Multivariate analyser
– bivariate og trivariate sammenhenger

Silje Synnøve Lyder Hermansen

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

Bivariate analyser

Jacobsen (2005), kap. 14

Samvariasjon - hva betyr det?

Samvariasjon betyr at verdier på to ulike variabler tenderer til å opptre sammen. Samvariasjon krever i praksis at vi ser på forksjeller mellom undergrupper enheter: Tenderer enkelte enheter til å få de samme verdiene på variabler? NB: Samvariasjon er *ikke* det samme som sammenheng.

- **Tilfeldigheter:** Det kan være tilfeldigheter som gjør at to fenomener opptrer sammen.
- **Spøriøs effekt:** To fenomener kan ha en felles bakenforliggende årsak som påvirker begge variabler.

Samvariasjon - hva betyr det (forts.)?

Årsak-virkning: Når et fenomen forårsaker et annet, vil vi observere samvariasjon (men all samvariasjon er ikke et tegn på årsaksforhold). I statistikk observerer vi skjelden mekanismene som forårsaker et fenomen, vi observerer konsekvensene av et årsaksforhold (samvariasjon). Da må vi bruke kompensasjonsstrategier:

- I et årsaks-virkningsforhold vil det ene fenomene opptre først, før det andre tar til.
- Vi lener oss på teori.
- Vi lener oss på tidligere kvalitativ forskning.

Bivariate analyser

Etter å ha gjennomført univariate analyser, er det naturlig å fortsette med bivariate analyser: Hvordan står variablene i forhold til hverandre? Vi er interessert i samvariasjon.
Måten vi analyserer samvariasjon på avhenger av målenivået på variabelen. Vi kan alltid behandle variabler med høyere målenivå med teknikker for lavere målenivå. Det motsatte er ikke tilfellet.
Binære variabler/dummyvariabler kan behandles på samtlige måter.

Kategoriske variabler

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

Kategoriske variabler

Vi har i utgangspunktet to/tre måter å gjøre bivariate analyser hvor en eller begge av variablene er kategoriske:

- **Når begge er kategoriske:** Vi kan lage krysstabeller hvor vi viser frekvensen (antall enheter) som har hver spesifikke verdikombinasjon. *Eksempel:* En tabell som viser hvor mange studenter med karkteren A på særemne som går lærerlinjen, hvor mange med karakteren B... osv. Med flere studieretninger vil vi få flere kolonner i tabellen vår.
- **Når en variabel er kategorisk og den andre er metrisk** (eller blir behandlet som metrisk): Gjennomsnitlig verdi for den metriske variabelen i hver kategori av den kategoriske variabelen. *Eksempel:* Gjennomsnitlig forventet lønn for studenter med karakteren A i særemne, gjennomsnitlig lønn for studenter med karakteren B... osv.
- **Når begge variabler er dikotomer:** Kan vi bruke *fi*.

