

# **Week 7:**

# **Statistical tests involving two variables (part I)**

Phase 3

# Key concepts

- Sometimes we want to make inferences that may involve **two** variables in our dataset.
- For example, testing differences in means between two or more populations.
- Remember: a population is just a group that you want to draw conclusions about.
- Different **data types** and **ASSUMPTIONS** → **a particular** statistical test.
- Here (and in all phase III), we are going to concentrate on **parametric** tests, which assume an **underlying distribution** to compute the p-values.

# Your dataset

Inferences based on the relation between two variables

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

# Your dataset

Inferences based on the relation between two variables

e.g. are there  
differences in ACT  
scores between  
men and women?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

# Your dataset

Inferences based on the relation between two variables

e.g. do men and women taking ACT exams exhibit differences in their education levels?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

# Your dataset

Inferences based on the relation between two variables

e.g. do people of  
greater ages tend to  
score ACT exams  
better?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

# Continuous vs categorical variable

- Here you'll probably want to test whether estimations in the continuous variable (e.g. **means**) differ across levels of the categorical variable.
- Key concept: Each level in the categorical variable represents a **different population** (e.g. healthy and disease, education levels, etc)
- The most famous tests in this scenario are the **two sample t-test** (two categories) and the **one-way** analysis of variance (**ANOVA**) test ( $\geq$  two categories).

# Two levels in your categorical variable

Research question:  
are there  
differences in ACT  
scores between  
men and women?

gender	education	age	ACT	SATV	SATQ	
2		3	19	24	500	500
2		3	23	35	600	500
2		3	20	21	480	470
1		4	27	26	550	520
1		2	33	31	600	550
1		5	26	28	640	640
2		5	30	36	610	500
1		3	19	22	520	560
2		4	23	22	400	600
2		5	40	35	730	800



# Two levels in your categorical variable

Research question:  
are there  
differences in ACT  
scores between  
men and women?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

$X_1$



# Two levels in your categorical variable

Research question:  
are there  
differences in ACT  
scores between  
men and women?

gender	education	age	ACT	SATV	SATQ
2		3	19	24	500
2		3	23	35	600
2		3	20	21	480
1		4	27	26	550
1		2	33	31	600
1		5	26	28	640
2		5	30	36	610
1		3	19	22	520
2		4	23	22	400
2		5	40	35	730

$X_2$

# Two levels in your categorical variable

Research question:  
are there  
differences in ACT  
scores between  
men and women?

gender	education	age	ACT	SATV	SATQ	
2		3	19	24	500	500
2		3	23	35	600	500
2		3	20	21	480	470
1		4	27	26	550	520
1		2	33	31	600	550
1		5	26	28	640	640
2		5	30	36	610	500
1		3	19	22	520	560
2		4	23	22	400	600
2		5	40	35	730	800

$X_1$

$X_2$

# Welch's t-test

- **Categorical variable**: only two levels → **Two** samples (two populations!) of the **continuous** variable,  $\mathbf{X}_1$  and  $\mathbf{X}_2$ .
- It tests  $\mu_1 - \mu_2$ , the **difference between the means** of populations with respect to a hypothesized value  $\Delta_0$ .

$$\mathbf{H}_0: \mu_1 - \mu_2 = \Delta_0$$

$$\mathbf{H}_A: \mu_1 - \mu_2 > \text{ or } < \Delta_0 \text{ (one-sided)}$$

$$\mu_1 - \mu_2 \neq \Delta_0 \text{ (two-sided)}$$

# Welch's t-test

- **Categorical variable**: only two levels → **Two** samples (two populations!) of the **continuous** variable,  $\mathbf{X}_1$  and  $\mathbf{X}_2$ .
- It tests  $\mu_1 - \mu_2$ , the **difference between the means** of populations with respect to a hypothesized value  $\Delta_0$ .
- Example: “For the insula, is its average activation for healthy people different from its average activation in Autistic subjects?”
- There exists a less general version, the **Student's t-test** (see next slides).

# Welch's t-test

## Assumptions:

1. Independence: Observations are not correlated with each other (e.g. in *cross-sectional* studies), both **within and between** samples  $X_1$  and  $X_2$ .
2. Normality. Both values in  $X_1$  and  $X_2$  follow a gaussian distribution, or their sample sizes are big enough (thanks, Central Limit Theorem!)

# Welch's t-test

## Assumptions:

1. Independence: Observations are not correlated with each other (e.g. in *cross-sectional* studies), both **within and between** samples  $X_1$  and  $X_2$ .
2. Normality. Both values in  $X_1$  and  $X_2$  follow a gaussian distribution, or their sample sizes are big enough (thanks, Central Limit Theorem!)
3. (Extra assumption) Variances in  $X_1$  and  $X_2$ , are different i.e.  $\sigma^2_1 \neq \sigma^2_2$ .

