

Kvantitativ metode II: ANOVA

PSY2117 / PSYK4317

Frederick Anyan

Førsteamanuensis i kvantitativ forskningsmetode

[NTNU](#) | [Personlig nettside](#)

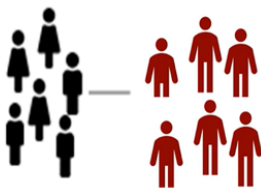
NTNU, Institutt for psykologi

H2026

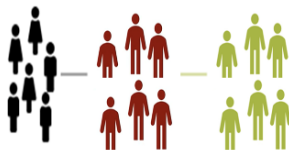
Plan for i dag

- Kort repetisjon: hvorfor t-testen ikke strekker til med flere grupper
- Hva er en ANOVA, og hvilken "familie" av teknikker tilhører den?
- Logikken: $SS_T = SS_M + SS_R$, frihetsgrader, F -forhold
- Forutsetninger
- Praktisk eksempel: Viagra-dose og libido (ekte data, JASP)
- Effektstørrelse (η^2 , ω^2) og post-hoc-tester
- ANOVA som regresjon – broen videre til ANCOVA og faktoriell ANOVA

Kort repetisjon: begrensningen ved t-tester



t-testen: 2 grupper



Men hva med 3+?

- t-testen sammenligner **kun to** gjennomsnitt om gangen.
- Mange studier har tre, fire eller flere grupper/betingelser.
- Løsning som *ikke* fungerer: mange parvise t-tester – Type-I-feilraten eksploderer (jf. forrige forelesning).

Så hva er en ANOVA?

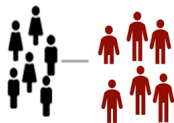
Kjernen

ANOVA studerer forskjeller mellom **tre eller flere** grupper på en kontinuerlig avhengig variabel – ved å sammenligne **varians**, ikke bare gjennomsnitt direkte.

Brukes blant annet til å analysere resultater fra eksperimenter med flere betingelser.

Teknikk	Grupperingsvariabel(-er)	Eksempel
t-test	1 variabel, 2 nivåer	Placebo vs. Viagra
Enveis ANOVA	1 variabel, ≥ 3 nivåer	Placebo / Lav / Høy dose
ANCOVA	1 variabel + kovariat	Dose, kontroll for alder
Faktoriell ANOVA	≥ 2 variabler	Dose $\times$ kjønn
ANOVA m/ rep. målinger	Samme personer, flere tidspunkt	Libido: før / rett etter / oppfølging

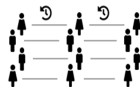
Designfamilien – visualisert



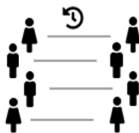
t-test
(2 uavh. grupper)



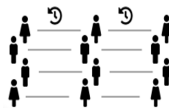
Enveis ANOVA
(≥ 3 uavh. grupper)



Faktoriell / mixed
(between $\times$ within)



Avhengig t-test
(2 tidspunkt, samme pers.)



RM-ANOVA
(≥ 3 tidspunkt, samme pers.)

ANOVA – hvordan?

Logikken

I stedet for å sammenligne grupper direkte, ser vi på **variansen av gruppegjennomsnittene**: er denne større enn vi forventer av ren tilfeldighet?

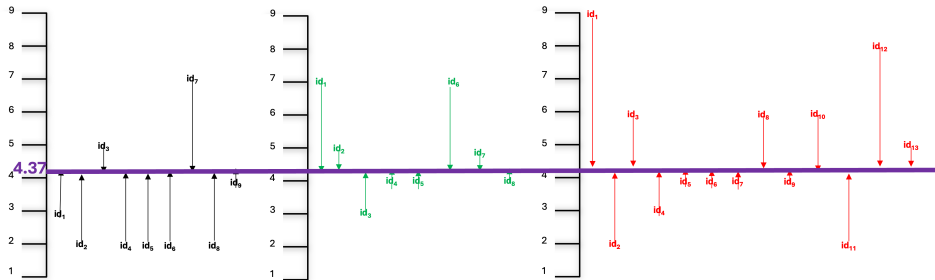
En ANOVA sammenligner to typer varians:

