- linear regression continued
- two-way/multiple linear regression

# linear regression

- a linear regression (or a linear model) is a model that fits a line to a set of data points
  - Y = aX + b
  - Y: dependent variable
  - X: independent variable
  - a? b?
- a: slope, b: intercept
- sometimes, we reorder this equation:
  - $y = \beta_0 + \beta_1 x$
  - $\beta_0$: intercept (where the line cuts the y-axis)
  - $\beta_1$: slope (the change in y due to x)
- in this framework, the null hypothesis ($H_0$) is that $\beta_1$ = 0, i.e., there is no change in y due to x
  - $H_0$: $\beta_1 = 0$




https://lindeloev.github.io/tests-as-linear/

# linear regression in R

- predict height by weight
- print the summary of the model
- what is the equation of the line?

```
women_model = lm(data = women, height ~ weight)
```

```
summary(women_model)
```

```
Call:
lm(formula = height ~ weight, data = women)

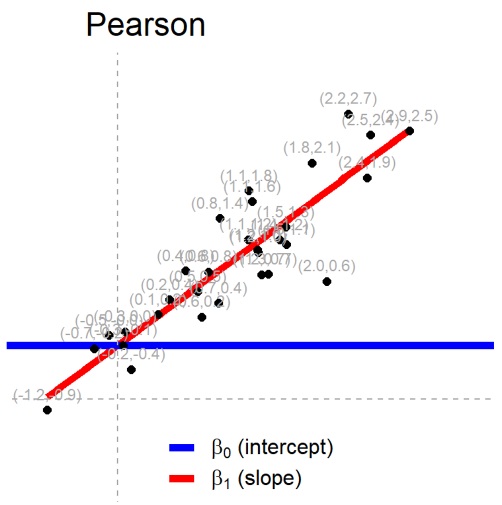
Residuals:
     Min       1Q   Median       3Q      Max 
-0.83233 -0.26249  0.08314  0.34353  0.49790 

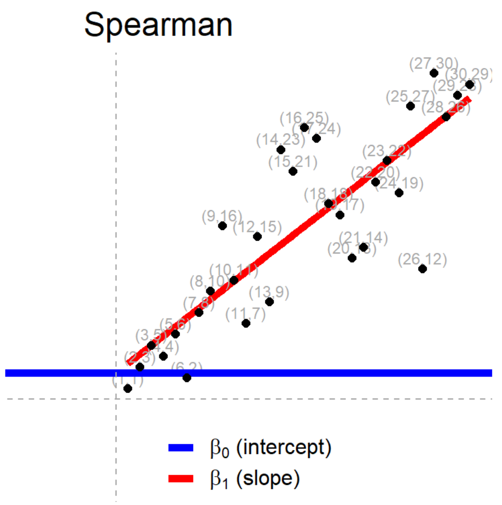
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.723456   1.043746   24.64 2.68e-12 ***
weight       0.287249   0.007588   37.85 1.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.44 on 13 degrees of freedom
Multiple R-squared:  0.991,	Adjusted R-squared:  0.9903 
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

# linear regression and correlation

- correlations also describe the relationship between Y and X, so what's the difference?
- mathematically, correlations are equivalent to a linear model where a line is being fit to a set of data points
- two common correlation
    - Pearson's *r*: *r* = slope if x and y have the same standard deviation
    - Spearman's *rho* = same linear model but with ranks of x and Y
        - $\text{rank}(y) = \beta_0 + \beta_1\,\text{rank}(x)$




https://lindeloev.github.io/tests-as-linear/

# linear regression and correlation

- compute the standard deviation of the height and weight columns
- create two new columns that contain the z-scored height and weight
- compute the standard deviation of the z-scored height and weight columns

```
sd(women$height)
sd(women$weight)
```

```
women = women %>%
  mutate(z_height = scale(height),
         z_weight = scale(weight))
```

```
sd(women$height)
sd(women$weight)
```

# linear regression and correlation

- predict the z-scored height with the z-scored weight using linear regression
- now compute the correlation between the two columns using summarize() and cor()

```
women_model_2 = lm(data = women, z_height ~ z_weight)
summary(women_model_2)
```

```
Call:
lm(formula = z_height ~ z_weight, data = women)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.18611 -0.05869  0.01859  0.07682  0.11133 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -8.268e-16  2.541e-02    0.00        1    
z_weight     9.955e-01  2.630e-02   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