# Welch's t-test

## Assumptions:

1. Independence: Observations are not correlated with each other (e.g. in *cross-sectional* studies), both **within and between** samples  $X_1$  and  $X_2$ .
2. Normality. Both values in  $X_1$  and  $X_2$  follow a gaussian distribution, or their sample sizes are big enough (thanks, Central Limit Theorem!)
3. (Extra assumption) The opposite, i.e.  $\sigma^2_1 = \sigma^2_2$ , corresponds to the **Student's t-test**.



# Welch's t-test

➤ Null Hypothesis  $H_0: \mu_1 - \mu_2 = \Delta_0$

➤ Test statistic: 
$$t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \Delta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

➤ Alternative Hypothesis  $H_A$

$\mu_1 - \mu_2 > \Delta_0$  (one-sided right tail)

$\mu_1 - \mu_2 < \Delta_0$  (one-sided left tail)

$\mu_1 - \mu_2 \neq \Delta_0$  (two-sided)

# Welch's t-test

➤ Null Hypothesis  $H_0: \mu_1 - \mu_2 = \Delta_0$

➤ Test statistic:  $t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \Delta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}} \sim \text{Student's } t \text{ (df)}$

➤ Alternative Hypothesis  $H_A$

$\mu_1 - \mu_2 > \Delta_0$  (one-sided right tail)

$\mu_1 - \mu_2 < \Delta_0$  (one-sided left tail)

$\mu_1 - \mu_2 \neq \Delta_0$  (two-sided)

# Welch's t-test

- Null Hypothesis  $H_0: \mu_1 - \mu_2 = \Delta_0$

- Test statistic: 
$$t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \Delta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

$$\frac{(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2)^2}{(\hat{\sigma}_1^2/N_1)^2/(N_1 - 1) + (\hat{\sigma}_2^2/N_2)^2/(N_2 - 1)}$$

~ Student's t (df)

- Alternative Hypothesis  $H_A$

$\mu_1 - \mu_2 > \Delta_0$  (one-sided right tail)

$\mu_1 - \mu_2 < \Delta_0$  (one-sided left tail)

$\mu_1 - \mu_2 \neq \Delta_0$  (two-sided)

# Welch's t-test

- Null Hypothesis  $H_0: \mu_1 - \mu_2 = \Delta_0$

- Test statistic: 
$$t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \Delta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}} \sim \text{Student's } t \text{ (df)}$$

$\frac{(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2)^2}{(\hat{\sigma}_1^2/N_1)^2/(N_1 - 1) + (\hat{\sigma}_2^2/N_2)^2/(N_2 - 1)}$   
↑

- Alternative Hypothesis  $H_A$       Rejection region for  $\alpha$

$\mu_1 - \mu_2 > \Delta_0$  (one-sided right tail)

$$t \geq t_{\alpha, \text{df}}$$

$\mu_1 - \mu_2 < \Delta_0$  (one-sided left tail)

$$t \leq t_{\alpha, \text{df}}$$

$\mu_1 - \mu_2 \neq \Delta_0$  (two-sided)

$$|t| \geq |t_{\alpha/2, \text{df}}|$$

# Welch's t-test

- Null Hypothesis  $H_0: \mu_1 - \mu_2 = \Delta_0$

```
t.test(X1, X2 mu = Δ0,  
alternative="greater")
```

- Test statistic: 
$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}} \sim \text{Student's } t \text{ (df)}$$

- Alternative Hypothesis  $H_A$       Rejection region for  $\alpha$  (in P-VALUES)

$\mu_1 - \mu_2 > \Delta_0$  (one-sided right tail)

$$P(T \geq t \mid H_0) \leq \alpha$$

$\mu_1 - \mu_2 < \Delta_0$  (one-sided left tail)

$$t \leq t_{\alpha, df}$$

$\mu_1 - \mu_2 \neq \Delta_0$  (two-sided)

$$|t| \geq |t_{\alpha/2, df}|$$

# Welch's t-test

- Null Hypothesis  $H_0: \mu_1 - \mu_2 = \Delta_0$

```
t.test(X1, X2 mu = Δ0,  
alternative="less")
```

- Test statistic:  $t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}} \sim \text{Student's } t \text{ (df)}$

- Alternative Hypothesis  $H_A$       Rejection region for  $\alpha$  (in P-VALUES)

$\mu_1 - \mu_2 > \Delta_0$  (one-sided right tail)

$$t \geq t_{\alpha, df}$$

$\mu_1 - \mu_2 < \Delta_0$  (one-sided left tail)

$$P(T \leq t \mid H_0) \leq \alpha$$

$\mu_1 - \mu_2 \neq \Delta_0$  (two-sided)

$$|t| \geq |t_{\alpha/2, df}|$$

# Welch's t-test

- Null Hypothesis  $H_0: \mu_1 - \mu_2 = \Delta_0$

```
t.test(X1, X2 mu = Δ0,  
alternative="two.sided")
```

- Test statistic:  $t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}} \sim \text{Student's } t \text{ (df)}$

- Alternative Hypothesis  $H_A$       Rejection region for  $\alpha$  (in P-VALUES)

$\mu_1 - \mu_2 > \Delta_0$  (one-sided right tail)

$$t \geq t_{\alpha, df}$$

$\mu_1 - \mu_2 < \Delta_0$  (one-sided left tail)

$$t \leq t_{\alpha, df}$$

$\mu_1 - \mu_2 \neq \Delta_0$  (two-sided)

$$P(T \geq |t| \mid H_0) \leq \alpha$$

## Brief note: Student's t-test

- As we said, there exists a less general version of this test, the Student's t-test, that assumes equal variances.
- In this case:  $t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \Delta_0}{\hat{\sigma}_P \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \sim \text{Student's } t \text{ (df)}$



## Brief note: Student's t-test

- As we said, there exists a less general version of this test, the Student's t-test, that assumes equal variances.

