

Kvantitativ metode II: t-tester

PSY2117 / PSYK4317

Frederick Anyan

Førsteamanuensis i kvantitativ forskningsmetode

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

NTNU, Institutt for psykologi

H2026

- Kort historie: hvorfor heter det "Student's t-test"?
- Hensikten med en t-test – hvilket spørsmål svarer den på?
- Tre varianter: ett utvalg, uavhengig, avhengig
- Forutsetninger og Student vs. Welch
- Praktisk eksempel for hver variant (ekte data, JASP)
- Statistisk vs. praktisk signifikans – effektstørrelse (Cohen's d)
- Hvorfor vi **ikke** bare kjører mange t-tester (Type-I inflasjon)

- Guinness drev vitenskapelig ølbrygging og ansatte topp statistikk- og biokjemistudenter fra Oxford og Cambridge.
- En av dem: **William Sealy Gosset**.
- I 1908 introduserte Gosset t-testen for å sammenligne stout-brygg med *små utvalg*.
- Guinness tillot ikke ansatte å publisere under eget navn → Gosset publiserte som "*Student*".
- Derfor: **Student's t-test** – den dag i dag.



William Gosset ("Student")

Én type spørsmål, tre varianter

En t-test tester om en forskjell i **gjennomsnitt** er større enn det vi forventer ut fra tilfeldig variasjon alene.

Vi kan sammenligne:

- 1 gjennomsnittet i én gruppe mot en antatt/kjent verdi
- 2 gjennomsnittet i **to uavhengige grupper**
- 3 gjennomsnittet på **to måletidspunkt** for de samme personene

De tre variantene – oversikt

Test	Spørsmål	Eksempel
Ett utvalg	Skiller gruppesnittet seg fra en kjent/antatt verdi?	Gir 1000 ekstra kalorier/dag 16 pund vektøkning?
Uavhengig	Har to uavhengige grupper ulikt snitt?	Er aggresjonsskåren ulik hos psykologi- vs. matematikkstudenter?
Avhengig (paret)	Endrer samme person seg mellom to tidspunkt?	Endrer uro seg hos demenspasienter fra vår til høst?

Forutsetninger

	Ett utv.	Uavh.	Avh.
Tilfeldig utvalg	✓	✓	✓
Avhengig variabel er kvantitativ	✓	✓	✓
Normalfordelte residualer	✓	✓	✓
Uavhengige feilledd	–	mellom gruppene	mellom <i>par</i>
Homoskedastisitet (lik varians)	–	✓	–

Parametrisk vs. ikke-parametrisk

Parametriske tester (som t-testen) antar en bestemt fordeling og bruker gjennomsnitt/SD. Ikke-parametriske tester gjør færre fordelingsantagelser og bruker ofte *rangordning* i stedet for rådata – mer robust mot skjevhet/uteliggere, men noe lavere styrke når antagelsene faktisk holder.

Hva hvis normalitet brytes?

Med lite utvalg og klart skjeve data: vurder ikke-parametriske alternativer – *Mann-Whitney U* (uavhengig) eller *Wilcoxon signed-rank* (avhengig).

Student vs. Welch's t-test

To strategier:

- 1 Sjekk Levene's test for lik varians.
Signifikant → bruk Welch.
Ikke signifikant → bruk Student.
- 2 Eller: **rapporter alltid Welch** – den korregerer automatisk for ulik varians.

Anbefaling

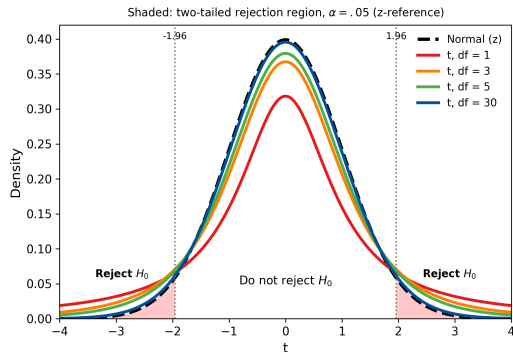
De fleste statistikkfagfolk anbefaler nå **Welch som standard**: den koster nesten ingenting når variansene er like, og redder deg når de ikke er det.