- **Systematisk varians** – forskjeller *mellom* gruppene (skyldes behandlingen/grupperingen)
- **Uforklart varians** – forskjeller *innad i* gruppene (tilfeldig støy, individuelle forskjeller)

Total varians: SS_T

$$SS_T = \sum_{i=1}^N (x_i - \bar{X}_{GM})^2 = 110.95$$

N = totalt antall deltakere | x_i = observasjon for person i | $\bar{X}_{GM}$ = grand mean (gjennomsnitt over alle)

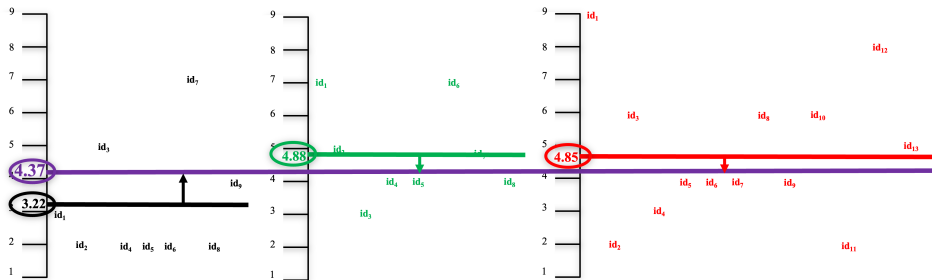


Hver id er **én deltaker** i sin dosegruppe (svart = placebo, grønn = lav dose, rød = høy dose); hvert individs avvik er fra **gjennomsnittet over alle** ($GM = 4.37$)

Systematisk varians: SS_M

$$SS_M = \sum_{k=1}^K n_k (\bar{X}_k - \bar{X}_{GM})^2 = 16.844$$

K = antall grupper | n_k = antall deltakere i gruppe k | $\bar{X}_k$ = gjennomsnitt i gruppe k

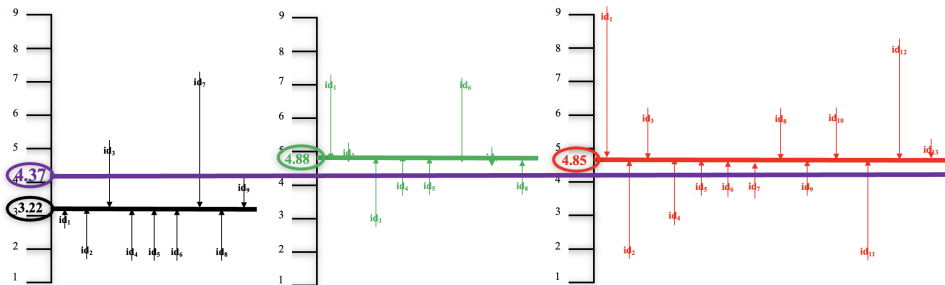


Hver id er **én deltaker** i sin dosegruppe (svart = placebo, grønn = lav dose, rød = høy dose); linjene viser hvert **gruppegjennomsnitt** sitt avvik fra grand mean ($GM = 4.37$)

Uforklart varians: SS_R

$$SS_R = \sum_{k=1}^K \sum_{i=1}^{n_k} (x_{ik} - \bar{X}_k)^2 = 94.123$$

x_{ik} = observasjon for person i i gruppe k | $\bar{X}_k$ = gjennomsnitt i gruppe k

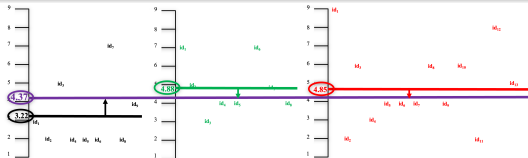


Hver id er **én deltaker** i sin dosegruppe (svart = placebo, grønn = lav dose, rød = høy dose); pilene viser hvert individs avvik fra **sitt eget gruppegjennomsnitt**

