Gyldigheten av analysen – Hvor gode er konklusjonene vi kan trekke?

Silje Synnøve Lyder Hermansen

s.s.l.hermansen@stv.uio.no

Gyldigheten av analysen

Jacobsen (2005), kap. 15

Fire elementer for en god undersøkelse

Jacobsen (2005: 345) understreker fire elementer som bør være til stede når vi vurderer kvaliteten på en analyse.

- **Begrepsmessig gyldighet:** Operasjonaliseringen er god, slik at vi vet vi måler det vi sier vi måler.
- **Intern gyldighet:** Vi har påvist en samvariasjon (korrelasjon), og vi kan sannsynliggjøre at det finnes et kausalt forhold mellom variablene.
- **Ekstern gyldighet:** Resultatene kan generaliseres/overføres til andre områder.
- **Reliabilitet:** Dataene og analysen er pålitelige.

Begrepsmessig gyldighet
Hvor god er operasjonaliseringen?

Jacobsen (2005), kap. 15

Begremsmessig gyldighet

Jacobsen (2005, 348-352) foreslår flere strategier for å sikre den begrepsmessige gyldigheten:

- **Gyldighet ved første øyekast**
- **Flere indikatorer**
- **Kriterievaliditet**
- **Annen forskning/kumulativ forskning**

Begremsmessig gyldighet – gyldighet ved første øyekast

Gyldighet ved første øyekast ("face value") vil bli bedre om man sjekker operasjonaliseringer med forskerkolleger og med forskningsobjektene. Dette er en kontroll via *intersubjektivitet*: Snakker vi om det samme?

- *Eksempel:* Hva er en prestisjetung komité? Man diskuterer med kolleger, og sjekker hva stortingsrepresentantene selv mener gir status.

Operasjonaliseringer er også et punkt hvor Mertons norm om organisert skepsis er svært tilstede: Det er lett å felle et forskningsprosjekt på grunnlag av operasjonaliseringen.

Begrepsmessig gyldighet – flere indikatorer

Flere indikatorer: Komplekse begreper krever at forarbeid for å gjøres målbare: Man må operasjonalisere. Ofte ender man opp med flere indikatorer som alle måler hver sin del av fenomenet (men ikke nødvendigvis hele).

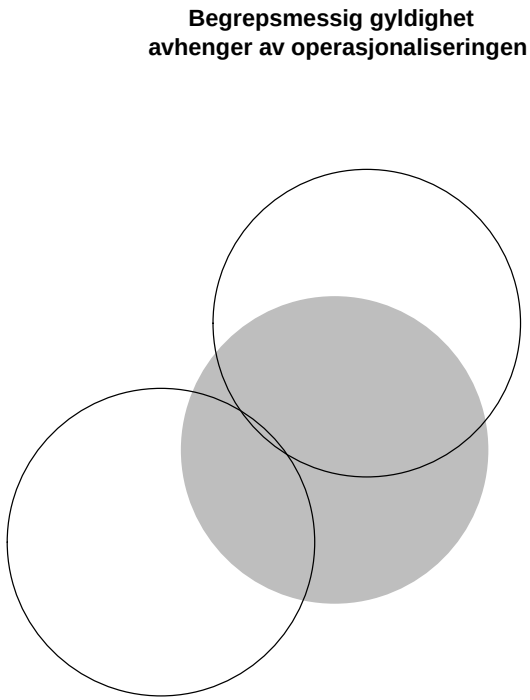


Figure : To eller flere indikatorer (variabler) kan måle deler av et underliggende begrep, og noen operasjonaliseringer er bedre enn andre.

Begrepsmessig gyldighet – flere indikatorer

Jo mer komplekst begrepet er, jo bedre er det å måle det på flere ulike måter. I noen tilfeller kan man teste om disse indikatorene måler det samme ved å se om de korrelerer. Men det kan også være slik at at et abstrakt begrep kan bestå av flere elementer som ikke alltid opptrer sammen. Da vil man ikke finne korrelasjon.