- In this case: 
$$t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \Delta_0}{\hat{\sigma}_P \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \sim \text{Student's } t \text{ (df)}$$

*estimated pooled standard deviation*  $\nearrow \hat{\sigma}_P$

$\downarrow$   
 $N_1 + N_2 - 2$

## Brief note: Student's t-test

- As we said, there exists a less general version of this test, the Student's t-test, that assumes equal variances.

- In this case: 
$$t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \Delta_0}{\hat{\sigma}_P \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \sim \text{Student's } t \text{ (df)}$$

*estimated pooled standard deviation*       $N_1 + N_2 - 2$

`t.test(X1, X2 mu = Δ0, var.equal = TRUE, alternative = "greater")`

`t.test(X1, X2 mu = Δ0, var.equal = TRUE, alternative = "less")`

`t.test(X1, X2 mu = Δ0, var.equal = TRUE, alternative = "two.sided")`

# What if our data is longitudinal?

- What if our two samples include the same subjects? This is a typical situation in **longitudinal studies**.
- Example: “Does physical exercise affect brain activity?”
- Here, we can’t apply the usual two sample t-test, since this assumes that observations are independent.
- Instead, we can run a **paired sample t-test**.
- This test is just a **one sample t-test** applied to the **difference** between the two samples,  $D_i = X_{i1} - X_{i2}$ .

# Reminder: One sample t-test

➤ Null Hypothesis  $H_0: \mu = \mu_0$

➤ Test statistic:  $t = \frac{\langle X \rangle - \mu_0}{\hat{\sigma} / \sqrt{N}}$

➤ Alternative Hypothesis  $H_A$

$\mu > \mu_0$  (one-sided right tail)

$\mu < \mu_0$  (one-sided left tail)

$\mu \neq \mu_0$  (two-sided)

# Paired sample t-test $\equiv$ One sample t-test on differences

➤ Null Hypothesis  $H_0: \mu_D = \mu_{D0}$

➤ Test statistic:  $t = \frac{\langle D \rangle - \mu_{D0}}{\hat{\sigma}_D / \sqrt{N}}$

➤ Alternative Hypothesis  $H_A$

$\mu_D > \mu_{D0}$  (one-sided right tail)

$\mu_D < \mu_{D0}$  (one-sided left tail)

$\mu_D \neq \mu_{D0}$  (two-sided)

# Paired sample t-test

- Null Hypothesis  $H_0: \mu_D = \mu_{D0}$
- Test statistic:  $t = \frac{\langle D \rangle - \mu_{D0}}{\hat{\sigma}_D / \sqrt{N}} \sim \text{Student's } t \text{ (df=N-1)}$
- Alternative Hypothesis  $H_A$ 
  - $\mu_D > \mu_{D0}$  (one-sided right tail)
  - $\mu_D < \mu_{D0}$  (one-sided left tail)
  - $\mu_D \neq \mu_{D0}$  (two-sided)

# Paired sample t-test

➤ Null Hypothesis  $H_0: \mu_D = \mu_{D0}$

➤ Test statistic:  $t = \frac{\langle D \rangle - \mu_{D0}}{\hat{\sigma}_D / \sqrt{N}} \sim \text{Student's } t \text{ (df=N-1)}$

➤ Alternative Hypothesis  $H_A$       Rejection region for  $\alpha$

$\mu_D > \mu_{D0}$  (one-sided right tail)       $t \geq t_{\alpha, N-1}$

$\mu_D < \mu_{D0}$  (one-sided left tail)       $t \leq t_{\alpha, N-1}$

$\mu_D \neq \mu_{D0}$  (two-sided)       $|t| \geq |t_{\alpha/2, N-1}|$

# Paired sample t-test

- Null Hypothesis  $H_0: \mu_D = \mu_{D0}$

- Test statistic:  $t = \frac{\langle D \rangle - \mu_{D0}}{\hat{\sigma}_D / \sqrt{N}}$

```
t.test(X1, X2, mu =  $\mu_{D0}$ ,  
paired = TRUE,  
alternative="greater")
```

- Alternative Hypothesis  $H_A$       Rejection region for  $\alpha$  (in P-VALUES)