ANOVA-terminologi: mellom vs. innad i gruppene

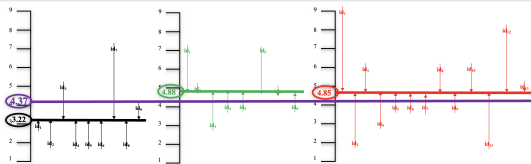
Mellom gruppene

- *Between-group differences*
- SS_M : avvik fra grand mean til hvert gruppegjennomsnitt
- Måler systematisk varians (pga. gruppetilhørighet/behandling)



Innad i gruppene

- *Within-group differences*
- SS_R : avvik fra eget gruppegjennomsnitt til hver observasjon
- Måler uforklart/tilfeldig feilvarians



Hver id er én deltaker i sin dosegruppe – svart = placebo, grønn = lav dose, rød = høy dose.

Fra kvadratsum til gjennomsnittlig kvadratsum

Spørsmålet

SS_M og SS_R avhenger av **hvor mange** grupper og deltakere vi har – flere deltakere gir større kvadratsummer, uavhengig av om effekten faktisk er sterk. Hvordan går vi fra en kvadratsum til et "gjennomsnittlig" mål vi kan sammenligne på tvers av studier?

- Løsning: del hver kvadratsum på sine **frihetsgrader** – resultatet kalles *Mean Square (MS)*.

Frihetsgrader (degrees of freedom)

Frihetsgradene i en ANOVA partisjoneres på samme måte som kvadratsummene:

$$df_T = N - 1, \quad df_M = k - 1, \quad df_R = N - k$$

N = totalt antall deltakere | k = antall grupper

Med våre data ($N = 30$, $k = 3$ grupper)

$$df_T = 29, \quad df_M = 2, \quad df_R = 27$$

Mean Squares (MS)

Kvadratsummene avhenger av antall observasjoner – vi "gjennomsnitter" dem med frihetsgradene:

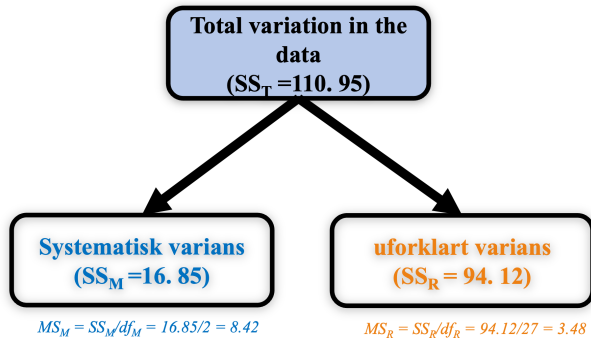
$$MS_M = \frac{SS_M}{df_M}, \quad MS_R = \frac{SS_R}{df_R}$$

MS_M = gjennomsnittlig systematisk varians | MS_R = gjennomsnittlig uforklart varians (feilvariens)

Innsatt

$$MS_M = \frac{16.844}{2} = 8.422, \quad MS_R = \frac{94.123}{27} = 3.486$$

Fra kvadratsum til variansforhold



Hvor mye varians er **systematisk** (pga gruppetilhørighet) og hvor mye er «tilfeldig» feilvariens?

F-testen: formel

$$F = \frac{MS_M}{MS_R}$$

Innsatt

$$F(2, 27) = \frac{8.422}{3.486} = 2.416$$

- $df_M = \text{antall grupper} - 1$ $df_R = \text{antall deltakere} - \text{antall grupper}$
- Hvis eksperimentet er "vellykket", vil SS_M (og dermed MS_M) være stor relativt til MS_R , og F blir stor.
- F -fordelingen avhenger av *to* typer frihetsgrader (df_M , df_R) – i motsetning til t -fordelingen som bare har én.