- *Eksempel:* Man kan lage en korrelasjonsmatrise hvor man regner ut Pearsons R mellom alle indikatorene:

	Perioder	Saksordfører	Regjering.omk
Perioder	1.00	0.70	0.21
Saksordfører	0.70	1.00	-0.05
Regjering.omk	0.21	-0.05	1.00

Table : Korrelasjonsmatrise mellom tre operasjonaliseringer av prestisje i Stortinget.

Begrepsmessig gyldighet – kriterie-validitet

Kriterie-validitet: Særlig når man ikke forventer at indikatorene korrelerer, vil det være aktuelt å sjekke kriterie-validiteten til en operasjonalisering.

- Når teorien sier at to fenomener henger sammen, kan man sjekke korrelasjonen mellom disse.
- Dette forutsetter at de to fenomenene er separate, ikke to aspekter/indikatorer for samme underliggende begrep (da ville vi falt tilbake på testen av flere indikatorer).

Eksempel: Sosiolog Harriet Holter (1976; "Om kvinneundertrykkelse, mannsundertrykkelse og herskerteknikker") hadde en tese om krympende institusjoner: Kvinner får innpass i institusjoner med synkende makt. Da forventer vi at alle operasjonaliseringer av prestisjefylte komiteer vil resultere i en negativ korrelasjon med andel kvinner i disse komiteene.

Begrepsmessig gyldighet – kumulativ forskning

Kommunisme/kumulativ forskning: Den begrepsmessige gyldigheten styrkes av tidligere forskning; enten dette er teori eller tidligere empiriske undersøkelser.

- **Teori:** Særlig i kvantitativ forskning hvor vi ikke observerer årsaksmekanismene, men resultatene, vil man lene seg på teorier om hvordan ting henger sammen/årsaksforhold. Men operasjonaliseringen blir dermed ikke bedre enn teorien den lener seg på. Teorien eksisterer før vi foretar undersøkelsen, og er ofte ikke utarbeidet av oss selv: vi står på kjempers skuldre (ref. Mertons norm om kommunisme)

Begrepsmessig gyldighet – kumulativ forskning (forts.)

- **Empirisk forskning:** Når man måler det samme fenomenet med andre metoder, og får det samme resultatene, vil konklusjonene våre komme styrket ut. Dermed er *metodetriangulering* svært effektiv for å styrke den begrepsmessige gyldigheten. Ofte er forskere gode på en eller to metoder, slik at metodetrianguleringen i praksis gjøres av ulike forskere (ref. Mertons norm om kommunisme).

Eksempel: Hege Skjeie (1991, "The Rhetoric of Difference: On Women's Inclusion into Political Elites") har tidligere diskutert og operasjonalisert "prestisje" (og referer i sin tur til andre forskere).

Intern gyldighet
Hva er årsakssammenhengen?

Jacobsen (2005), kap. 15

Intern gyldighet – tre kriterier for å finne en årsakssammenheng

Hvordan kan man forklare et fenomen? Samvariasjon (korrelasjon) er ikke det samme som årsakssammenheng. Hvordan kan man sannsynliggjøre at man har funnet en konsekvens og en (eller flere) årsaker? Man har tre kriterier som bør tilfredsstilles:

- **Samvariasjon (korrelasjon):** Alle årsakssammenhenger krever samvariasjon, men all samvariasjon er ikke en årsakssammenheng.
- **Tidsrekkefølge:** En årsak vil alltid komme før konsekvensen.
- **Kontrollvariabler:** Man må kontrollere for andre mulige årsaker til fenomenet man måler. Dette kommer fra idealet om et eksperiment hvor man har en kontrollgruppe. Det vil si at man gjør et *naturlig kvasieksperiment*. Man forsøker å finne enheter som er like på alle punkter minus ett; årsaksvariabelen.

Intern gyldighet – tre kriterier for å finne en årsakssammenheng (forts.)

Eksempel: Hva forklarer forventning om lønn?

