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 (>= two categories).

Two levels in your categorical variable

<u>Research question:</u>
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

<u>Research question:</u>
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	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_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 Δ_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 Δ_0.
- <u>Example</u>: “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. <u>Independence</u>: Observations are not correlated with each other (e.g. in *cross-sectional* studies), both **within and between** samples X_1 and X_2.
2. <u>Normality.</u> 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. $\boldsymbol{\sigma_1 \neq \sigma_2}$.

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. $\boldsymbol{\sigma_1 = \sigma_2}$, corresponds to the **Student's t-test.**

Welch's t-test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 - \mu_2 = \Delta_0$

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

- Alternative Hypothesis $\mathbf{H_A}$

  $\mu_1 - \mu_2 > \Delta_0$ (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 $\mathbf{H_0}$: $\mu_1 - \mu_2 = \Delta_0$

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

- Alternative Hypothesis $\mathbf{H_A}$

  $\mu_1 - \mu_2 > \Delta_0$ (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 $\mathbf{H_0}$: $\mu_1 - \mu_2 = \Delta_0$

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

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

- Alternative Hypothesis $\mathbf{H_A}$

  $\mu_1 - \mu_2 > \Delta_0$ (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 $\mathbf{H_0}$: $\mu_1 - \mu_2 = \Delta_0$

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

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

- Alternative Hypothesis $\mathbf{H_A}$ — <u>Rejection region for α</u>

Alternative Hypothesis $\mathbf{H_A}$	Rejection region for α
$\mu_1 - \mu_2 > \Delta_0$ (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)	$\lvert t \rvert \geq \lvert t_{\alpha/2,df} \rvert$

Welch's t-test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 - \mu_2 = \Delta_0$

t.test(X_1, X_2 mu = Δ_0, alternative="greater")

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

- Alternative Hypothesis $\mathbf{H_A}$ — Rejection region for α (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)	$\lvert t \rvert \geq \lvert t_{\alpha/2,df} \rvert$

Welch's t-test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 - \mu_2 = \Delta_0$

t.test(X_1, X_2 mu = Δ_0, alternative="less")

- Test statistic: $t = \dfrac{(<X_1> - <X_2>) - \Delta_0}{\sqrt{\dfrac{\hat{\sigma}_1^2}{N_1} + \dfrac{\hat{\sigma}_2^2}{N_2}}}$ ~ Student's t (df)

- Alternative Hypothesis $\mathbf{H_A}$ — Rejection region for α (in P-VALUES)

Alternative Hypothesis $\mathbf{H_A}$	Rejection region for α (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)	$\lvert t \rvert \geq \lvert t_{\alpha/2,df} \rvert$

Welch's t-test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 - \mu_2 = \Delta_0$

t.test(X_1, X_2 mu = Δ_0, alternative="two.sided")

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

- Alternative Hypothesis $\mathbf{H_A}$ — Rejection region for α (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: $t = \dfrac{(< X_1 > - < X_2 >) - \Delta_0}{\hat{\sigma}_P \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$ ~ Student's t (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: $t = \dfrac{(< X_1 > - < X_2 >) - \Delta_0}{\hat{\sigma}_P \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$ ~ Student's t (df)

estimated pooled standard deviation

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: $t = \dfrac{(< X_1 > - < X_2 >) - \Delta_0}{\hat{\sigma}_P \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$ ~ Student's t (df)

estimated pooled standard deviation

N_1+N_2-2

t.test(X_1, X_2 mu = Δ_0, **var.equal = TRUE**, alternative = "greater")

t.test(X_1, X_2 mu = Δ_0, **var.equal = TRUE**, alternative = "less")

t.test(X_1, X_2 mu = Δ_0, **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 $\mathbf{H_0}$: $\mu=\mu_0$
- Test statistic: $t = \dfrac{< X > - \mu_0}{\hat{\sigma}/\sqrt{N}}$
- Alternative Hypothesis $\mathbf{H_A}$

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

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

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

Paired sample t-test ≣ One sample t-test on differences

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

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

- Alternative Hypothesis $\mathbf{H_A}$

  $\mu_D > \mu_{D0}$ (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 $\mathbf{H_0}$: $\mu_D = \mu_{D0}$

- Test statistic: $t = \dfrac{< D > -\mu_{D0}}{\hat{\sigma}_D/\sqrt{N}}$ ~ Student's t (df=N-1)