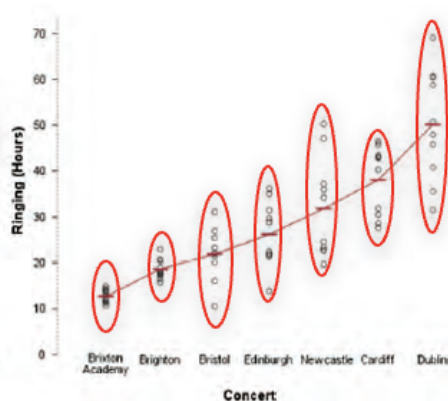
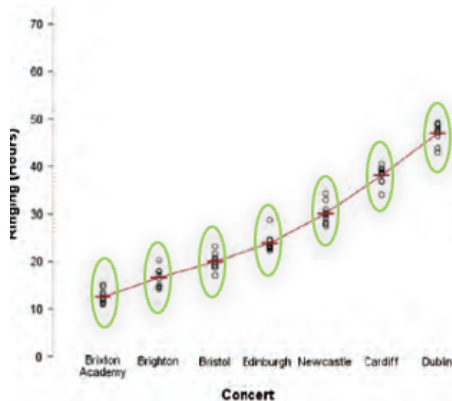
Forutsetninger for ANOVA

Forutsetning	Hvordan sjekke
Uavhengige feilledd (statistisk uavhengighet – én observasjon forteller ingenting om en annen)	Studiedesign – ikke testbart i etterkant
Lik varians på tvers av gruppene (homoskedastisitet)	Levene's test; sammenlign standardavvik
Normalfordelte data innad i hver gruppe	QQ-plott; Shapiro–Wilk eller Kolmogorov–Smirnov
Tilfeldig utvalg / tilfeldig tildeling til grupper	Studiedesign

Hva hvis forutsetningene brytes?

Ulik varians: bruk Welch- eller Brown–Forsythe-korrigert ANOVA (JASP kan rapporterer begge, akkurat som Welch's t-test). Klart ikke-normalfordelte data: vurder *Kruskal–Wallis H-test* som ikke-parametrisk alternativ.

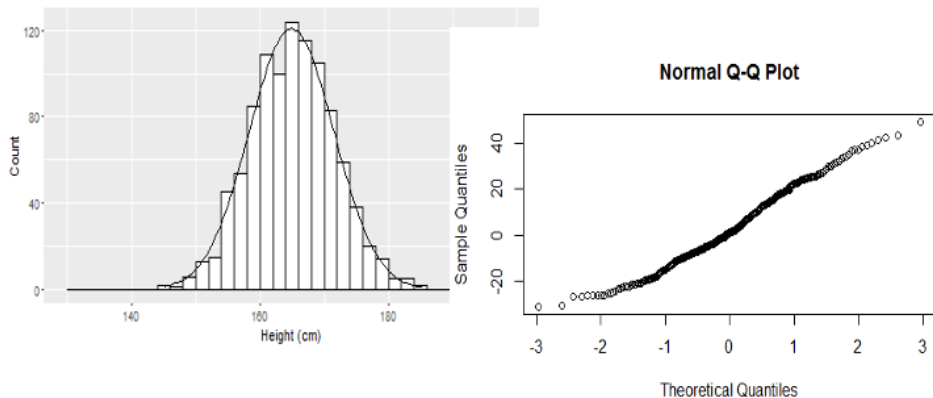
Homogenitet av varians – visualisert



Ringetid for gitarforsterkere etter sju ulike konserter (Field).

Grønt: lik spredning på tvers av grupper – **homogeneity of variance**. Rødt: spredningen øker med gjennomsnittet – **heterogeneity of variance**.

Normalitet – visualisert



Histogram og QQ-plott for en tilnærmet normalfordelt variabel (høyde, cm).



Enveis ANOVA

Skiller mer enn to gruppegjennomsnitt seg fra hverandre?