- **Samvariasjon (korrelasjon):** *Eksempel:* Finner man en korrelasjon mellom forventet lønn og kjønn? Karakterer? Foreldres utdanning?
- **Tidsrekkefølge:** Man kan enten gjøre en tidsserieanalyse hvor man observerer enheter over tid, eller man kan observere enheter på ett tidspunkt, men dedusere hvilke fenomener som opptrer først. *Eksempel:* Hva kommer først av forventet lønn og foreldres utdanning? Karakterer? Studieretning? Kjønn? Ofte har mann grensetilfeller: Valgte man et studium fordi man ønsker høy lønn, eller forventer man et lønnsnivå på grunn av studieretningen?

Intern gyldighet – tre kriterier for å finne en årsakssammenheng (forts.)

Eksempel: Hva forklarer forventning om lønn? (forts.)

- **Kontrollvariabler:** Når enheter er like på alle punkter/har samme variabelverdier på to forklaringsvariabler: Finner vi fortsatt en korrelasjon mellom uavhengige variabelen (årsaksvariabelen) og den avhengige variabelen? *Eksempel:* Ved likt karakternivå: Har fortsatt kvinner og menn ulike forventninger til framtidig lønn? Ved likt nivå på foreldres utdanning: Er det fortsatt kjønnsforskjeller?

Intern gyldighet – tre kriterier for å finne en årsakssammenheng (forts.)

Kausal feilslutning: Man risikerer å konkludere at samvariasjon er sammenheng. Det er undersøkelsesdesignet som gjør det mulig å konkludere rundt årsakssamenhengen. Undersøkelsesdesignet gjør at vi kan sammenlikne enheter som er like på alle andre punkter enn ett. I vitenskap konkluderer man med å sammenlikne enheter, ellers risikerer man å bli *normativ*.

- **Man trenger respondenter med ulike verdier:** lønnsforventninger, av ulikt kjønn, med ulike karakterer, foreldre med utdannelse osv. Da kan vi si at en lønnsforventning er "høyere enn".. eller "lavere enn"... (ikke "høyt" og "lavt"; det ville vært normativt.)
- **Man trenger i tillegg nok verdikombinasjoner:** Man trenger kvinner med ulike lønnsforventninger. Man trenger studieretninger med både kvinner og menn osv.

Variasjon er informasjon!

Reliabilitet
Hvor pålitelige er resulatatene?

Jacobsen (2005), kap. 15

Reliabilitet

For å sikre kvaliteten på konklusjonen, må vi sjekke om den er resultatet av undersøkelsesmetoden, heller enn et reelt fenomen. Noen elementer er enkle å sikre seg mot:

- **Regnefeil:** Trekk ved selve analysen kan påvirke svaret. Man gjør regneoperasjoner på variabler med feil målenivå (eks. tar gjennomsnittet av kategoriske variabler). Man feiltolker de statistiske resultatene. Man benytter ikke informasjonen som ligger i dataene (eks. man glemmer å kontrollere for andre årsaker).
- **Kodefeil:** Man har lagt dataene feil inn i datasettet (eks. har lagt inn en ekstra null på lønnsforventning for noen respondenter).

Reliabilitet (forts.)

Andre elementer er vanskeligere å sikre seg mot:

- **Intervjuereffekt:** Svarer respondenten anderledes pga. hvem som spør? Eks: Oppgir mannlige respondenter høyere lønn for å imponere en kvinnelig intervjuer? Man kan standardisere samhandlingen (gi opplæring), og kontrollere ved å bruke ulike intervjuere.
- **Intervjuskjema:** Påvirker spørsmålsstillingen eller -rekkefølgen svaret man får? Man teste ut skjemaet i forkant, eller designe to skjemaer slik at man i praksis får en "kontrollgruppe".
- **Trekk ved respondenten:** Respondenten kan svare strategisk, man har egentlig ingen mening/vet ikke, eller svarer i "hytt og vær". Den beste måten å demme opp for dette på, er å stille konkrete spørsmål. Dette er ikke alltid mulig.