- Alternative Hypothesis $\mathbf{H_A}$

  $\mu_D > \mu_{D0}$ (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 $\mathbf{H_0}$: $\mu_D = \mu_{D0}$

- Test statistic: $t = \dfrac{<D> - \mu_{D0}}{\hat{\sigma}_D/\sqrt{N}}$ ~ Student's t (df=N-1)

- Alternative Hypothesis $\mathbf{H_A}$ — Rejection region for α

Alternative Hypothesis $\mathbf{H_A}$	Rejection region for α
$\mu_D > \mu_{D0}$ (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)	$\lvert t \rvert \geq \lvert t_{\alpha/2,N-1} \rvert$

Paired sample t-test

- Null Hypothesis **H_0**: $\mu_D = \mu_{D0}$
- Test statistic: $t = \dfrac{< D > - \mu_{D0}}{\hat{\sigma}_D / \sqrt{N}}$

t.test(X_1, X_2, mu = μ_{D0},
paired = TRUE,
alternative="greater")

- Alternative Hypothesis **H_A** — Rejection region for α (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)	$\lvert t \rvert \geq \lvert t_{\alpha/2,N-1} \rvert$

Paired sample t-test

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

- Test statistic: $t = \dfrac{< D > - \mu_{D0}}{\hat{\sigma}_D / \sqrt{N}}$

t.test(X_1, X_2, mu = μ_{D0},
paired = TRUE,
alternative="less")

- Alternative Hypothesis $\mathbf{H_A}$ — Rejection region for α (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)	$\lvert t \rvert \geq \lvert t_{\alpha/2,N-1} \rvert$

Paired sample t-test

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

t.test(X_1, X_2, mu = μ_{D0},
paired = TRUE,
alternative="two.sided")

- Test statistic: $t = \dfrac{<D> - \mu_{D0}}{\hat{\sigma}_D/\sqrt{N}}$

- Alternative Hypothesis $\mathbf{H_A}$ — <u>Rejection region for α</u> (in P-VALUES)

Alternative Hypothesis	Rejection region
$\mu_D > \mu_{D0}$ (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 (>= 2) → *j* subsets of the **continuous** variable $\mathbf{X}_j$.
- Why one-way? Only one categorical variable (More? in a future lesson…)
- It tests whether the **means** of the *j* populations, μ_j, differ.

$$\mathbf{H_0}: \mu_1 = \mu_2 = \ldots = \mu_j$$

$\mathbf{H_A}$: at least two of the μ_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

<u>Research question:</u>
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	$\boxed{2}$	33	$\boxed{31}$	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

X_2

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

<u>Research question:</u>
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	$\boxed{5}$	26	$\boxed{28}$	640	640
2	$\boxed{5}$	30	$\boxed{36}$	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	$\boxed{5}$	40	$\boxed{35}$	730	800

X_4

Several levels in your categorical variable

Inferences based on the relation between two variables

<u>Research question:</u>
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

X_2

X_3

X_4

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
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

X_3

X_4

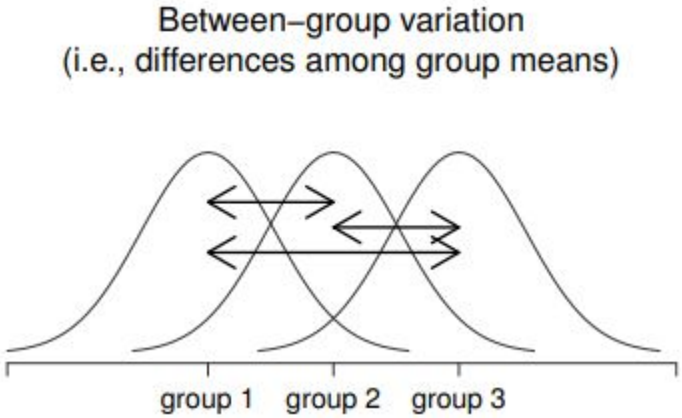
one way ANOVA test

- It is based on separating the **total variance,** V_{tot}**,** in the continuous variable into two terms: the **between-group variance,** V_B, and the **within groups variances,** V_w, i.e. $\mathbf{V_{tot} = V_b + V_w}$
- It quantifies changes in between-group variation with respect the within-group variation.
- Our statistic will be ~ 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* $<X_j>$ with respect to the total (also known as grand) mean $<X>$.