$\mu_D > \mu_{D0}$  (one-sided right tail)

$$P(T \geq t \mid H_0) \leq \alpha$$

$\mu_D < \mu_{D0}$  (one-sided left tail)

$$t \leq t_{\alpha, N-1}$$

$\mu_D \neq \mu_{D0}$  (two-sided)

$$|t| \geq |t_{\alpha/2, N-1}|$$



# Paired sample t-test

- Null Hypothesis  $H_0: \mu_D = \mu_{D0}$

- Test statistic: 
$$t = \frac{\langle D \rangle - \mu_{D0}}{\hat{\sigma}_D / \sqrt{N}}$$

```
t.test(X1, X2, mu =  $\mu_{D0}$ ,  
paired = TRUE,  
alternative="less")
```

- Alternative Hypothesis  $H_A$       Rejection region for  $\alpha$  (in P-VALUES)

$\mu_D > \mu_{D0}$  (one-sided right tail)

$$t \geq t_{\alpha, N-1}$$

$\mu_D < \mu_{D0}$  (one-sided left tail)

$$P(T \leq t \mid H_0) \leq \alpha$$

$\mu_D \neq \mu_{D0}$  (two-sided)

$$|t| \geq |t_{\alpha/2, N-1}|$$

# Paired sample t-test

- Null Hypothesis  $H_0: \mu_D = \mu_{D0}$

- Test statistic:  $t = \frac{\langle D \rangle - \mu_{D0}}{\hat{\sigma}_D / \sqrt{N}}$

```
t.test(X1, X2, mu =  $\mu_{D0}$ ,  
paired = TRUE,  
alternative="two.sided")
```

- Alternative Hypothesis  $H_A$       Rejection region for  $\alpha$  (in P-VALUES)

$\mu_D > \mu_{D0}$  (one-sided right tail)

$$t \geq t_{\alpha, N-1}$$

$\mu_D < \mu_{D0}$  (one-sided left tail)

$$t \leq t_{\alpha, N-1}$$

$\mu_D \neq \mu_{D0}$  (two-sided)

$$P(T \geq |t| \mid H_0) \leq \alpha$$

# one way ANOVA test

- **Categorical variable:**  $j$  levels (  $\geq 2$  )  $\rightarrow j$  subsets of the **continuous** variable  $X_j$ .
- Why one-way? Only one categorical variable (More? in a future lesson...)
- It tests whether the **means** of the  $j$  populations,  $\mu_j$ , differ.

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

$H_A$ : at least two of the  $\mu_j$ 's are  
different

- Example: “Is the average activation of the insula different across healthy, Autistic and Alzheimer subjects?”

# Several levels in your categorical variable

Inferences based on the relation between two variables

Research question:

Are there  
differences in ACT  
scores between  
education levels?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

# Several levels in your categorical variable

Inferences based on the relation between two variables

Research question:

Are there  
differences in ACT  
scores between  
education levels?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

$X_1$



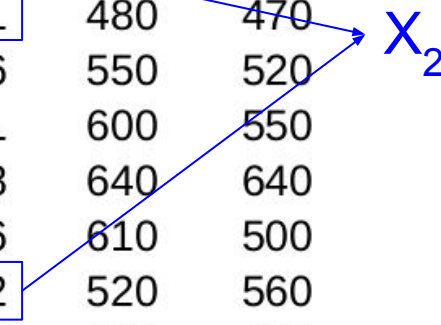
# Several levels in your categorical variable

Inferences based on the relation between two variables

Research question:

Are there  
differences in ACT  
scores between  
education levels?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800



# Several levels in your categorical variable

Inferences based on the relation between two variables

Research question:

Are there  
differences in ACT  
scores between  
education levels?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

$X_3$

# Several levels in your categorical variable

Inferences based on the relation between two variables

Research question:

Are there  
differences in ACT  
scores between  
education levels?

gender	education	age	ACT	SATV	SATQ
2		3	19	24	500
2		3	23	35	600
2		3	20	21	480
1		4	27	26	550
1		2	33	31	600
1		5	26	28	640
2		5	30	36	610
1		3	19	22	520
2		4	23	22	400
2		5	40	35	730

$X_4$



# Several levels in your categorical variable

Inferences based on the relation between two variables

Research question:

Are there  
differences in ACT  
scores between  
education levels?

gender	education	age	ACT	SATV	SATQ	
2	3	19	24	500	500	$X_1$
2	3	23	35	600	500	
2	3	20	21	480	470	$X_2$
1	4	27	26	550	520	
1	2	33	31	600	550	$X_3$
1	5	26	28	640	640	
2	5	30	36	610	500	$X_4$
1	3	19	22	520	560	
2	4	23	22	400	600	$X_4$
2	5	40	35	730	800	

# Several levels in your categorical variable

Inferences based on the relation between two variables

## Change of notation:

- $X_{ij}$  = observation  $i$  in category  $j$
- $n_j$  = number of observations in category  $j$
- $N$  = total number of observations

gender	education	age	ACT	SATV	SATQ	
2	3	19	24	500	500	$X_1$
2	3	23	35	600	500	
2	3	20	21	480	470	$X_2$
1	4	27	26	550	520	
1	2	33	31	600	550	$X_3$
1	5	26	28	640	640	
2	5	30	36	610	500	$X_4$
1	3	19	22	520	560	
2	4	23	22	400	600	
2	5	40	35	730	800	