Ekstern gyldighet
Hvor sant er dette utenfor utvalget vårt?

Jacobsen (2005), kap. 15

Ekstern gyldighet

Representativiteten til utvalget i forhold til populasjonen vil avgjøre muligheten vi har for å generalisere:

- **Frafall:** Når noen typer enheter faller fra utvalget, mens andre blir igjen, får vi ikke lenger et utvalg som reflekterer populasjonen. Det kan være vanskelig å rette på.
- **Tilfeldige uoverstemmelser:** Alt er mulig i naturen, men bare noen ting er sannsynlig. Dette kan vi ta hensyn til for å regne ut sikkerhetsmarginer. Det er svært praktisk.

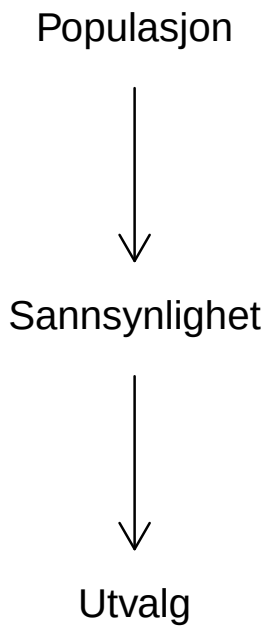
Ekstern gyldighet – hvor tilfeldig er resultatet?

Utvalget vårt avhenger av sannsynligheten for å trekke en enhet fra populasjonen. Når vi trekker flere enheter fra populasjonen, vil typen enheter vi trekker avhenge av sannsynligheten for å trekke dem.

- Man regner ut sannsynligheten for alle mulige variabelverdier.
- Den mest sannsynlige *variabelverdien* er beregnet ved å regne gjennomsnittet av enhetene i populasjonen.
- Sannsynligheter går alltid fra 0 (ikke sannsynlig) til 1 (garantert).
- Summen av alle sannsynligheter er alltid 1.

Ekstern gyldighet – hvor tilfeldig er resultatet?

Hvilke enheter som befinner seg i utvalget avhenger av sannsynligheten for å trekke dem fra populasjonen.

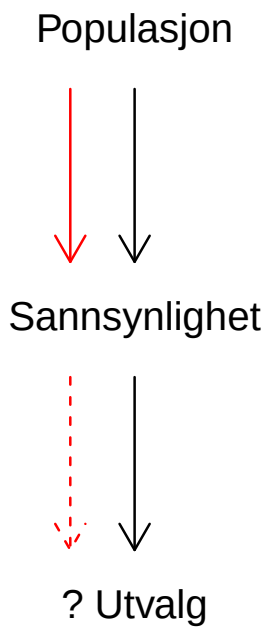


Ekstern gyldighet – hvor tilfeldig er resultatet?

Eksempel 1: Når jeg kjenner kjønnsfordelingen på Stortinget, vet jeg også hva sannsynligheten er å trekke en kvinne i komiteene. Jeg vet det er 67 kvinnelige representanter av i alt 169. Det betyr at jeg har en sannsynlighet på $\frac{67}{169} = 0.4$ for å trekke en kvinne til utvalget mitt. Av og til vil utvalget mitt inneholde flere kvinner, av og til mindre, men disse variasjonene er tilfeldige.

Ekstern gyldighet – hvor tilfeldig er resultatet?

Når jeg kjenner populasjonen, kan jeg regne ut sannsynligheten for å finne liknende enheter i utvalget (fourtsatt at utvalget er tilfeldig).



Ekstern gyldighet – hvor tilfeldig er resultatet?

Eksempel 2: I spørreundersøkelsen vet jeg hvor mange som har fått karakteren 1 på særemne. Jeg vet også at det totale antallet enheter i datasettet er 115. Det vil si at jeg har $\frac{1}{115} = 0.01$ sjanse for å trekke vedkommende.