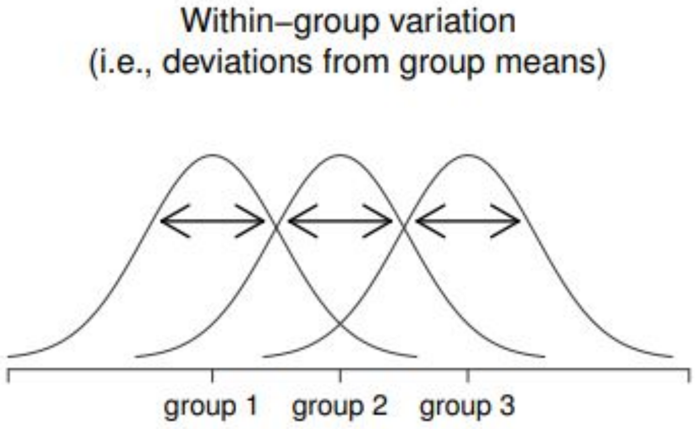
$$V_b = \sum_{j=1}^{k} N_j \left(< X_j > - < X >\right)^2$$



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.

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



one way ANOVA test

Ass**umptions:**

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 $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_j$

- Test statistic: $F = \dfrac{\text{between-group variance}/\mathrm{df}_b}{\text{within-group variance}/\mathrm{df}_w}$

- Alternative Hypothesis $\mathbf{H_A}$

  at least two of the μ_j's are different

one way ANOVA test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_j$

- Test statistic: $F = \dfrac{V_b/\mathrm{df}_b}{V_w/\mathrm{df}_w}$

- Alternative Hypothesis $\mathbf{H_A}$

  at least two of the μ_j's are different

one way ANOVA test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_j$

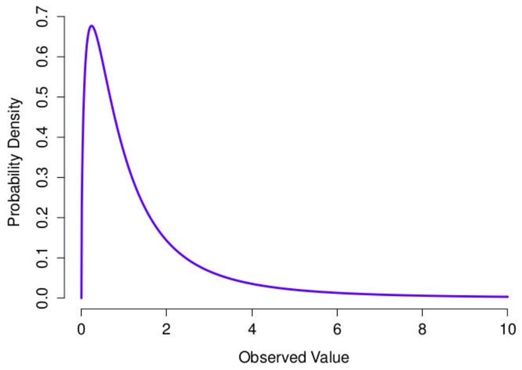
- Test statistic: $F = \dfrac{V_b/\mathrm{df}_b}{V_w/\mathrm{df}_w}$ ~ F-distribution (df_b, df_w)

- Alternative Hypothesis $\mathbf{H_A}$

  at least two of the μ_j's are different

Reminder: F-distribution

- It is related to the χ^2 distribution; specifically as the **ratio** between two χ^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 $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_j$

- Test statistic: $F = \dfrac{V_b/\mathrm{df}_b}{V_w/\mathrm{df}_w}$ ~ F-distribution (df_b, df_w)

- Alternative Hypothesis $\mathbf{H_A}$

  at least two of the μ_j's are different

one way ANOVA test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_j$

- Test statistic: $F = \dfrac{V_b/\mathrm{df}_b}{V_w/\mathrm{df}_w}$ ~ F-distribution (df_b, df_w)