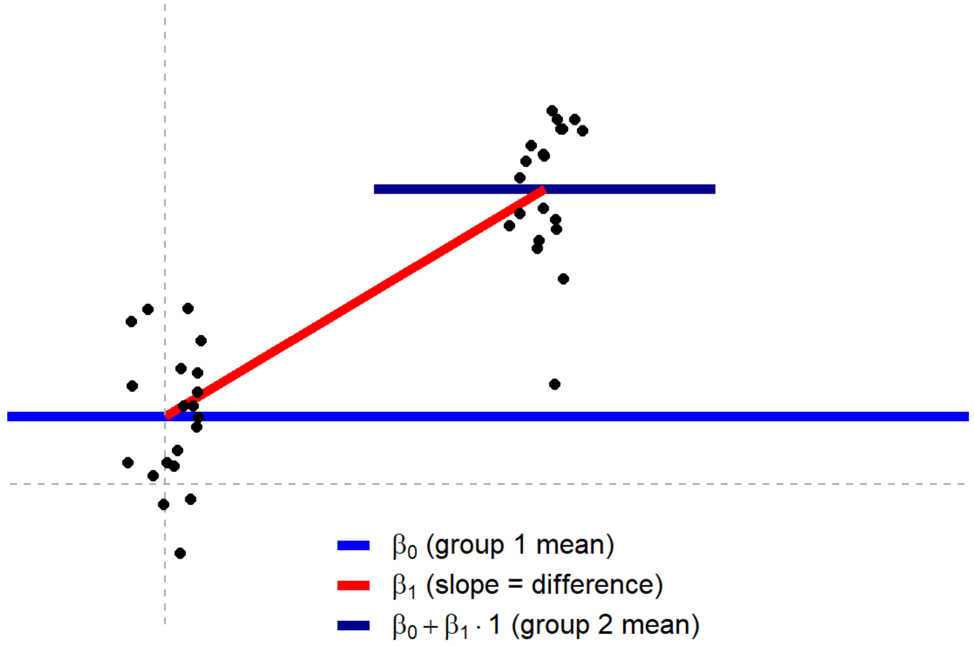
Residual standard error: 0.0984 on 13 degrees of freedom
Multiple R-squared:  0.991,	Adjusted R-squared:  0.9903 
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

```
women %>%
  summarise(r = cor(z_height, z_weight))
```

```
          r
1 0.9954948
```

# linear regression and t-tests

- unpaired/independent samples t-test
  - $y = \beta_0 + \beta_1 x$
  - x = 0 or 1 (which group)
  - $H_0$: $\beta_1 = 0$
  - comparing paired differences and testing whether the difference is significantly different from 0
  - note that "x" here contains information about group membership for each y



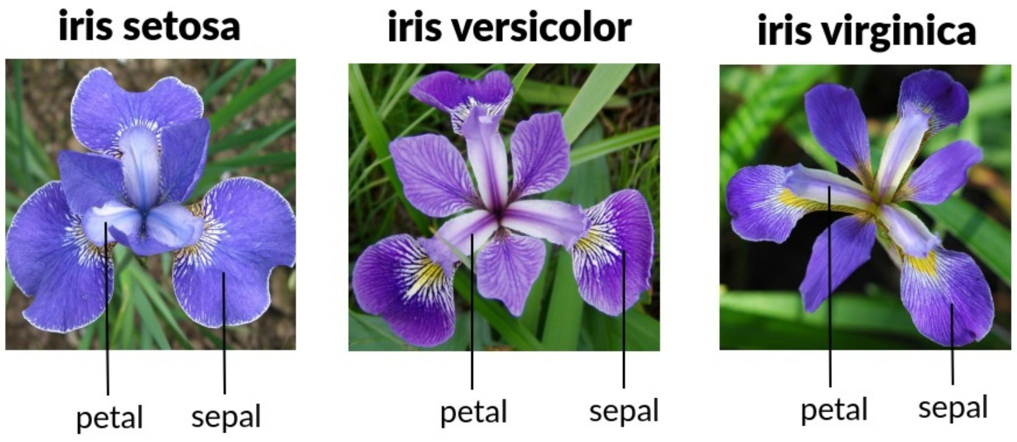

https://lindeloev.github.io/tests-as-linear/

# revisiting iris

- recall that iris contains flower petal and sepal information for three species

```
data("iris")
View(iris)
```