Utvalgsfordelingen under H_0

- Vi bygger en **utvalgsfordeling** for gjennomsnittet, gitt at H_0 er sann.
- Spørsmålet: hvor havner *vårt* gjennomsnitt på denne fordelingen?
 - Godt innenfor det sannsynlige området $\rightarrow$ ikke forkast H_0
 - Langt ute i halen $\rightarrow$ forkast H_0
- Formen avhenger av frihetsgrader (df) – færre observasjoner gir tyngre haler.

Prøv selv

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





Ett utvalg t-test

Skiller gruppesnittet seg fra en kjent eller antatt verdi?

Eksempel: Vektøkning

Data (Moore, McCabe & Craig, 2012)

16 deltakere fikk et overskudd på 1000 kalorier/dag i 8 uker. Utfallsvariabel: vektøkning i pund (lb).

Statistikk	Verdi
n	16
M	10.41
SD	3.84
SE_M	0.96

Forskningsspørsmål: Gir 1000 ekstra kalorier/dag **16 pund** vektøkning, som en enkel kalorimodell forutsier?

$$H_0 : \mu = 16 \text{ lb}$$

$$H_1 : \mu < 16 \text{ lb}$$

Ett utvalg t-test: formel

$$t = \frac{M - \mu_0}{SE_M}, \quad SE_M = \frac{SD}{\sqrt{n}}$$

- M = utvalgsgjennomsnitt μ_0 = antatt/hypotetisk verdi
- SD = utvalgets standardavvik n = utvalgsstørrelse

Innsatt

$$t = \frac{10.41 - 16}{0.96} = -5.82, \quad df = 15$$

JASP: kjøre testen

- Legg utfallsvariabelen i **Test variable(s)**.
- Velg **Student** for klassisk t-test.
- Endre **Test value** til den hypotesen du tester (her: 16).

Prøv selv

https://istats.shinyapps.io/Inference_mean/

JASP-output: Ett utvalg t-test

Test variable

Select 'Student' for t-test

Change to test value

$H_a: \mu < 16$

Computes the difference between $(\bar{x} - \mu_0)$

One Sample T-Test

Variables: weight

Tests: ☒ Student, ☐ Wilcoxon signed-rank, ☐ Z Test

Test value: 16

Std. deviation: 1

Alternative Hypothesis: ☐ ≠ Test value, ☐ > Test value, ☒ < Test value

Assumption checks: ☒ Normality, ☒ Q-Q plot residuals

Missing Values: ☒ Exclude cases per variable, ☐ Exclude cases listwise

Additional Statistics: ☒ Location estimate, ☐ Confidence interval 95 %, ☒ Effect size, ☐ Confidence interval 95 %, ☒ Descriptives, ☐ Vovk-Sellke maximum p-ratio

Plots: ☐ Descriptives plots, ☐ Raincloud plots, ☐ Bar plots

Results

One Sample T-Test

	t	df	p	Mean Difference	Cohen's d	SE Cohen's d
weight	-5.823	15	< .001	-5.591	-1.456	0.359

Note: For the Student t-test, the alternative hypothesis specifies that the mean is less than 16.
Note: Student's t-test.

Assumption Checks

Test of Normality (Shapiro-Wilk)

	W	p
weight	0.938	.325

Note: Significant results suggest a deviation from normality.

Q-Q Plots

weight

Descriptives

	N	Mean	SD	SE	Coefficient of variation
weight	16	10.41	3.841	0.960	0.369

Sannsynlighetstolkning

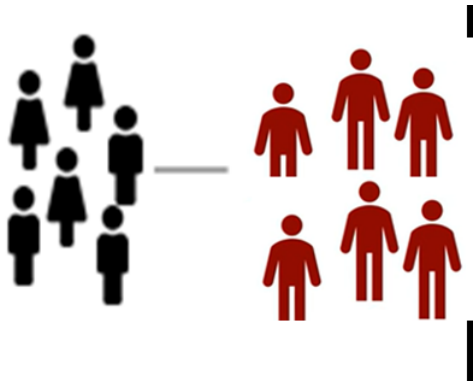
Sannsynligheten for å observere $t = -5.82$ eller mer ekstremt, **hvis H_0 er sann**, er $p < .001$.