Krysstabeller - datasettet

	Komite	Posisjon	Navn	Kvinne	Regjering	Per
1	Familie- og kulturkomiteen	Leder	Harberg, Svein	0	Nei	
2	Familie- og kulturkomiteen	Første nestleder	Bekkevold, Geir Jørgen	0	Nei	
3	Familie- og kulturkomiteen	Andre nestleder	Aasrud, Rigmor	1	Statsrad	
4	Familie- og kulturkomiteen	Medlem	Liadal, Hege Haukeland	1	Nei	
5	Familie- og kulturkomiteen	Medlem	Løvaas, Kårstein Eidem	0	Nei	
6	Familie- og kulturkomiteen	Medlem	Mandt, Sonja	1	Nei	
7	Familie- og kulturkomiteen	Medlem	Grande, Arild	0	Nei	
8	Familie- og kulturkomiteen	Medlem	Stordalen, Morten	0	Nei	
9	Familie- og kulturkomiteen	Medlem	Thomsen, Ib	0	Nei	
10	Familie- og kulturkomiteen	Medlem	Tønder, Mette	1	Nei	
11	Utenriks- og forsvarskomiteen	Leder	Huitfeldt, Anniken	1	Statsrad	
12	Utenriks- og forsvarskomiteen	Første nestleder	Halleraker, Øyvind	0	Nei	
13	Utenriks- og forsvarskomiteen	Andre nestleder	Norheim, Kristian	0	Nei	
14	Utenriks- og forsvarskomiteen	Medlem	Agdestein, Elin Rodum	1	Nei	
15	Utenriks- og forsvarskomiteen	Medlem	Alexandrova, Regina	1	Nei	
16	Utenriks- og forsvarskomiteen	Medlem	Graham, Sylvi	1	Statssekretær	
17	Utenriks- og forsvarskomiteen	Medlem	Grande, Trine Skei	1	Nei	
18	Utenriks- og forsvarskomiteen	Medlem	Hansen, Svein Roald	0	Nei	
19	Utenriks- og forsvarskomiteen	Medlem	Helleland, Trond	0	Nei	
20	Utenriks- og forsvarskomiteen	Medlem	Hareide, Knut Arild	0	Statsrad	
21	Utenriks- og forsvarskomiteen	Medlem	Navarsete, Liv Signe	1	Statsrad	
22	Utenriks- og forsvarskomiteen	Medlem	Nybakk, Marit	1	Nei	
23	Utenriks- og forsvarskomiteen	Medlem	Sandberg, Per	0	Nei	
24	Utenriks- og forsvarskomiteen	Medlem	Simensen, Kåre	0	Nei	
25	Utenriks- og forsvarskomiteen	Medlem	Solhjell, Bård Vegar	0	Statsrad	
26	Utenriks- og forsvarskomiteen	Medlem	Stoltenberg, Jens	0	Statsrad	
27	Utenriks- og forsvarskomiteen	Medlem	Tybring-Gjedde, Christian	0	Nei	

Krysstabeller - når begge variabler er kategoriske

Når begge variabler er kategoriske er krysstabeller en god start for en bivariat analyse.
La oss si at jeg er interessert i å definere hvilke komiteer som er prestisjetunge i Stortinget. Jeg har gjort et utvalg på tre komiteer for perioden 2013-2017, og operasjonaliserer "prestisjetung" som antall komitemedlemmer med regjeringserfaring. Det finnes flere typer regjeringserfaringer: For enheter som har hatt flere stillinger i regjeringen har jeg bare beholdt den øverste stillingen.

Krysstabeller - når begge variabler er kategoriske

- Jeg ønsker med andre ord å lage en krysstabell for variablene "Komite" og "Regjering". Jeg får en tabell hvor jeg teller antall enheter innen hver verdikombinasjon:

	Nei	Statsrad	Statssekretær
Arbeids- og sosialkomiteen	10	1	2
Familie- og kulturkomiteen	9	1	0
Helse- og omsorgskomiteen	15	1	0
Utenriks- og forsvarskomiteen	11	5	1

Table : Frekvensfordeling i undergrupper av enheter

- I øverste rute til venstre finner jeg antall enheter med verdien "Arbeids- og sosialkomiteen" , og som har verdien "Nei".
- Det er ulikt antall tidligere regjeringsmedlemmer i komiteene, men det er vanskelig å finne en klar tendens fordi det også er et ulikt antall medlemmer i hver komité.

Krysstabeller - når begge variabler er kategoriske

- **Marginalfordeling:** Det er vanlig å legge til en kolonne og en rad med totalsummen av fordelingen. Med disse radene kan vi finne frekvensfordelingen for begge variabler. Dette kan vi bruke bl.a. for å lage søylediagram.
- **Antall enheter:** I den siste ruten, nederst til høyre, kan vi også sjekke at vi har regnet riktig: Den viser summen av enheter.