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 4.8 | 3.4 | 1.6 | 0.2 | setosa |
| 4.8 | 3.0 | 1.4 | 0.1 | setosa |
| 4.3 | 3.0 | 1.1 | 0.1 | setosa |
| 5.8 | 4.0 | 1.2 | 0.2 | setosa |
| 5.7 | 4.4 | 1.5 | 0.4 | setosa |



# subset of iris

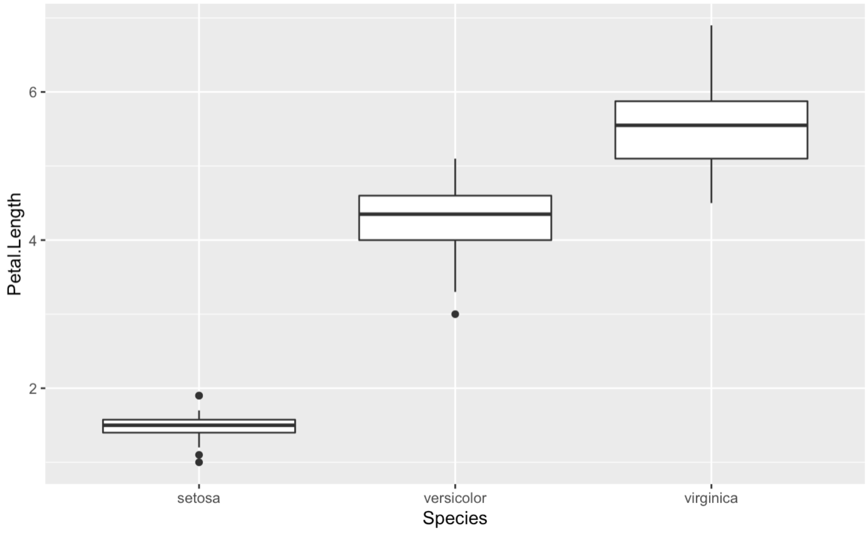
- create a subset of iris that only contains setosa and virginica
- plot the petal lengths by species in a boxplot

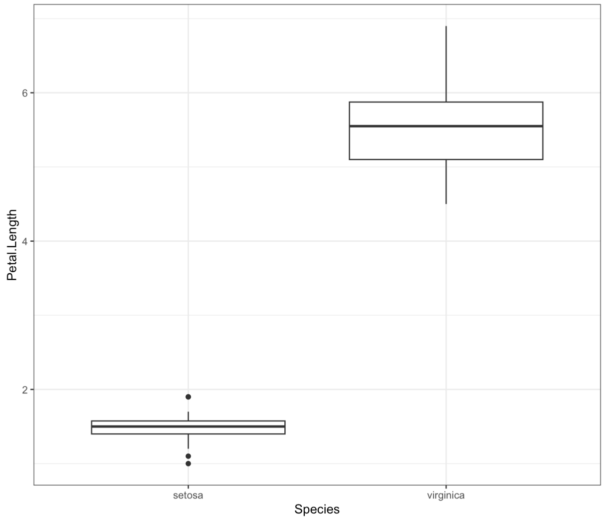
```
## t -test

```{r}
iris_subset = iris %>%
  filter(Species %in% c("setosa", "virginica"))

```
```

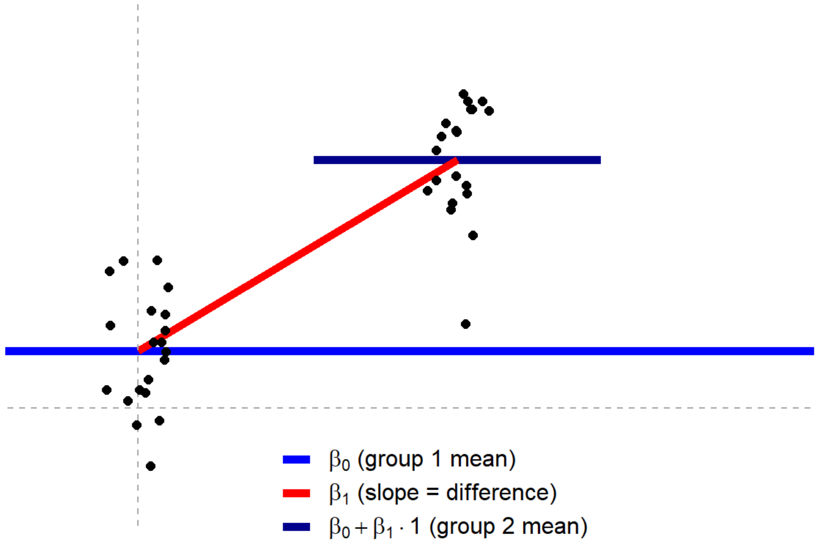
```
iris_subset %>%
  ggplot(aes(x = Species, y = Petal.Length))+
  geom_col()
```