"Deltakernes vektøkning ($M = 10.41$, $SD = 3.84$) var signifikant lavere enn den forventede verdien på 16 pund, $t(15) = -5.82$, $p < .001$, og representerte en stor effekt, $d = 1.46$."

Konklusjon

Den svært lille p-verdien gir bevis mot $H_0 \rightarrow$ forkast H_0 .

Uavhengig t-test



Uavhengig t-test

To uavhengige grupper – har de samme gjennomsnitt?

Eksempel: Usynlighetskappe og rampestreker

Studie (Andy Field)

24 deltakere. 12 fikk *ikke* usynlighetskappe, 12 fikk kappe. Utfall: antall rampete handlinger på en uke.

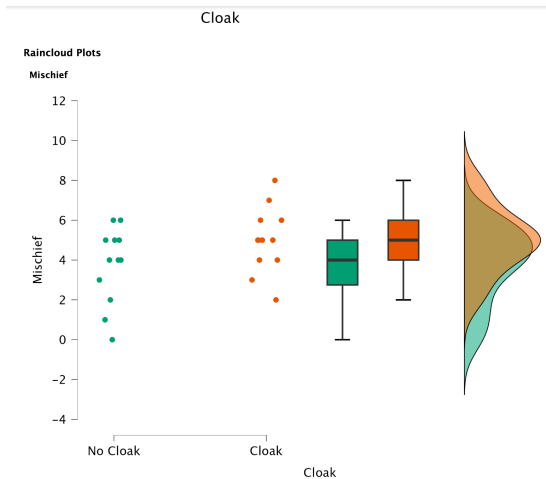
	Uten kappe (0)	Med kappe (1)
n	12	12
M	3.75	5.00
SD	1.91	1.65
SE_M	0.55	0.48

Forskningsspørsmål: Fører det å ha på seg en usynlighetskappe til flere rampestreker enn å ikke ha kappe?

$$H_0 : \mu_{\text{uten}} = \mu_{\text{med}}$$

$$H_1 : \mu_{\text{uten}} \neq \mu_{\text{med}}$$

Fordelingene visualisert



Uavhengig t-test: formel

$$t = \frac{M_1 - M_2}{SE_{\text{pooled}}}$$

Sammenslått (pooled) standardavvik

I motsetning til ett-utvalgstesten må vi her kombinere variansen fra **begge** gruppene til ett samlet mål på spredning.

Resultat (Welch-korrigert)