	Karakter	Respondenter
1	1	1
2	2	2
3	3	16
4	4	37
5	5	45
6	6	14

Table : Frekvenstabell: Karakterfordeling blant respondentene.

Ekstern gyldighet – hvor tilfeldig er resultatet?

Tilsvarende vil sjansen for å trekke en respondent med karakteren 3 være $\frac{16}{115} = 0.14$.

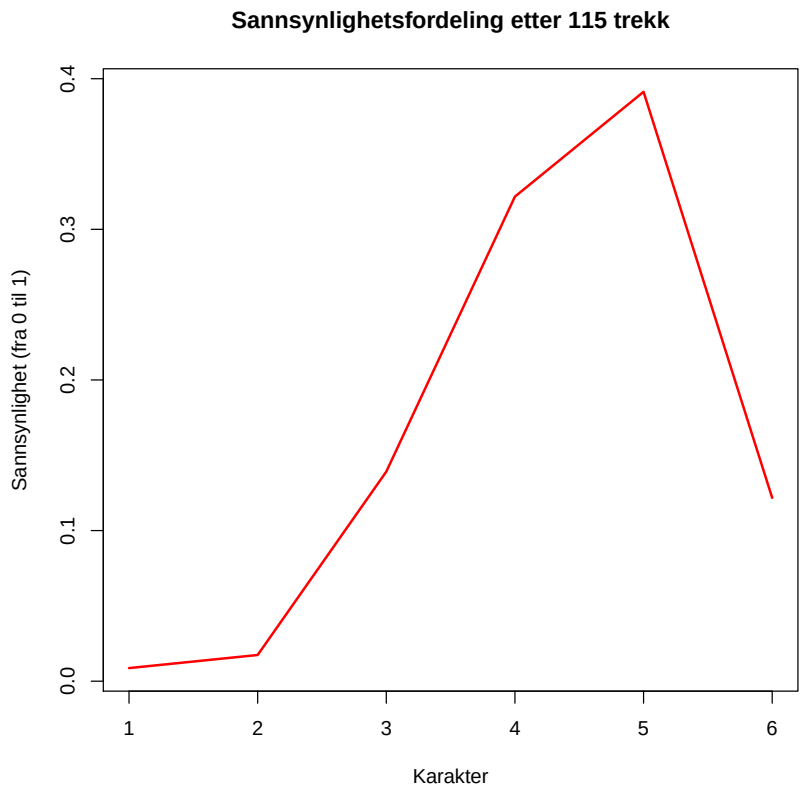
	Karakter	Respondenter	Sannsynlighet
1	1	1	0.01
2	2	2	0.02
3	3	16	0.14
4	4	37	0.32
5	5	45	0.39
6	6	14	0.12

Table : Frekvenstabell: Karakterfordeling blant respondentene.

Summen av den ytterste kolonnen vil alltid bli 1.

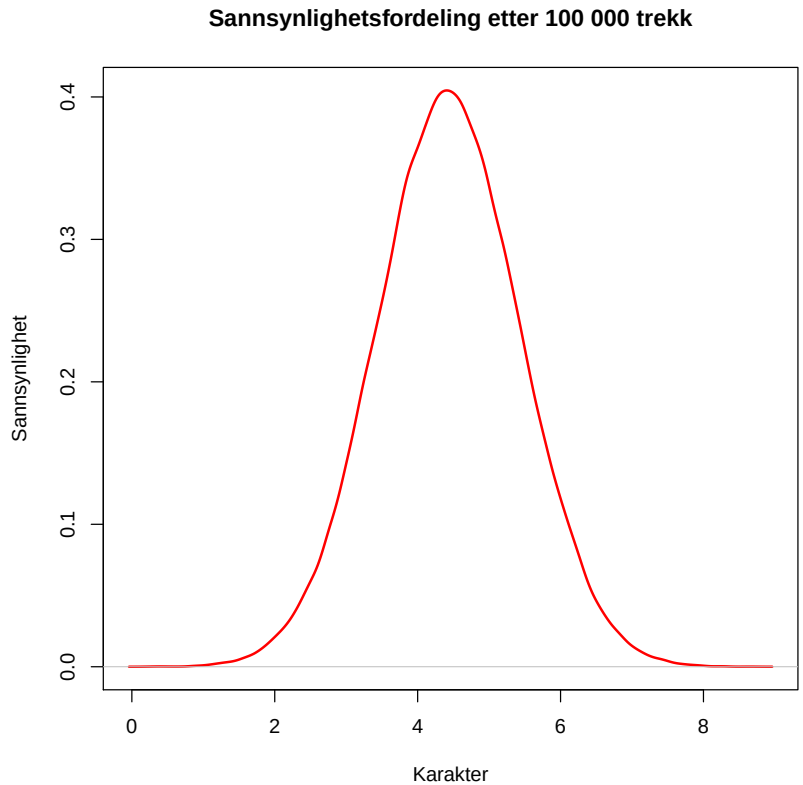
Ekstern gyldighet – frekvensfordeling/sannsynlighetsfordeling