# comparing

- create linear model
- conduct t-test




```r
iris_subset_lm = lm(data = iris_subset, Petal.Length ~ Species)
summary(iris_subset_lm)
```

```
Call:
lm(formula = Petal.Length ~ Species, data = iris_subset)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.0520 -0.1620  0.0380  0.1405  1.3480 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.46200    0.05786   25.27   <2e-16 ***
Speciesvirginica   4.09000    0.08182   49.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4091 on 98 degrees of freedom
Multiple R-squared:  0.9623,	Adjusted R-squared:  0.9619 
F-statistic:  2499 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
t.test(Petal.Length ~ Species, data = iris_subset)
```

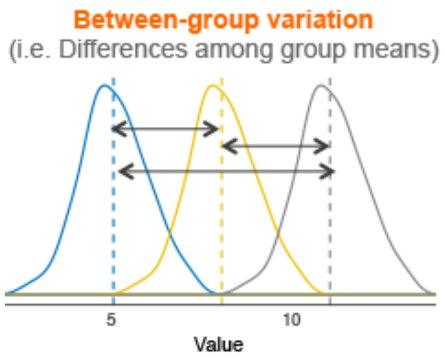
```
	Welch Two Sample t-test

data:  Petal.Length by Species
t = -49.986, df = 58.609, p-value < 2.2e-16
alternative hypothesis: true difference in means between group setosa and group virginica is not equal to 0
95 percent confidence interval:
 -4.253749 -3.926251
sample estimates:
   mean in group setosa mean in group virginica 
                  1.462                   5.552 
```

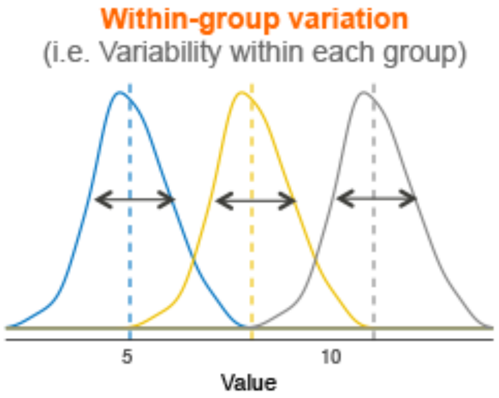
# testing more than two groups

- a t-test is a special case of linear models
- it is *also* a special case of only comparing two groups
- example of comparing more than two groups?

# ANOVA: Analysis of Variance

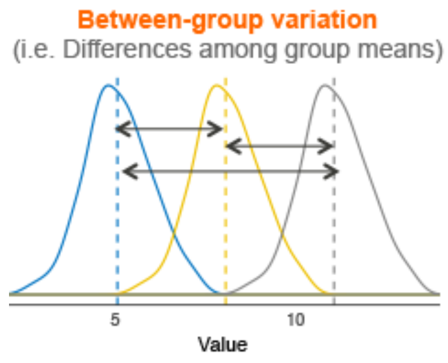
- a *generalized* t-test for more than two means/groups!
- key idea: we will try to understand the difference between groups and whether it can be attributed to our "conditions" or randomness
- $SS_{between}$ = variation between groups
- $SS_{within}$ = variation within groups
- $F = SS_{between}/Ss_{within}$
- If F > 1, the group differences are greater than what would be expected as random variation within groups