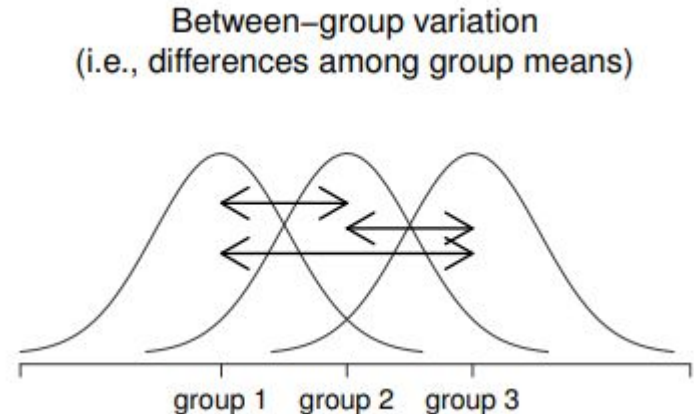
# one way ANOVA test

- It is based on separating the **total variance**,  $V_{\text{tot}}$ , in the continuous variable into two terms: the **between-group variance**,  $V_B$ , and the **within groups variances**,  $V_w$ , i.e.  $V_{\text{tot}} = V_b + V_w$
- It quantifies changes in between-group variation with respect the within-group variation.
- Our statistic will be  $\sim$  between-group variance/within-group variance

# Between-group variance

- It measures how **separated** each category level's data are.
- It is calculated as the differences of the means in each category  $j$   $\langle X_j \rangle$  with respect to the total (also known as grand) mean  $\langle X \rangle$ .

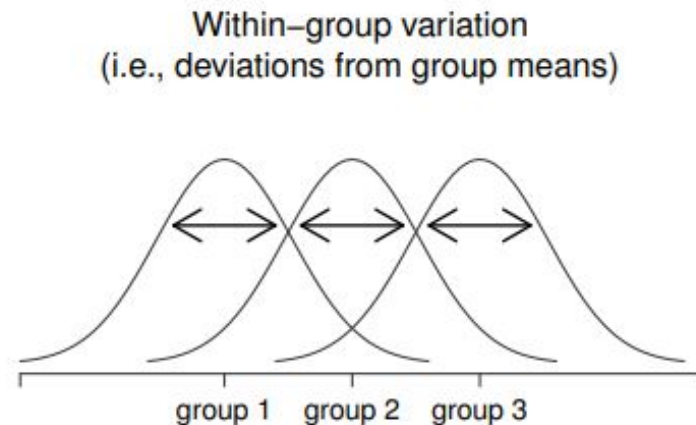
$$V_b = \sum_{j=1}^k N_j (\langle X_j \rangle - \langle X \rangle)^2$$



# Within-group variance

- It measures the spread of the data **within** each category level.
- It is just the summation of the variances in each category level.

$$V_w = \sum_{j=1}^k \sum_{i=1}^{N_j} (X_{ij} - \langle X_j \rangle)^2$$



# one way ANOVA test

## Assumptions:

1. Independence: Observations are not correlated with each other (e.g. in *cross-sectional* studies), both **within and between** samples  $X_{ij}$ .
2. Normality. Values in  $X_{ij}$  follow a gaussian distribution, or their sample sizes are big enough (thanks, Central Limit Theorem!)
3. Homoscedasticity. The variance of data across  $X_{ij}$  should be the same.

# one way ANOVA test

➤ Null Hypothesis  $H_0$ :  $\mu_1 = \mu_2 = \dots = \mu_j$

➤ Test statistic:  $F = \frac{\text{between-group variance}/df_b}{\text{within-group variance}/df_w}$

➤ Alternative Hypothesis  $H_A$

at least two of the  $\mu_j$ 's are different

# one way ANOVA test

➤ Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_j$

➤ Test statistic:  $F = \frac{V_b / df_b}{V_w / df_w}$

➤ Alternative Hypothesis  $H_A$   
at least two of the  $\mu_j$ 's are different



# one way ANOVA test

➤ Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_j$

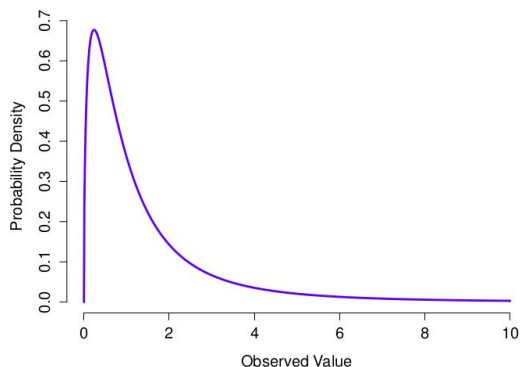
➤ Test statistic:  $F = \frac{V_b/df_b}{V_w/df_w} \sim \text{F-distribution } (df_b, df_w)$

➤ Alternative Hypothesis  $H_A$

at least two of the  $\mu_j$ 's are different

# Reminder: F-distribution

- It is related to the  $\chi^2$  distribution; specifically as the **ratio** between two  $\chi^2$  statistics.  $\theta_i = \{df_1, df_2\}$
- It usually arises as the ratio between variances. This ratio is common in testing **mean differences** across groups (ANOVA test).