	Nei	Statsrad	Statssekretær	Totalt
Arbeids- og sosialkomiteen	10	1	2	13
Familie- og kulturkomiteen	9	1	0	10
Helse- og omsorgskomiteen	15	1	0	16
Utenriks- og forsvarskomiteen	11	5	1	17
Totalt	45	8	3	56

Table : Frekvensfordeling i undergrupper av enheter – med marginalfordelingen

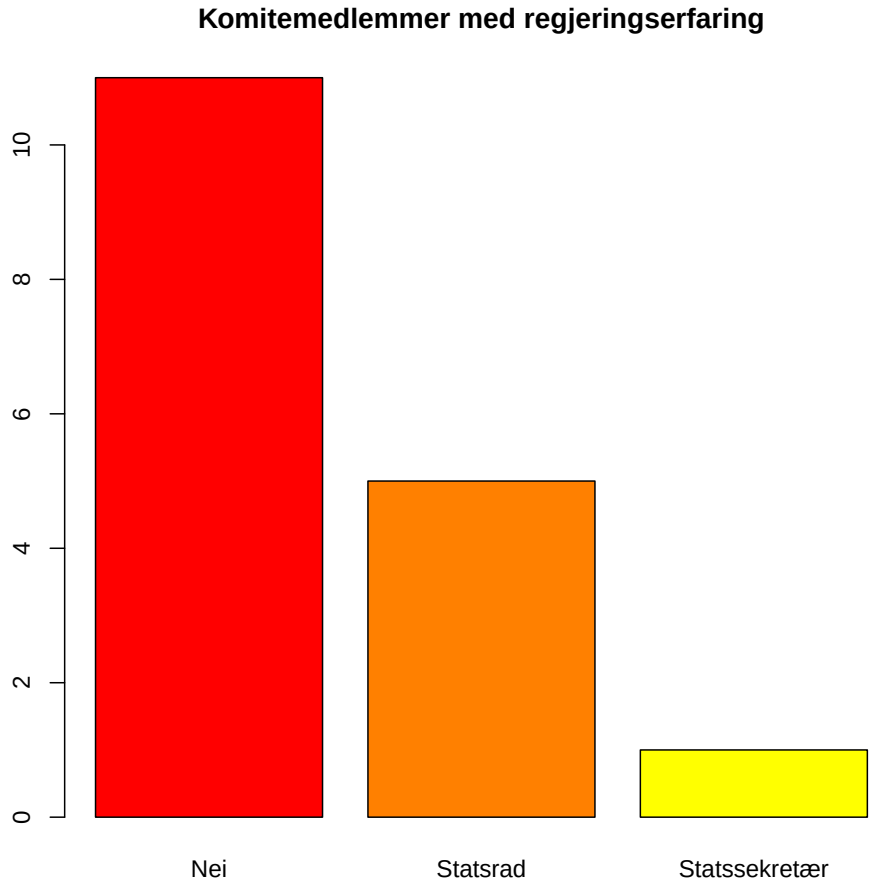


Figure : Univariat frekvensfordeling trukket fra ”total”-raden i en krysstabell.

Krysstabeller - når begge variabler er kategoriske

Relativ fordeling: Den beste måten å gjøre kategoriene sammenliknbare er å prosentuere. Da deler jeg frekvensen for hver verdikategori på det totale antallet enheter i en av kategoriene.

	Nei	Statsrad	Statssekretær
Arbeids- og sosialkomiteen	77	8	15
Familie- og kulturkomiteen	90	10	0
Helse- og omsorgskomiteen	94	6	0
Utenriks- og forsvarskomiteen	65	29	6

Table : Prosentfordeling av komitemedlemmer med regjeringserfaring

Krysstabeller - når begge variabler er kategoriske

- NB: Vi sammenlikner alltid på tvers av prosentueringsretningen. I dette tilfellet ønsker jeg å sammenlikne regjeringserfaring (kolonne). Derfor prosentuerer jeg ved å dele på antall komitemedlemmer (summen for hver linje). Eksempel: $\frac{10}{13} = 0.77$. Om jeg ønsker prosent, må jeg gange med 100. Det er et smaksspørsmål.
- Fra den relative fordelingen ser vi at Utenriks- og forsvarskomiteen er komiteen med størst andel medlemmer med regjeringserfaring, og Familie- og kulturkomiteen har flest medlemmer uten regjeringserfaring. Likevel kan tolkningen bli enda klarere:

Krysstabeller - når begge variabler er kategoriske

Jacobsen foreslår å slå sammen kategorier i analysen for å fremheve resultater. Vi kan for eksempel slå sammen kategorien "Statsråd" og "Statssekretær":

	Nei	Regjering	Total
Arbeids- og sosialkomiteen	10	3	13
Familie- og kulturkomiteen	9	1	10
Helse- og omsorgskomiteen	15	1	16
Utenriks- og forsvarskomiteen	11	6	17
Total	45	11	56

Table : Frekvensfordeling av komitemedlemmer med regjeringserfaring

Krysstabeller - når begge variabler er kategoriske

Med den relative fordelingen kan vi lett rangere komiteene etter prestisje: Utenrikskomiteen er mest prestisjefylt, Arbeids- og sosialkomiteen kommer på andreplass, mens Familie- og kulturkomiteen kommer på tredjeplass og Helse- og omsorgskomiteen havner sist.

	Nei	Regjering
Arbeids- og sosialkomiteen	77	23
Familie- og kulturkomiteen	90	10
Helse- og omsorgskomiteen	94	6
Utenriks- og forsvarskomiteen	65	35

Table : Relativ fordeling av komitemedlemmer med regjeringserfaring

Krysstabeller - en kategorisk og en metrisk variabel

La oss forsøke oss på en annen operasjonalisering av "prestisje": Antall perioder komitemedlemmene har bak seg som fullt medlem av Stortinget (ikke vara). Dermed har vi en metrisk variabel (tellevariabel) som vi ønsker å analysere i forhold til komitemedlemsskap; vi har en kategorisk og en metrisk variabel.

Krysstabeller - en kategorisk og en metrisk variabel

La oss forsøke oss på en annen operasjonalisering av "prestisje": Antall perioder komitemedlemmene har bak seg som fullt medlem av Stortinget (ikke vara). Dermed har vi en metrisk variabel (tellevariabel) som vi ønsker å analysere i forhold til komitemedlemsskap; vi har en kategorisk og en metrisk variabel. Den beste måten å analysere en slik sammenheng på, er å regne ut gjennomsnittsverdien av prestisje blandt medlemmene i hver enkelt komité.

	Perioder
Arbeids- og sosialkomiteen	0.8
Familie- og kulturkomiteen	0.9
Helse- og omsorgskomiteen	1.1
Utenriks- og forsvarskomiteen	2.1

Table : Gjennomsnitlig antall stortingsperioder blant komitemedlemmer

Krysstabeller - en kategorisk og en metrisk variabel

Med denne operasjonaliseringen får vi en annen rangering av komiteenes prestisje: Utenriks- og forsvarskomiteen kommer fortsatt først, siden kommer Familie- og kulturkomiteen, tett etterfulgt av Arbeids- og sosialkomiteen.

Krysstabeller - to dummyvariabler

- *Fi*: Når vi har to binære variabler (dummyvariabler), kan vi lage en enkel firefeltstabell, og regne ut ett enkelt mål for samvariasjon. *Fi* (engelsk: *phi*) går fra 0 til 1, og uttrykker samvariasjonen mellom to binære variabler. 0 viser ingen samvariasjon, 1 viser fullstendig samvariasjon.
- Siden dummyvariabler alltid kan behandles som metriske, vil man ofte ende opp med å bruke andre mål en *fi*.