$$t(21.54) = -1.71, \quad p = .101$$

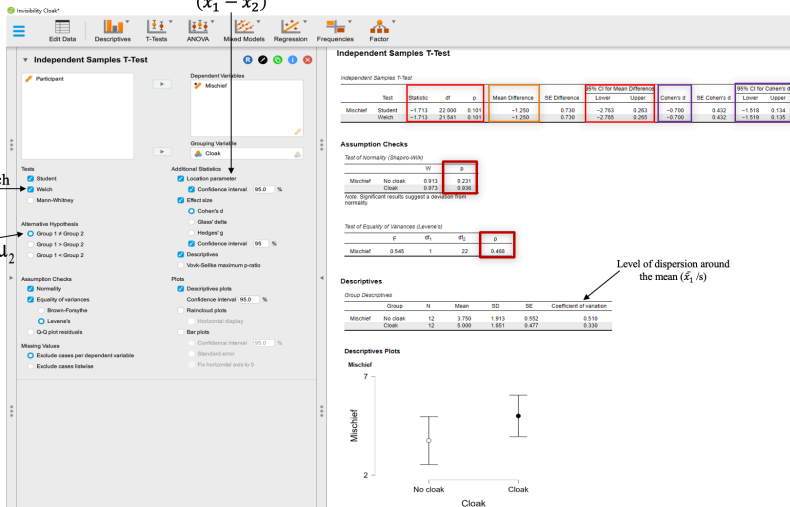
JASP-output: Uavhengig t-test

Computes the difference between

$$(\bar{x}_1 - \bar{x}_2)$$

Report Welch
test

$$H_a: \mu_1 \neq \mu_2$$



Sannsynlighetstolkning

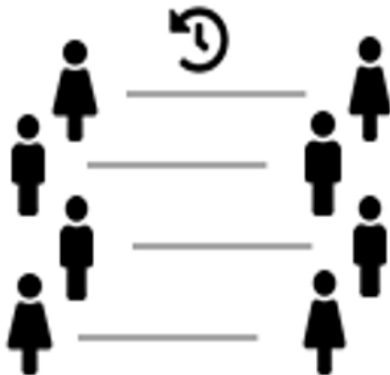
Sannsynligheten for å observere $t = -1.71$ eller mer ekstremt, **hvis H_0 er sann**, er $p = .101$.

"Deltakere med usynlighetskappe utførte i gjennomsnitt flere rampestreker ($M = 5.00$, $SE = 0.48$) enn de uten kappe ($M = 3.75$, $SE = 0.55$). Levene's test var ikke signifikant, $F(1, 22) = 0.54$, $p = .468$.

Gjennomsnittsforskjellen, -1.25 , 95% KI $[-2.76, 0.26]$, var ikke signifikant, $t(21.54) = -1.71$, $p = .101$, men representerte en middels effekt, $d = 0.70$, 95% KI $[-1.52, 0.13]$."

Konklusjon

$p = .101 > .05 \rightarrow$ vi forkaster **ikke** H_0 .



Avhengig t-test

Samme personer, to måletidspunkt – endrer de seg?

Eksempel: Demens og forstyrrende atferd

Data (Moore, McCabe & Craig, 2012)

15 pasienter med demens. Antall forstyrrende atferder registrert om våren og om høsten for de samme pasientene.

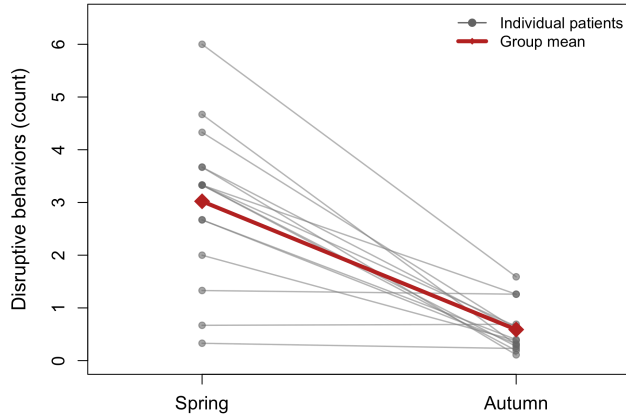
Statistikk	Gjennomsnittlig differanse
n	15
M_d	2.43
SE_d	0.38

Forskningsspørsmål: Endrer antall forstyrrende atferder seg hos demenspasienter fra vår til høst?