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

$$H_1 : \text{minst én } \mu_i \neq \mu_j$$

Eksempel: Viagra-dose og libido

Data (Field, 2013, fiktivt datasett)

30 menn ble tilfeldig fordelt til tre grupper: *placebo*, *lav dose* eller *høy dose* Viagra.
Utfallsvariabel: selvrapportert libido.

	Placebo	Lav dose	Høy dose	Totalt
n	9	8	13	30
M	3.22	4.88	4.85	4.37
SD	1.79	1.46	2.12	1.96

Forskningsspørsmål: Påvirker Viagra-dose libidonivå?

$$H_0 : \mu_{\text{placebo}} = \mu_{\text{lav}} = \mu_{\text{høy}}$$

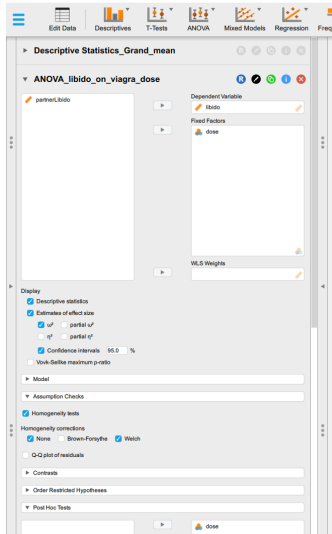
$$H_1 : \text{minst én } \mu \text{ er forskjellig}$$

- Legg den avhengige variabelen (libido) i **Dependent Variable**.
- Legg grupperingsvariabelen (dose) i **Fixed Factors**.
- Kryss av **Homogeneity tests** (Levene's) under *Assumption Checks*.
- Kryss av ω^2 (og evt. η^2) med 95% KI under *Estimates of effect size*.
- Under *Post Hoc Tests*: flytt dose over, velg korreksjon (Bonferroni/Tukey/Scheffe).

Prøv selv

<https://istats.shinyapps.io/ANOVA/>

JASP-meny: ANOVA-panelet



Sjekker forutsetningene for eksempelet

Levene's F	df_1	df_2	p
0.711	2	27	.500

Tolkning

$p = .500 > .05 \rightarrow$ vi forkaster **ikke** nullhypotesen om lik varians. Homogenitetsforutsetningen er tilfredsstilt for dette eksempelet.

ANOVA_libido_on_viagra_dose

ANOVA - libido

Homogeneity Correction	Cases	Sum of Squares	df	Mean Square	F	p	ω^2	95% CI for ω^2	
								Lower	Upper
None	dose	16.844	2.000	8.422	2.416	0.108	0.086	0.000	0.296
	Residuals	94.123	27.000	3.486					
Welch	dose	16.844	2.000	8.422	2.566	0.106	0.086	0.000	0.296
	Residuals	94.123	17.331	5.431					

Note. Type III Sum of Squares

Descriptives

Descriptives - libido

dose	N	Mean	SD	SE	Coefficient of variation
Placebo	9	3.222	1.787	0.596	0.555
Low Dose	8	4.875	1.458	0.515	0.299
High Dose	13	4.846	2.115	0.587	0.436

Assumption Checks

Test for Equality of Variances (Levene's)

F	df1	df2	p
0.711	2.000	27.000	0.500

Sannsynlighetstolkning

Sannsynligheten for å observere $F = 2.57$ eller mer ekstremt, hvis H_0 er sann (alle gruppegjennomsnitt er like), er $p = .106$.

"Det var ingen signifikant effekt av Viagra-dose på libido, $F(2, 17.33) = 2.57$, $p = .106$, $\omega^2 = .09$, 95% KI [.00, .29]."

Konklusjon

$p = .106 > .05 \rightarrow$ vi forkaster ikke H_0 . Vi finner ikke bevis for at Viagra-dose påvirker libido i dette (fiktive) datasettet.