# one way ANOVA test

- Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_j$
- Test statistic:  $F = \frac{V_b/df_b}{V_w/df_w} \sim \text{F-distribution } (df_b, df_w)$
- Alternative Hypothesis  $H_A$   
at least two of the  $\mu_j$ 's are different

# one way ANOVA test

➤ Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_j$

➤ Test statistic:  $F = \frac{V_b / df_b}{V_w / df_w} \sim \text{F-distribution } (df_b, df_w)$

Degrees of freedom

$$df_b = k - 1$$

$$df_w = N - k$$

➤ Alternative Hypothesis  $H_A$

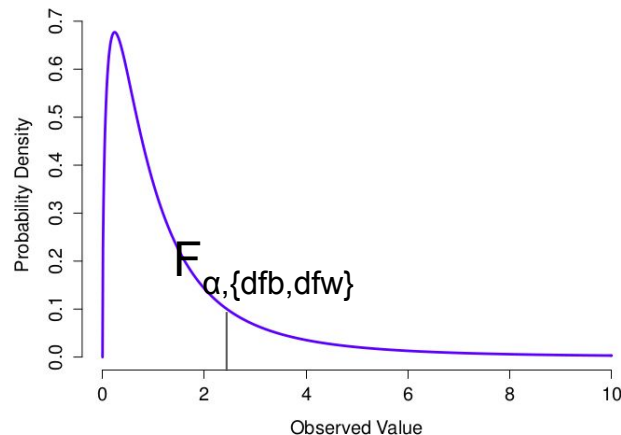
at least two of the  $\mu_j$ 's are different

# one way ANOVA test

➤ Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_i$

➤ Test statistic:  $F = \frac{V_b/df_b}{V_w/df_w}$

➤ Alternative Hypothesis  $H_A$   
at least two of the  $\mu_j$ 's are different



Rejection region for  $\alpha$

$$F \geq F_{\alpha, \{df_b, df_w\}}$$

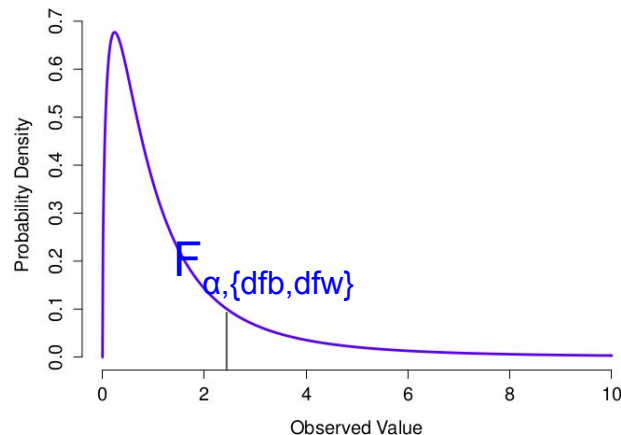
# one way ANOVA test

➤ Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_i$

➤ Test statistic:  $F = \frac{V_b / df_b}{V_w / df_w}$

$$qf(\alpha, df_b, df_w)$$

➤ Alternative Hypothesis  $H_A$   
at least two of the  $\mu_i$ 's are different



Rejection region for  $\alpha$

$$F \geq F_{\alpha, \{df_b, df_w\}}$$

# one way ANOVA test

➤ Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_i$

➤ Test statistic:  $F = \frac{V_b / df_b}{V_w / df_w}$

```
pf(F, df_b, df, lower.tail =  
FALSE)
```

➤ Alternative Hypothesis  $H_A$   
at least two of the  $\mu_i$ 's are different

Rejection region for  $\alpha$  (in P-VALUES)

$P(f > F \mid H_0)$

# one way ANOVA test

➤ Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_i$

➤ Test statistic:  $F = \frac{V_b / df_b}{V_w / df_w}$

`aov(formula, data)`

➤ Alternative Hypothesis  $H_A$   
at least two of the  $\mu_i$ 's are different

Rejection region for  $\alpha$  (in P-VALUES)

$P(f > F \mid H_0)$



# one way ANOVA test

➤ Null Hypothesis  $H_0: \mu_1 = \mu_2 = \dots = \mu_i$

➤ Test statistic:  $F = \frac{V_b / df_b}{V_w / df_w}$

➤ Alternative Hypothesis  $H_A$   
at least two of the  $\mu_i$ 's are different

TUTORIAL!!!