<u>Degrees of freedom</u>

$df_w = k-1$

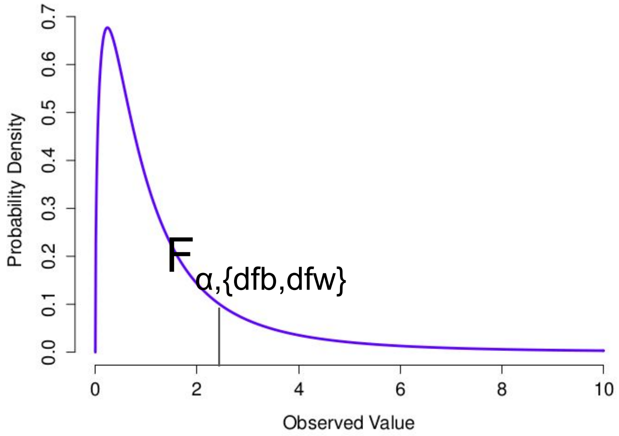
$df_b = N - k$

- Alternative Hypothesis $\mathbf{H_A}$

  at least two of the μ_j's are different

one way ANOVA test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_i$



- Test statistic: $F = \dfrac{V_b/\mathrm{df}_b}{V_w/\mathrm{df}_w}$

- Alternative Hypothesis $\mathbf{H_A}$

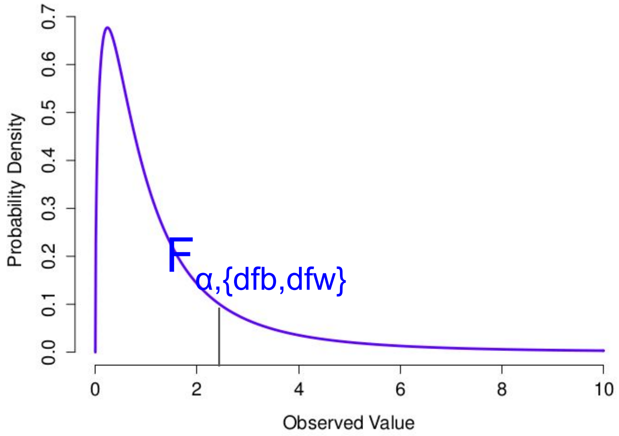
  at least two of the μ_j's are different

<u>Rejection region for α</u>

$F \geq F_{\alpha,\{dfb,dfw\}}$

one way ANOVA test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_i$



- Test statistic: $F = \dfrac{V_b/\mathrm{df}_b}{V_w/\mathrm{df}_w}$

$\boxed{qf(\alpha,\ df_b,\ df_w)}$

- Alternative Hypothesis $\mathbf{H_A}$

  at least two of the μ_i's are different

<u>Rejection region for α</u>

$F \geq F_{\alpha,\{dfb,dfw\}}$

one way ANOVA test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_i$

- Test statistic: $F = \dfrac{V_b/\mathrm{df}_b}{V_w/\mathrm{df}_w}$

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

- Alternative Hypothesis $\mathbf{H_A}$

at least two of the μ_i's are different

Rejection region for α (in P-VALUES)

$P(f > F \mid H_0)$

one way ANOVA test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_i$

- Test statistic: $F = \dfrac{V_b/\mathrm{df}_b}{V_w/\mathrm{df}_w}$

$\boxed{\textit{aov}\text{(formula, data)}}$

- Alternative Hypothesis $\mathbf{H_A}$

  at least two of the μ_i's are different

Rejection region for α (in P-VALUES)

$P(f > F \mid H_0)$

one way ANOVA test

- Null Hypothesis $\mathbf{H_0}$: $\mu_1 = \mu_2 = \ldots = \mu_i$

TUTORIAL!!!

- Test statistic: $F = \dfrac{V_b/\mathrm{df}_b}{V_w/\mathrm{df}_w}$

aov(formula, data)

- Alternative Hypothesis $\mathbf{H_A}$

  at least two of the μ_i's are different

Rejection region for α (in P-VALUES)

$P(f > F \mid H_0)$

one way ANOVA test ↔ two sample t-test

- When testing if the means differences in **two categories** are different from zero (Δ_0=0), an **ANOVA is similar to the Student t-test**.
- In this case: $\boxed{F = t^2}$
- In the end, the statistic is almost similar to a ratio between variances as well.

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

one way ANOVA test ↔ two sample t-test

- When testing if the means differences in **two categories** are different from zero (Δ_0=0), an **ANOVA is similar to the Student t-test**.
- In this case: $\boxed{F = t^2}$
- In the end, the statistic is almost similar to a ratio between variances as well.

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

Between-group

one way ANOVA test ↔ two sample t-test

- When testing if the means differences in **two categories** are different from zero (Δ_0=0), an **ANOVA is similar to the Student t-test**.
- In this case: $\boxed{F = t^2}$
- In the end, the statistic is almost similar to a ratio between variances as well.

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

Within-group

one way ANOVA test ↔ two sample t-test

- When testing if the means differences in **two categories** are different from zero (Δ_0=0), an **ANOVA is similar to the Student t-test**.
- In this case: $\boxed{F = t^2}$
- In the end, the statistic is almost similar to a ratio between variances as well.

$$t = \frac{\underbrace{(<X_1> - <X_2>) - \cancel{\Delta_0}}_{\text{Between-group}}}{\underbrace{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}_{\text{Within-group}}}$$

ANOVA rejects the null: what next?

We run our ANOVA test

ANOVA rejects the null: what next?

We run our ANOVA test

$p > \alpha$

We stick to the null

END

ANOVA rejects the null: what next?

We run our ANOVA test

$p \leq \alpha$

We reject the null
(at least two of the μ_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 μ_j's are different)

Which ones?

ANOVA rejects the null: what next?