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d \neq 0$$

Endringen visualisert



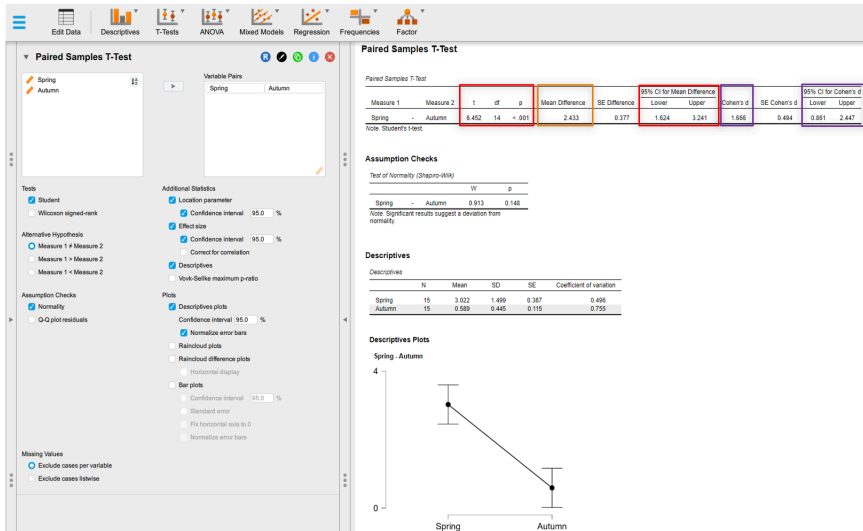
Avhengig t-test: formel og resultat

$$t = \frac{M_d - 0}{SE_d}, \quad M_d = \text{gjennomsnittlig differanse}$$

Innsatt

$$t(14) = \frac{2.43}{0.38} = 6.45, \quad p < .001$$

JASP-output: Avhengig t-test



Sannsynlighetstolkning

Sannsynligheten for å observere $t = 6.45$ eller mer ekstremt, hvis H_0 er sann, er $p < .001$.

"Antall forstyrrende atferder blant demenspasienter sank fra ($M = 3.02$, $SE = 0.39$) om våren til ($M = 0.59$, $SE = 0.12$) om høsten. Differansen på 2.43, 95% KI [1.63, 3.24], var signifikant, $t(14) = 6.45$, $p < .001$, og representerte en stor effekt, $d = 1.67$, 95% KI [0.87, 2.45]."

Konklusjon

Den lille p-verdien gir sterke bevis mot $H_0 \rightarrow$ forkast H_0 .

Diskusjon: Kaffe og konsentrasjon

En klasse designer tre mini-studier om kaffe og konsentrasjon:

Studie A Samme studenter testes med og uten kaffe.

Studie B Populasjonssnittet (uten kaffe) er kjent $= 50$. Testes kaffedrikkende studenter mot denne verdien.

Studie C Studenter randomisert til kaffe- vs. ikke-kaffegruppe.

Diskuter

Hvilken t-test passer hvert design? Hvilket design gir sterkest bevis, og hvorfor? Forventer du like resultater fra alle tre?

Statistisk $\neq$ praktisk signifikans

- En signifikanstest spør: skiller den sanne parameteren seg fra H_0 ?
- P-verdien oppsummerer **bevis mot H_0** – den sier ingenting om praktisk relevans.
- En liten p-verdi ($p = .001$) er statistisk signifikant, men innebærer ikke automatisk et *viktig* funn.

Moderne påminnelse

Med store nok utvalg blir *alt* statistisk signifikant. Effektstørrelse er derfor ikke valgfritt – det er nødvendig.

Effektstørrelse: hvorfor ikke bare gjennomsnittsdifferansen?

- Enkleste mål: rå gjennomsnittsdifferanse (f.eks. 1.25 flere rampestreker).
- Problem 1: avhenger av variabelens skala (cm vs. km gir ulik "størrelse").
- Problem 2: tar ikke hensyn til spredningen i dataene.

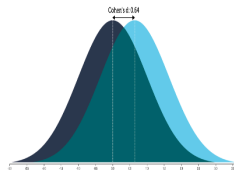
Løsning: standardisert effektstørrelse

Cohen's d uttrykker forskjellen i standardavviksenheter – sammenlignbar på tvers av studier og variabler.

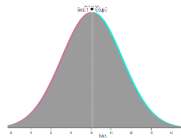


Jacob Cohen (1923–1998)

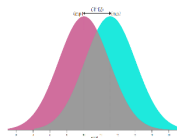
Cohen's d – visualisert



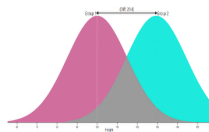
liten effekt



mellomstor effekt



stor effekt



Cohen's d for de tre testene

$$d_{\text{ett utvalg} / \text{avhengig}} = \frac{M - \mu_0}{SD}$$

$$d_{\text{uavhengig}} = \frac{M_1 - M_2}{SD_{\text{pooled}}}$$

- Formelen for ett-utvalgs- og avhengig t-test er **identisk** – logikken i hypotesetestene er den samme.
- Husk: en avhengig t-test er egentlig en ett-utvalgs t-test på differanseskårene, testet mot $\mu_d = 0$.