`aov(formula, data)`

Rejection region for  $\alpha$  (in P-VALUES)

$P(f > F \mid H_0)$

# one way ANOVA test $\leftrightarrow$ two sample t-test

- When testing if the means differences in **two categories** are different from zero ( $\Delta_0=0$ ), an **ANOVA is similar to the Student t-test**.
- In this case:  $F = t^2$
- In the end, the statistic is almost similar to a ratio between variances as well.

$$t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \cancel{\Delta_0}}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

# one way ANOVA test $\leftrightarrow$ two sample t-test

- When testing if the means differences in **two categories** are different from zero ( $\Delta_0=0$ ), an **ANOVA is similar to the Student t-test**.
- In this case:  $F = t^2$
- In the end, the statistic is almost similar to a ratio between variances as well.

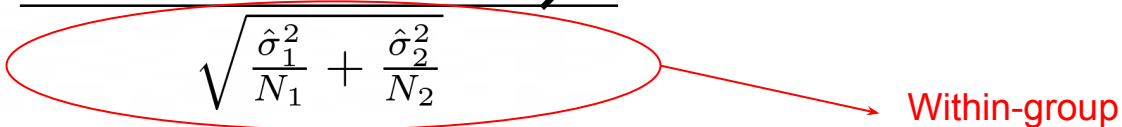
$$t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \cancel{\Delta_0}}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

Between-group

# one way ANOVA test $\leftrightarrow$ two sample t-test

- When testing if the means differences in **two categories** are different from zero ( $\Delta_0=0$ ), an **ANOVA is similar to the Student t-test**.
- In this case:  $F = t^2$
- In the end, the statistic is almost similar to a ratio between variances as well.

$$t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \cancel{\Delta_0}}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

 Within-group

# one way ANOVA test $\leftrightarrow$ two sample t-test

- When testing if the means differences in **two categories** are different from zero ( $\Delta_0=0$ ), an **ANOVA is similar to the Student t-test**.
- In this case:  $F = t^2$
- In the end, the statistic is almost similar to a ratio between variances as well.

$$t = \frac{(\langle X_1 \rangle - \langle X_2 \rangle) - \cancel{\Delta_0}}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

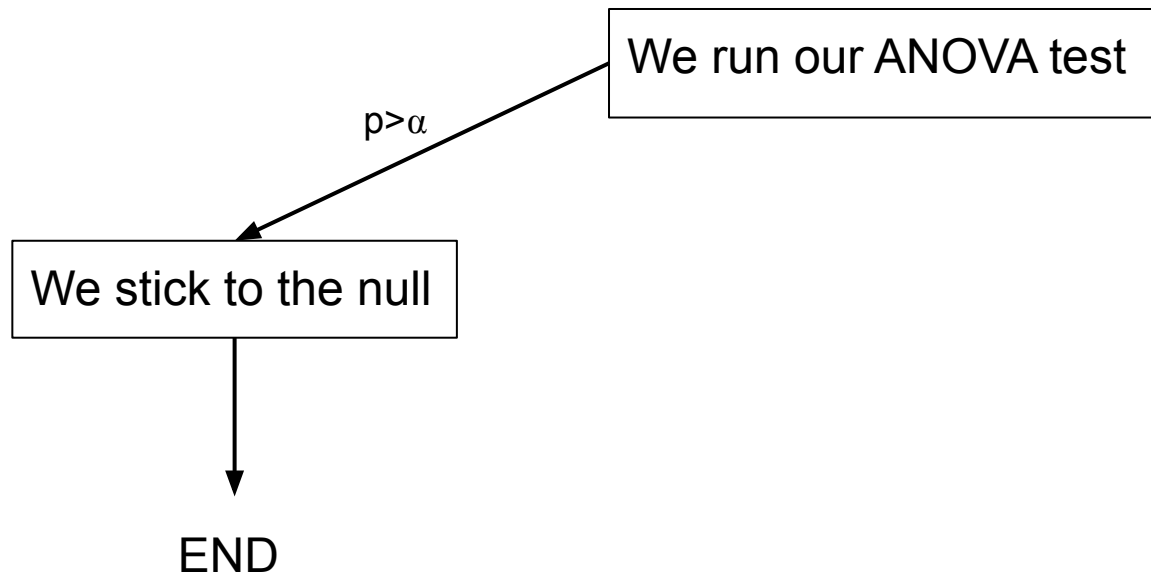
Between-group

Within-group

# ANOVA rejects the null: what next?

We run our ANOVA test

# ANOVA rejects the null: what next?



# ANOVA rejects the null: what next?

We run our ANOVA test

$p \leq \alpha$

We reject the null  
(at least two of the  $\mu_j$ 's are different)



# ANOVA rejects the null: what next?

We run our ANOVA test

$p \leq \alpha$

We reject the null  
(at least two of the  $\mu_j$ 's are different)

Which ones?