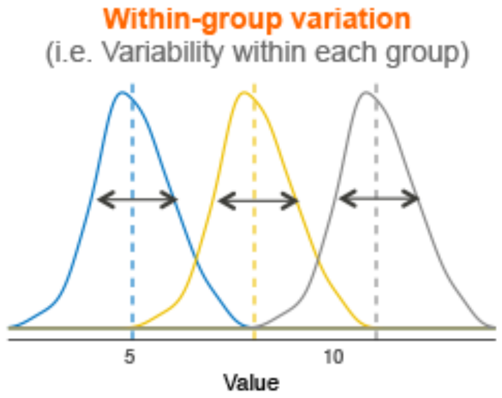


# types of ANOVAs

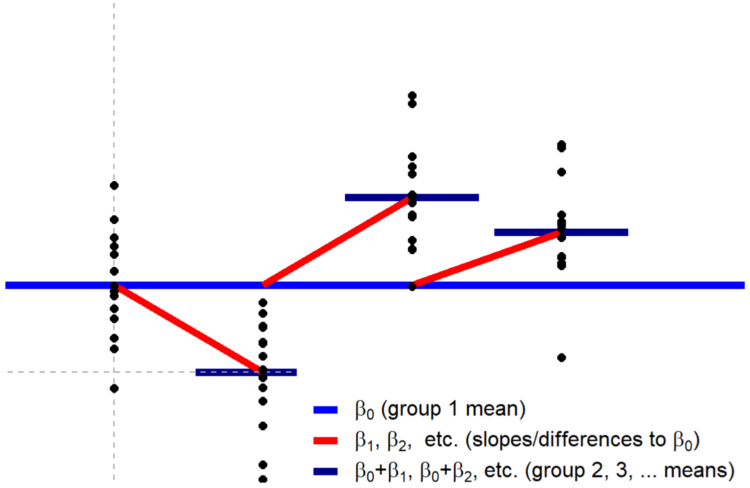
- n( independent variables)
  - one-way
  - two-way
  - three-way
- within or between subjects
  - between subjects: regular ANOVA
  - within-subjects: repeated measures ANOVA



# one-way ANOVA

- predict the petal lengths using the full iris dataset




```
full_iris_model = lm(data = iris, Petal.Length ~ Species)
summary(full_iris_model)
```

```
Call:
lm(formula = Petal.Length ~ Species, data = iris)

Residuals:
   Min     1Q Median     3Q    Max
-1.260 -0.258  0.038  0.240  1.348

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.46200    0.06086   24.02   <2e-16 ***
Speciesversicolor  2.79800    0.08607   32.51   <2e-16 ***
Speciesvirginica   4.09000    0.08607   47.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4303 on 147 degrees of freedom
Multiple R-squared:  0.9414,    Adjusted R-squared:  0.9406
F-statistic:  1180 on 2 and 147 DF,  p-value: < 2.2e-16
```

```
full_iris_aov = aov(data = iris, Petal.Length ~ Species)
summary(full_iris_aov)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

# follow-up tests

- when more than two groups are present, it can be useful to understand exactly which groups differ from each other
- install emmeans package
- load the package inline and compute pairwise differences
- compare to lm summary

```
Call:
lm(formula = Petal.Length ~ Species, data = iris)

Residuals:
   Min     1Q Median     3Q    Max 
-1.260 -0.258  0.038  0.240  1.348 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        1.46200    0.06086   24.02   <2e-16 ***
Speciesversicolor  2.79800    0.08607   32.51   <2e-16 ***
Speciesvirginica   4.09000    0.08607   47.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4303 on 147 degrees of freedom
Multiple R-squared:  0.9414,	Adjusted R-squared:  0.9406 
F-statistic:  1180 on 2 and 147 DF,  p-value: < 2.2e-16
```

```
#install.packages("emmeans")
emmeans::emmeans(full_iris_model,
                 pairwise ~ Species,
                 adjust="tukey")
```

```
$emmeans
 Species    emmean     SE  df lower.CL upper.CL
 setosa       1.46 0.0609 147     1.34     1.58
 versicolor   4.26 0.0609 147     4.14     4.38
 virginica    5.55 0.0609 147     5.43     5.67

Confidence level used: 0.95 

$contrasts
 contrast               estimate     SE  df t.ratio p.value
 setosa - versicolor       -2.80 0.0861 147 -32.510  <.0001
 setosa - virginica        -4.09 0.0861 147 -47.521  <.0001
 versicolor - virginica    -1.29 0.0861 147 -15.012  <.0001

P value adjustment: tukey method for comparing a family of 3 estimates
```

# next class

- **before** class
  - *resubmit:* formative assignment #2
  - *finalize:* experiment
  - submit: pre-registration
- **during** class
  - multiple regression in R
  - linear models for non-independent data