Effektstørrelse: η^2 og ω^2

$$\eta^2 = \frac{SS_M}{SS_T}$$

$$\omega^2 = \frac{SS_M - df_M \cdot MS_R}{SS_T + MS_R}$$

SS_M/SS_T = kvadratsum for hhv. modell og total | df_M = frihetsgrader for modell | MS_R = gjennomsnittlig feilvarians

Med våre data

$$\eta^2 = \frac{16.844}{110.967} = .152$$

$$\omega^2 = \frac{16.844 - 2(3.486)}{110.967 + 3.486} = .086$$

Konvensjon (Cohen, 1988)	.01	.06	.14
Styrke	liten	middels	stor

Hvorfor ω^2 , ikke η^2 ?

- η^2 er **positivt skjev**: overvurderer effekten, spesielt i små utvalg.
- ω^2 er bias-korrigert – anbefalt effektstørrelse i moderne rapportering.
- η^2 rapporteres likevel oftest i praksis: enklere å regne ut, lang tradisjon.
- Her: $\eta^2 = .152$ vs. $\omega^2 = .086$ – ganske ulikt bilde av effekten!

Hvorfor trenger vi multiple comparisons?

- ANOVA er en **omnibus-test**: den forteller bare at *ikke alle* gruppegjennomsnitt er like.
- Den forteller **ikke** *hvilke* gjennomsnitt som skiller seg fra hvilke.
- Løsning: post-hoc-tester *etter* en signifikant ANOVA, eller forhåndsspesifiserte hypoteser for planlagte kontraster.

Planlagte kontraster vs. post hoc

Planlagte kontraster

Planlagt *før* datainnsamling, basert på en spesifikk hypotese (f.eks. fra tidligere forskning).

Post hoc

Ikke planlagt på forhånd – i praksis t-tester mellom alle mulige gruppepar, med korrigert signifikansnivå.

Hvorfor korrigere?

Flere parvise tester → økt sjanse for minst én falsk positiv (jf. Type-I-inflasjon fra forrige forelesning). Post-hoc-tester korrigerer signifikansnivået for å kompensere.

Post-hoc-tester: hvilken skal jeg velge?

Test	Når brukes den
Bonferroni	Enklest og mest konservativ; deler α på antall tester. Vanligst i praksis.
Tukey (HSD)	God balanse mellom styrke og Type-I-kontroll ved lik variansforutsetning.
Games–Howell	Moderne standard når homogenitet av varians <i>ikke</i> holder (analogt til Welch's t-test).
Scheffe	Mer konservativ og fleksibel – tillater også komplekse kontraster, ikke bare parvise.

Post Hoc Tests

Standard (LSD)

Post Hoc Comparisons - dose

		Mean Difference	SE	t	Cohen's d	P _{bonf}
Placebo	Low Dose	-1.653	0.907	-1.822	-0.885	0.239
	High Dose	-1.624	0.810	-2.006	-0.870	0.165
Low Dose	High Dose	0.029	0.839	0.034	0.015	1.000

Note. P-value adjusted for comparing a family of 3

"Post-hoc-analyse med Bonferroni-korreksjon viste at placebogruppen rapporterte lavere libido enn både lav dose, $MD = -1.65$, $p = .239$, $d = 0.89$, 95% $KI [-2.16, 0.39]$, og høy dose, $MD = -1.62$, $p = .165$, $d = 0.87$, 95% $KI [-2.02, 0.28]$, mens lav dose-gruppen rapporterte marginalt høyere libido enn høy dose-gruppen, $MD = 0.03$, $p = 1.000$, $d = 0.02$, 95% $KI [-1.13, 1.16]$. Ingen av forskjellene var statistisk signifikante."

Konklusjon

Konsistent med den ikke-signifikante omnibus- F -en: ingen gruppear skiller seg signifikant fra hverandre når vi korrejerer for multiple sammenligninger.