Utforsk interaktivt

<https://rpsychologist.com/cohend/> – se hvordan d endres med gjennomsnittsforskjell og spredning.

t-testen som regresjon

Samme test, ny innpakning – broen til ANOVA og regresjon

Logikken bak dummy-regresjon

- Kod gruppetilhørighet som en **dummy-variabel**: $X = 0$ eller $X = 1$.
- Kjør enkel lineær regresjon: $\hat{Y} = b_0 + b_1X$.
- b_0 = gjennomsnittet i referansegruppen ($X = 0$).
- b_1 = **differansen** mellom gruppene – akkurat det t-testen tester!

Med usynlighetskappe-dataene

X : 0 = ingen kappe, 1 = kappe

$$\hat{Y} = 3.75 + 1.25X$$

Fra regresjonslikning til alle tre t-tester

$$\hat{Y} = b_0 + b_1X = 3.75 + 1.25X$$

$X = 0$ (ingen kappe)

$$\hat{Y} = 3.75 + 1.25(0) = 3.75$$

$X = 1$ (kappe)

$$\hat{Y} = 3.75 + 1.25(1) = 5.00$$

Test	Regresjonsmodell	Hypotese
Ett utvalg	$\hat{Y} = b_0$	$H_0 : b_0 = \mu_0$
Uavhengig	$\hat{Y} = b_0 + b_1X$ ($X = \text{gruppe}$)	$H_0 : b_1 = 0$
Avhengig (paret)	$\hat{D} = b_0$ (på differanseskårer D)	$H_0 : b_0 = 0$

Hvorfor dette betyr noe

- t -verdien for b_1 i regresjonen er **identisk** med t -verdien fra den uavhengige t -testen.
- Samme p -verdi, samme konklusjon – to språk for samme modell.
- Fordelen med regresjonsspråket: det generaliserer rett til flere grupper og flere prediktorer – dette **er** ANOVA og ANCOVA.

Hvor mange dummy-prediktorer trenger vi?

- t -test: X har **2 nivåer** (f.eks. kappe/ingen kappe) $\rightarrow$ én dummy-variabel er nok ($k - 1 = 1$ for $k = 2$ grupper).
- ANOVA med k grupper: gruppetilhørighet kodes med $k - 1$ dummy-variabler (én gruppe er referansekategori, $X = 0$ på alle dummyene).
- Eksempel: 4 betingelser $\rightarrow$ 3 dummy-variabler i én samlet regresjonsmodell, i stedet for 6 separate t -tester.

Ta med deg videre

t -test $\rightarrow$ regresjon med én dummy-prediktor $\rightarrow$ ANOVA $\rightarrow$ ANCOVA. Samme generelle lineære modell hele veien.

Begrensningen ved t-tester

- t-testen sammenligner **kun to grupper** om gangen.
- Mange studier har tre, fire eller flere grupper eller betingelser.
- Fristende løsning: bare kjør flere t-tester, én per par?

Faren

Med $\alpha = .05$ per test har hver enkelt test 5% sjanse for en falsk positiv (Type-I feil) – selv om H_0 er sann.

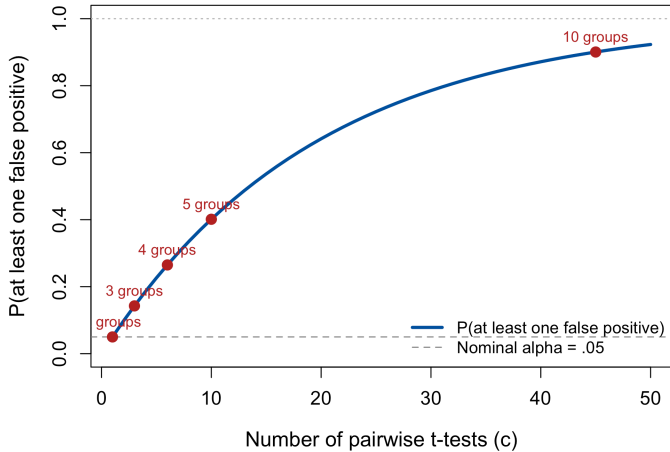
Inflasjon av Type-I feilrate

Sannsynlighet for **minst én** falsk positiv ved c parvise t-tester ($\alpha = .05$):