# ANOVA rejects the null: what next?

We run our ANOVA test

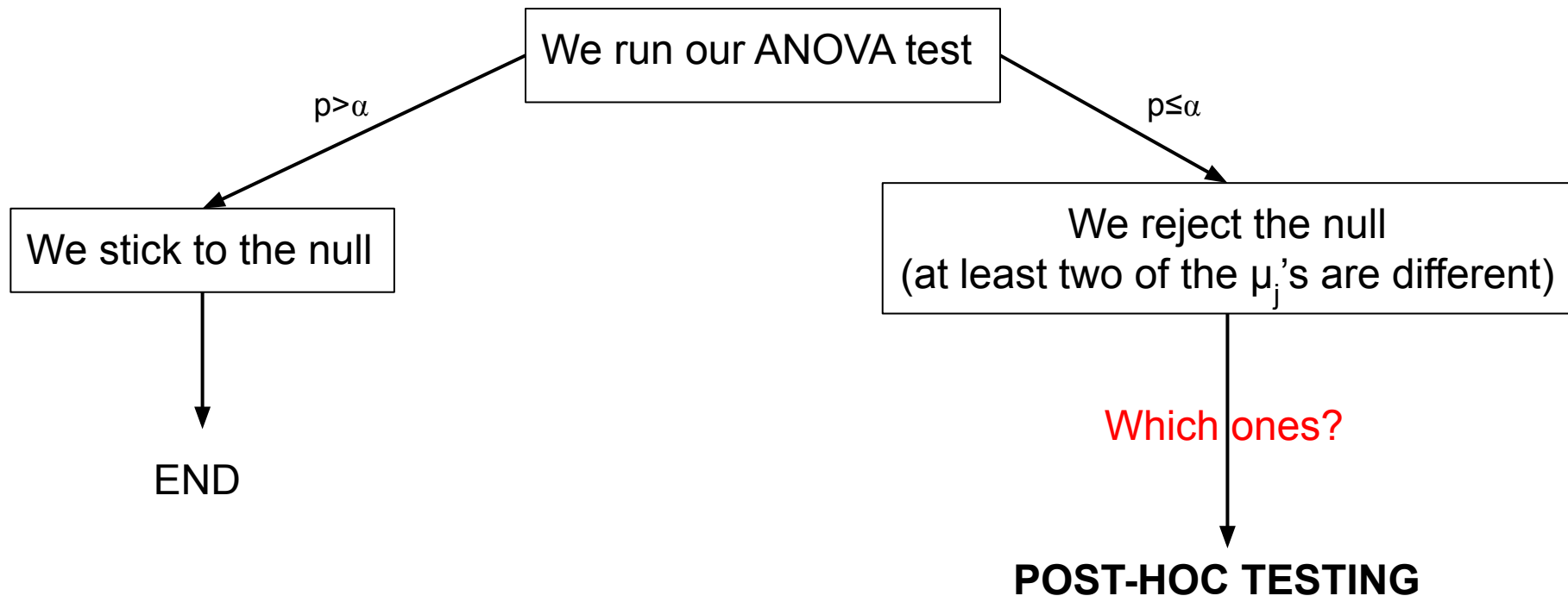
$p \leq \alpha$

We reject the null  
(at least two of the  $\mu_j$ 's are different)

Which ones?

**POST-HOC TESTING**

# ANOVA rejects the null: what next?

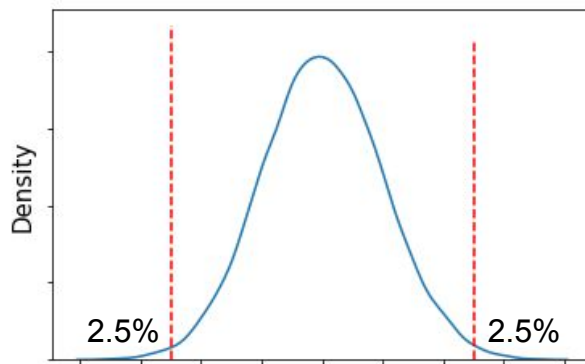


# ANOVA rejects the null: Post-hoc testing

- Post hoc (“after this”) testing involves performing **new** statistical analyses **after** the data have already been seen.
- Typical in one-way ANOVA to test **which pairs** of means are **significant**.
- A lot of care needs to be taken → **Multiple testing inflates Type I error  $\alpha$**

# ANOVA rejects the null: Post-hoc testing

- Post hoc (“after this”) testing involves performing **new** statistical analyses **after** the data have already been seen.
- Typical in one-way ANOVA to test **which pairs** of means are **significant**.
- A lot of care needs to be taken → **Multiple testing inflates Type I error  $\alpha$**  (remember when we adjusted  $\alpha$  when testing both tails?)



# ANOVA rejects the null: Post-hoc testing

One very simple recipe:

First, **pairwise** two sample (Student's) t-tests; then, **Bonferroni** procedure that keeps Type I error under control (We'll come back to this in the future).

```
pairwise.t.test(x=Our continuous variable,  
                g=Our categorical variable,  
                p.adjust.method="bonf")
```

# Recap

- A categorical variable usually encodes subpopulations we may want to draw conclusions about.
- If talking about a continuous estimation (e.g. the mean), we may use a **two-sample t-test** (2 populations) or a **one-way ANOVA** ( $\geq 2$  populations).
- If more than 2 populations, we may need to run **post-hoc analyses** followed by a procedure to keep Type I error,  $\alpha$ , under control.
- In the next lectures, we will cover the relationship between **pairs of categorical variables** and **pairs of continuous variables**.