Når vi regner ut sannsynligheten for alle utfall/alle verdier av karakter-variabelen vår, får vi en sannsynlighetsfordeling.



Ekstern gyldighet – sannsynlighetsfordeling/normalfordeling

Om vi skulle trukket et større utvalg et 115 enheter, så ville fordelingen til sist liknet på dette:



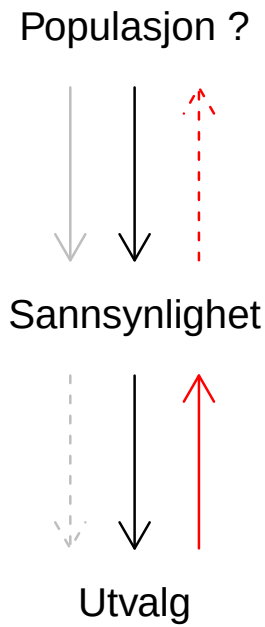
Ekstern gyldighet – sentralgrenseteoremet

Det er et produkt av en matematisk lov som vi kaller *sentralgrenseteoremet*. Vi kjenner alle til den fra før av: Når noe skjer igjen og igjen, vil vi til sist oppdage en "spennvidde" i sannsynlige utfall. Det mest sannsynlige utfallet er gjennomsnittet i populasjonen (karakteren 4, eller menn på Stortinget), men andre – og mer ekstreme – utfall vil også forekomme; de er bare ikke veldig sannsynlige.

Ekstern gyldighet – hvor tilfeldig er resultatet?

Vi bruker denne kunnskapen til å snu logikken på hodet: Vi vet at dataene våre er et utvalg, men vi kjenner ikke populasjonen. Forutsatt at dataene våre er et tilfeldig utvalg, beregner vi sannsynligheten for å få nettopp dette utvalget på grunnlag av frekvensfordelingen i dataene.

Ekstern gyldighet – hvor tilfeldig er resultatet?