$$P(\text{minst én falsk positiv}) = 1 - (1 - \alpha)^c$$

Antall grupper	Antall t-tester	Nominelt α	Familywise α
2 grupper	1 t-test	.05	.05
3 grupper	3 t-tester	.05	.14
4 grupper	6 t-tester	.05	.26
5 grupper	10 t-tester	.05	.40
...			
10 grupper	45 t-tester	.05	.90

Type-I-inflasjon – kurven



Family-wise error rate og korreksjon

- Denne inflasjonen kalles **family-wise error rate (FWER)**.
- Vanligste løsning: **Bonferroni-korreksjon** – del α på antall tester (α/k).
- Kostnad: redusert statistisk styrke – vanskeligere å finne en reell effekt.

Alternativ i praksis

Når man har mer enn to grupper, er løsningen sjelden "mange korrigerede t-tester" – det er **én ANOVA**, som håndterer flere grupper i samme rammeverk.

Eksempel



Er det signifikante forskjeller i hyppighet av **depresjon** mellom **menn og kvinner** bosatt i **forskjellige fylker**?

uavhengig variabler: fylke, kjønn

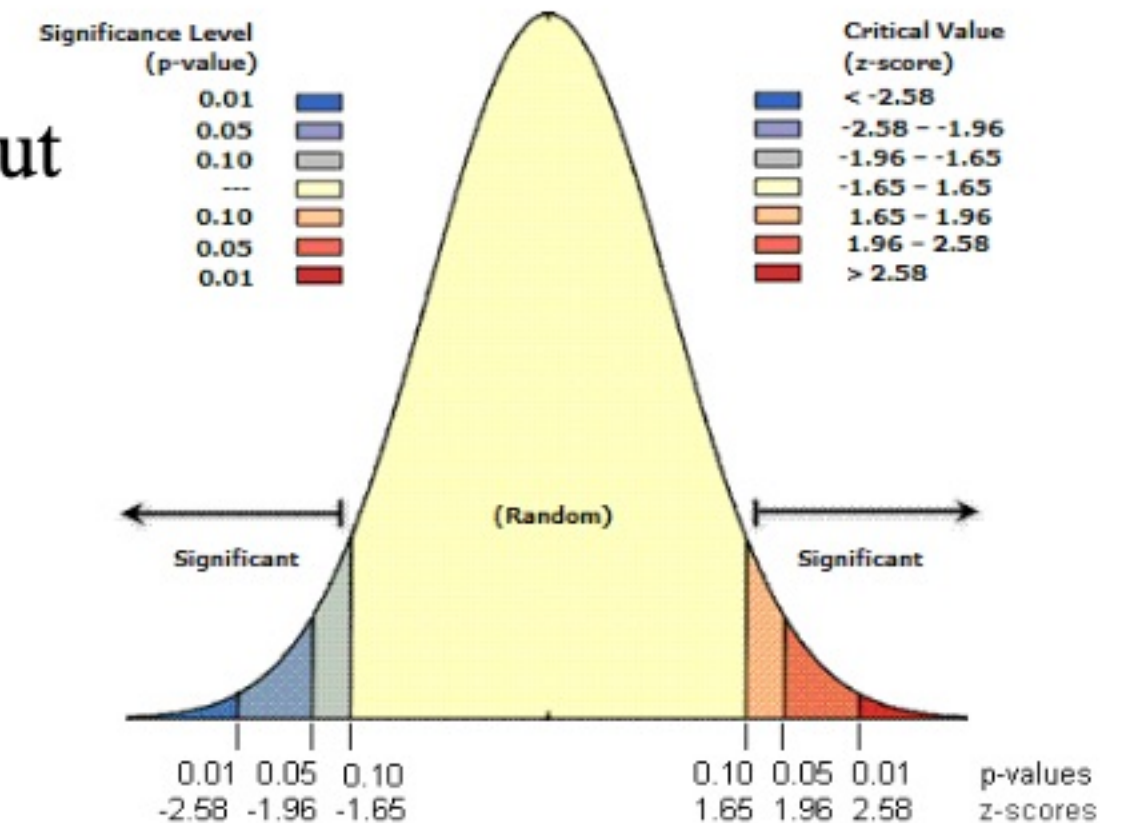
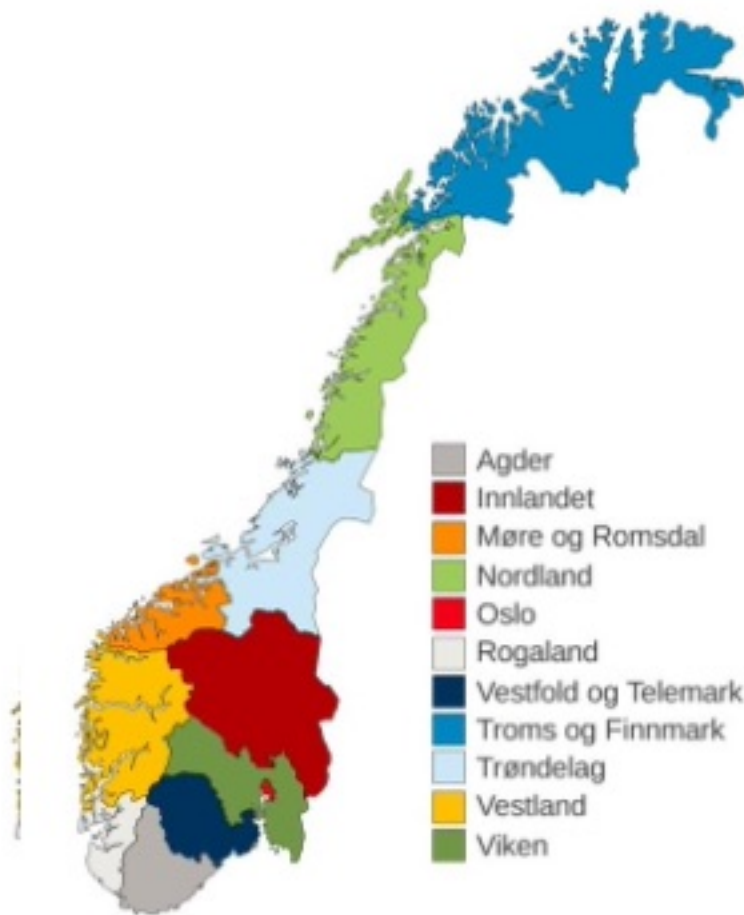
avhengig variabel: depresjonsskåre

Hvor mange t-tester trenger vi?

- 11 fylker, 2 kjønn, trenger 231 tester for å sammenlikne alle
- Bonferroni: $\alpha = \frac{0.05}{231} = 0.0002$

ANOVA gjør at vi får $\alpha = 0.05$ og vi kan få ut

- effekt av fylke
- effekt av kjønn
- effekt av interaksjonen mellom de to!



- Tre varianter av t-test: ett utvalg, uavhengig, avhengig – samme logikk, ulikt spørsmål.
- Sjekk forutsetninger; bruk Welch som trygg standard ved uavhengig t-test.
- Rapporter alltid effektstørrelse (d) og konfidensintervall – ikke bare p .
- Flere grupper $\rightarrow$ ikke flere t-tester. Det er derfor vi trenger ANOVA.

For å unngå Type-I-inflasjon ved flere grupper bruker vi **Analysis of Variance (ANOVA)**.

Takk for i dag!