We run our ANOVA test

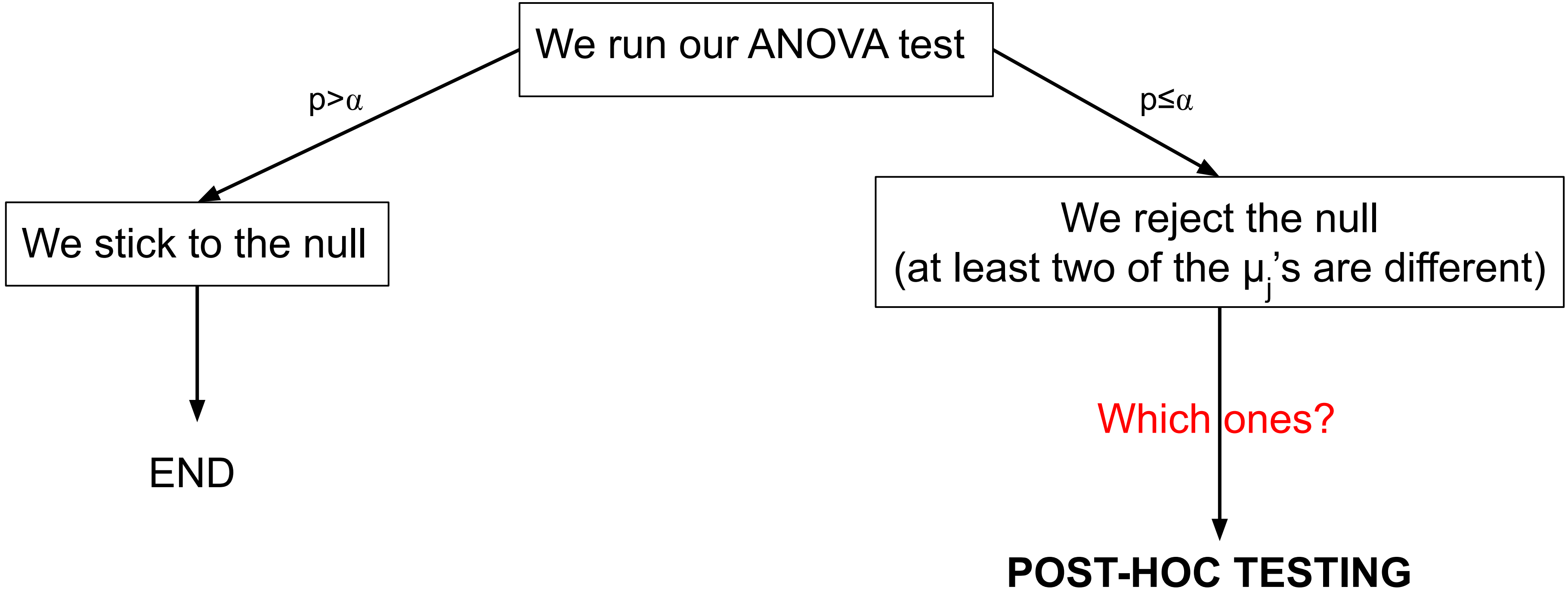
$p \leq \alpha$

We reject the null
(at least two of the μ_j's are different)

Which ones?

POST-HOC TESTING

ANOVA rejects the null: what next?

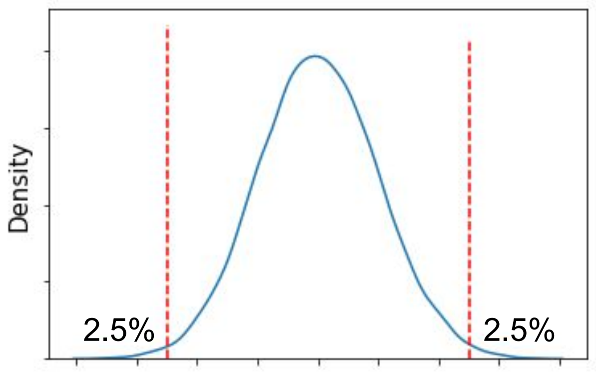


ANOVA rejects the null: Post-hoc testing

- 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 α**

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 α** (remember when we adjusted α 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** (>= 2 populations).
- ➢ If more than 2 populations, we may need to run **post-hoc analyses** followed by a procedure to keep Type I error, α, under control.
- ➢ In the next lectures, we will cover the relationship between **pairs of categorical variables** and **pairs of continuous variables**.