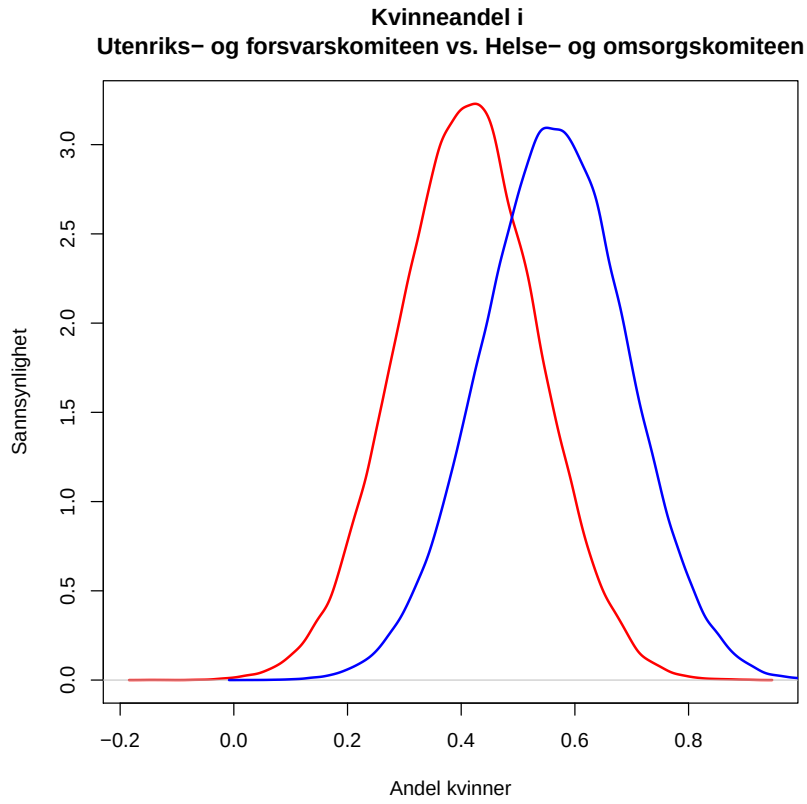
Ekstern gyldighet – hvor tilfeldig er resultatet?

Det betyr at når gjennomsnittet i en gruppe av enheter er svært forskjellig fra gjennomsnittet i en annen gruppe av enheter, så kan vi anta at de to gruppene ikke likner på hverandre i populasjonen heller. De er trukket fra to forskjellige deler av populasjonen. De er signifikant ulike.

Ekstern gyldighet – hvor tilfeldig er resultatet?

Eksempel 1: Ut fra denne teorien kan jeg dedusere dette: Om jeg hadde trukket 100 000 utvalg av populasjonen til Utenriks- og forsvarskomiteen, ville jeg fått 100 000 utvalg med disse gjennomsnittsverdiene for kjønnsfordelingen. Tilsvarende for Helse- og sosialkomiteen. De ser ut til å ha blitt trukket fra ulike populasjonsgrupper.

Ekstern gyldighet – hvor tilfeldig er resultatet?



Ekstern gyldighet – hypotesetesting

Jacobsen snakker om to typer kvantitativ hypotesetesting:

- **T-testen:** som vi nettopp gjennomførte, er laget for å sammenlikne gjennomsnitt. Man starter med en univariat analyse, og ser siden om den multivariate resultatene er signifikant forskjellige. Det vil si: Er de nye resultatene generert av tilfeldigheter, eller har de andre variablene bedret forklaringspotensialet?
- **Kjikvadrat-testen:** er laget for frekvenstabeller. Nok en gang sammenlikner testen hvilken fordeling man har i en univariat analyse med fordelingen i krysstabellen (hvor man har to variabler).

Ekstern gyldighet – feilmargin

For å uttrykke usikkerheten rundt tilfeldigheter beregner jeg en "feilmargin". Den er beregnet på grunnlag av tre ting:

- **Gjennomsnittet i utvalget:** Dette er den mest sannsynlige verdien.
- **Spredningen i utvalget:** Vi vet hvilke verdier som forekommer oftest, og vi vet hvilke verdier som forekommer skjelden. Jo større spredning jeg har i utvalget mitt, jo vanskeligere vil det være å gjette på riktig verdi.
- **Antall enheter:** Jeg vet også at jo større utvalget mitt er, jo nærmere vil gjennomsnittet i utvalget mitt likne på gjennomsnittet i populasjonen.

Ekstern gyldighet – feilmargin

Matematisk betyr det at jeg har en formel med tre elementer for å beregne feilmarginen. Målet er å beregne de mest sannsynlige gjennomsnittsverdiene i et utvalg, gitt informasjonen vi har.