Ordinale variabler

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

Ordinale variabler

- Jacobsen mener man nesten alltid kan behandle ordinale variabler som metriske (tellevariabler), men det er opp til oss som forskere å avgjøre hvorvidt det er holdbart. Jeg mener det er viktig å huske forskjellen.
- I de fleste tilfeller vil man avgjøre å *behandle* ordinale variabler enten som kategoriske, metriske eller omkode dem til dummyer. Svaret er ikke alltid gitt, og resultatene våre vil ofte avhenge av hvilken beslutning vi treffer.
- *Eksempel:* Hvordan ville dere målt prestisjen til komiteene? Man kunne tenkt seg å rangordne dem fra 1 til 3 (ordinal), men vi kunne også brukt prosentandelen av medlemmer med rejeringserfaring, eller gjennomsnitlig antall stortingsperioder (metrisk). Som vi har sett, blir resultatene anderledes.

Ordinale variabler

I tillegg til analysemetodene for kategoriske variabler, har man også egne mål på korrelasjon for rangordnede/ordinale variabler:

- **Rho** (engelsk: "Spearman"), **Tau** (engelsk: "Kendall") og **Gamma** er alle varianter av det samme målet. De kalles ofte rangkorrelasjonskoeffisienter.

Ordinale variabler

Vi snakker ofte om *korrelasjon* i stedet for samvariasjon når variablene er ordinale eller metriske. Tau, rho og gamma er korrelasjonsmål og varierer fra -1 til 1:

- **Ingen korrelasjon** 0 viser ingen samvariasjon mellom variablene.
- **Positiv korreslasjon** Verdier som nærmer seg 1 viser at enheter med høye verdier på den ene variabelen ofte har høye verdier på den andre variabelen.
- **Negativ korrelasjon** Verdier som nærmer seg -1 viser at enheter med lave verdier på den ene variabelen ofte har høye verdier på den andre variabelen.

Fordelen med disse er at vi har ett mål for samvariasjonen, i stedet for en helt tabell. Dette er spesielt interessant i tilfeller hvor vi har mange kategorier.

Metriske variabler (tellevariabler)

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

Metriske variabler

I tillegg til analysemetodene for variabler på lavere målenivå, har man også egne mål på korrelasjon for metriske variabler:

- **Pearsons R:** Er i likhet med rangkorrelasjonskoeffisientene et enkelt mål på korrelasjon som går fra -1 til 1.
- **Spredningsdiagram:** Man kan tegne/fremstille grafisk en sammenheng.

Metriske variabler - spredningsdiagram

Det er alltid en god idé å utforske en sammenheng grafisk før man bruker enklere/mer plassbesparende mål for korrelasjon. Det er to grunner til dette:

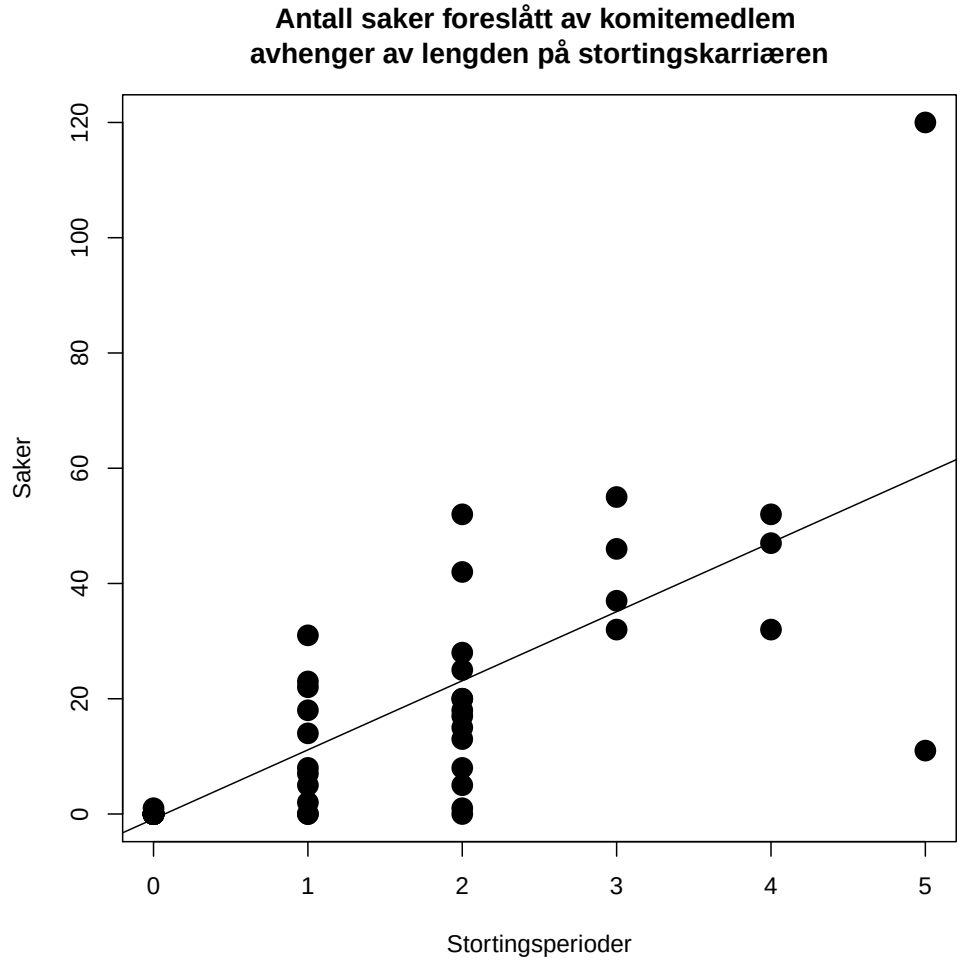
- Det er lett å forstå en grafisk sammenheng når man kan utforske den visuelt.
- Korrelasjonskoeffientene krever at en sammenheng er *lineær*. Dette kan vi se i et spredingngsdiagram.

Metriske variabler - spredningsdiagram

- Et spredningsdiagram viser enhetenes plassering i forhold til verdiene de har på to variabler. Når enhetene befinner seg langs diagonalen, ser vi en samvariasjon.
- Når linjen stiger (fra venstre til høyre) har vi en positiv korrelasjon. Når linjen synker, har vi en negativ korrelasjon.

Metriske variabler - spredningsdiagram

Spredningsdiagram: Pearsons R=0.76



Metriske variabler - Pearsons R

Pearsons R (eller Pearsons "produktmomentkorrelasjonskoeffisient") gir et enhetlig mål på lineære sammenhenger.