Diskusjon: Eksamensangst og KI-støtte

En forsker undersøker om ulik grad av KI-støtte før eksamen påvirker eksamensangst:

Gruppe 1 Ingen KI-støtte (kontroll)

Gruppe 2 Enkel FAQ-chatbot

Gruppe 3 Personlig AI-studieveileder (GPT-basert)

90 studenter randomisert likt til de tre gruppene. Utfall: eksamensangst-skår (0–100).

Diskuter

Hvorfor holder det ikke å kjøre tre t-tester her? Hva trenger vi i tillegg til selve F -testen for å si *hvilke* grupper som skiller seg? Hvordan ville du korrigert for multiple sammenligninger?

ANOVA som regresjon

Samme generelle lineære modell – nyere innpakning

Logikken bak dummy-koding for flere grupper

- Med k grupper trenger vi $k - 1$ dummy-variabler – én gruppe er referansekategori.
- Med våre data ($k = 3$): Placebo = referanse ($X_1 = 0, X_2 = 0$).

Gruppe	X_1 (Lav dose)	X_2 (Høy dose)
Placebo (referanse)	0	0
Lav dose	1	0
Høy dose	0	1

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

JASP-output: regresjonskoeffisienter

Coefficients

Model		Unstandardized	Standard Error
H_0	(Intercept)	4.367	0.357
H_1	(Intercept)	3.222	0.622
	dose (Low Dose)	1.653	0.907
	dose (High Dose)	1.624	0.810

^a Standardized coefficients can only be computed for continuous predictors.

H_0 = modell uten prediktorer (kun grand mean); H_1 = modell med de to dummy-variablene

Fra regresjonslikning til ANOVA- F -test

$$\hat{Y} = 3.222 + 1.653X_1 + 1.624X_2$$

Placebo	Lav dose	Høy dose
$\hat{Y} = 3.222$	$\hat{Y} = 3.222 + 1.653 = 4.875$	$\hat{Y} = 3.222 + 1.624 = 4.846$

Nøkkelpoeng

b_0 , b_1 og b_2 gjensker de tre gruppegjennomsnittene nøyaktig. Testen av om X_1 og X_2 *samlet* forbedrer modellen (ΔR^2 -test) gir nøyaktig samme F som ANOVA-tabellen: $F(2, 27) = 2.42$.

Hvorfor dette betyr noe

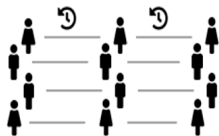
- Regresjonsspråket generaliserer rett til flere prediktorer – dette er ANCOVA og faktoriell ANOVA.
- Legg til en kontinuerlig kovariat (f.eks. alder) → ANCOVA.
- Legg til enda en grupperingsvariabel (f.eks. kjønn) → faktoriell ANOVA, med mulighet for interaksjonseffekter.

Ta med deg videre

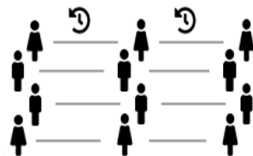
t-test → regresjon med én dummy-prediktor → enveis ANOVA → ANCOVA/faktoriell ANOVA. Samme generelle lineære modell (GLM) hele veien.

- ANOVA sammenligner **tre eller flere** gruppegjennomsnitt ved å partisjonere varians:
 $SS_T = SS_M + SS_R$.
- $F = MS_M / MS_R$ – stor F betyr at systematisk varians dominerer over uforklart varians.
- Sjekk forutsetninger; bruk Welch/Brown–Forsythe som trygg standard.
- ANOVA er en regresjon med $k - 1$ dummy-prediktorer – samme GLM som t-testen.
- Signifikant omnibus- $F \rightarrow$ post-hoc-tester for å lokalisere hvilke grupper som skiller seg.
- Rapporter alltid effektstørrelse (ω^2) i tillegg til p -verdi.

Neste uke: ANOVA 2 og 3



Faktoriell ANOVA & ANCOVA



ANOVA med repeterte målinger

Samme logikk – flere grupperingsvariabler, kovariater og måletidspunkt.

Takk for i dag!