- **Gjennomsnittet i utvalget** Vi starter med å beregne gjennomsnittet i utvalget vi har. I R gjøres dette med funksjonen `mean()`.
- **Standardfeilen (eng: standard error (SE))** Deretter beregner vi et mål på hvor ofte vi tror vi kommer til å ta feil. Det avhenger av hvordan sprednignen er i datasettet, og hvor stort utvalg vi har. Formelen blir $SE = \frac{sd}{\sqrt{(N)}}$
  - **Standardavviket i utvalget** I R gjøres dette med funksjonen `sd()`.
  - **Antall enheter:** I R gjøres dette med funksjonen `length()`.

Ekstern gyldighet – feilmargin

Kloke hoder har regnet ut formen på sannsynlighetsfordelinger og spredning. Når man lager en sikkerhetsmargin på omtrent to ganger standardfeilen rundt den mest sannsynlige verdien (gjennomsnittet), vil 95 av 100 utvalg ha en gjennomsnittsverdi som befinner seg innenfor sikkerhetsmarginen. Det nøyaktige tallet man ganger standardfeilen med kan man lese av en t-tabell. Den uttrykker i praksis sannsynligheten for et utfall (en variabelverdi), gitt utvalgsstørrelsen, spredningen og gjennomsnittet vårt:
$sikkerhetsmargin = gjennomsnittet \pm standardfeilen \times t - verdien$

Ekstern gyldighet – kjønnsfordeling på Stortinget

Eksempel 1:

- Jeg vet at sannsynligheten for å trekke en kvinne fra datasettet mitt om Stortinget er 0.4 (det er andelen kvinner i dataene).
- Jeg kan konstruere en 95 prosents feilmargin rundt denne sannsynligheten: Da får jeg en minimumsverdi på 0.28 og en maksimumsverdi på 0.52. 95 av 100 utvalg gjort fra disse enhetene vil ha en gjennomsnittsverdi innenfor dette spennet.
- Alle verdier utenfor dette spennet er svært usannsynlige, gitt tilfeldighetene. Vi kan dermed anta at i utvalg med andre gjennomsnittsverdier er trukket fra en annen populasjonsgruppe. De er signifikant ulike.

Ekstern gyldighet – kjønnsfordeling på Stortinget

Eksempel 1: Feilmargin for kjønnsfordeling i Stortingsdataene (68 enheter): 0.28 og 0.52. La oss anta at hver komite er et tilfeldig utvalg blant stortingsrepresentanter. Da bør komiteene ha en kvinneandel mellom disse to verdiene.

	Kvinneandel
Arbeids- og sosialkomiteen	0.31
Familie- og kulturkomiteen	0.40
Helse- og omsorgskomiteen	0.56
Kontroll- og konstitusjonskomiteen	0.25
Utenriks- og forsvarskomiteen	0.41

Ut fra tabellen ser vi at kjønnsbalansen i Helse- og omsorgskomiteen og Kontroll- og konstitusjonskomiteen neppe er tilfeldig.

Ekstern gyldighet – hypotesetesting

Hypotesetesting handler om å gjøre tilfeldighetene til vår venn. Vi formulerer to hypoteser:

- H_0 Nullhypotesen sier at resultatene våre er generert av tilfeldigheter. De befinner seg innenfor feilmarginen.
- H_{alt} Den alternative hypotesen sier at resultatene våre er svært *usannsynlige* (gitt tilfeldighetene). De er statistisk *signifikante*.

Hypotesetestingen handler om å teste nullhypotesen først. Når vi forkaster nullhypotesen, står vi igjen med den alternative hypotesen.

Ekstern gyldighet – kjikvadrattesten

Kjikvadrattesten

- **T-testen:** som vi nettopp gjennomførte, er laget for å sammenlikne gjennomsnitt. Man starter med en univariat analyse, og ser siden om den multivariate resultatene er signifikant forskjellige. Det vil si: Er de nye resultatene generert av tilfeldigheter, eller har de andre variablene bedret forklaringspotensialet?
- **Kjikvadrat-testen:** er laget for frekvenstabeller. Nok en gang sammenlikner testen hvilken fordeling man har i en univariat analyse med fordelingen i krysstabellen (hvor man har to variabler).