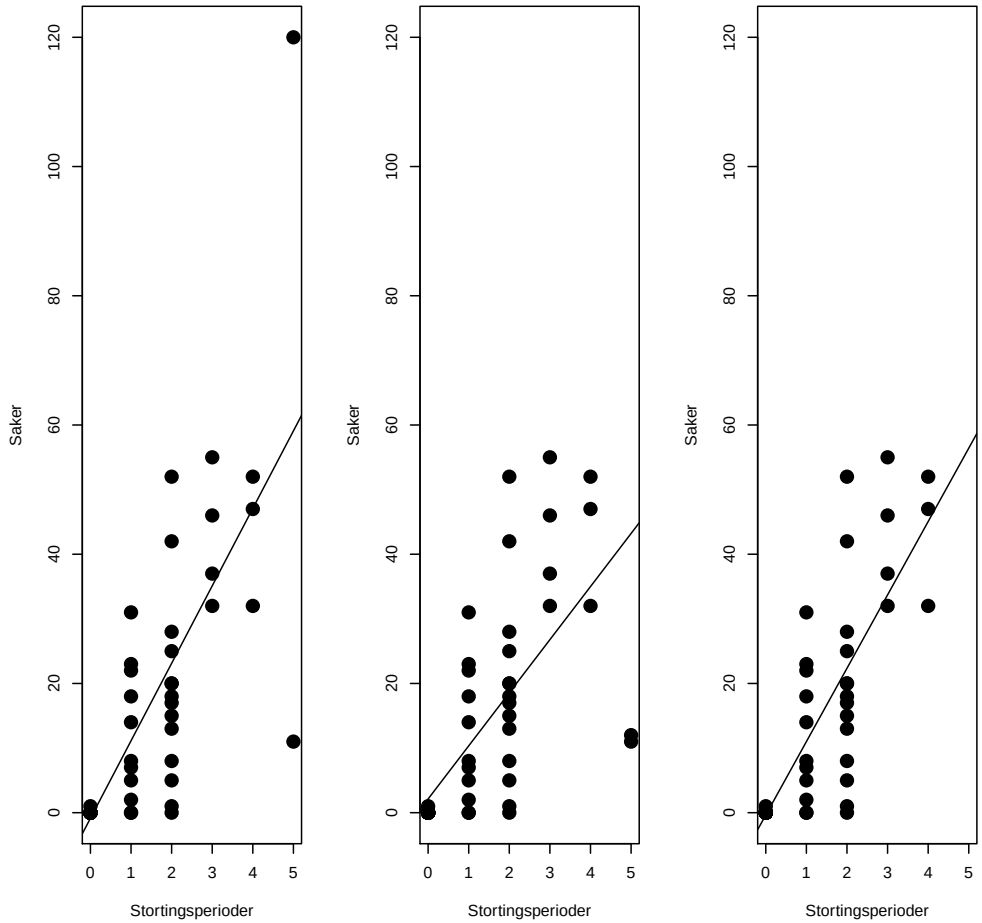
- Varierer fra -1 til 1. Når vi kvadrerer Pearsons R finner vi prosentandelen av variasjon de to variablene har til felles. Dette er veldig nyttig når vi tolker resultater: *Eksempel:* Pearsons R for representantforslag er $0.76^2 = 0.58$. Lengden på karriæren forklarer 58 prosent av variasjonen i lønnsforventning.

Metriske variabler - Pearsons R

- Når sammenhengen ikke er lineær eller når vi har "uteliggere" (ekstreme enheter, enheter som er svært ulike andre), vil Pearsons R ikke fange opp samvariasjonen.

Metriske variabler - kurvilineære sammenhenger

Spredningsdiagram for antall saker foreslått av komitemedlemmer: Pearsons R er hhv. 0.76, 0.69 og 0.82



Trivariate analyser

Jacobsen (2005), kap. 14

Trivariate analyser

Når vi har foretatt bivariate analyser, kan vi begynne med multivariate analyser. I dette faget vil vi begrense oss til analyser av tre variabler (trivariate analyser).

- Fordelen med multivariate analyser er at vi kan **kontrollere** for effekten av én variabel når vi analyserer en annen. Tanken er at substansielle sammenhenger av og til blir maskerte av mer trivielle samvariasjoner.

Trivariate analyser

Når vi har foretatt bivariate analyser, kan vi begynne med multivariate analyser. I dette faget vil vi begrense oss til analyser av tre variabler (trivariate analyser).

- Fordelen med multivariate analyser er at vi kan **kontrollere** for effekten av én variabel når vi analyserer en annen. Tanken er at substansielle sammenhenger av og til blir maskerte av mer trivielle samvariasjoner.
- Ved å kontrollere for andre variabler kan vi finne fram til **spuriøse sammenhenger**.

Trivariate analyser – starter med bivariat analyse

Når vi ser på absolutte tall, ser vi at menn foreslår langt oftere stortingsvedtak enn kvinner:

	Mann	Kvinne
Antall representantforslag	484	212
Prosent representantforslag	70	30

Table : Antall representantforslag

Trivariate analyser – starter med bivariat analyse

Det hele endrer seg litt når vi ser på gjennomsnittlig antall forslag; vi kontrollerer for at det er langt flere menn enn kvinner på stortinget. Noe av sammenhengen mellom forslag og kjønn var spuriøs: Den bakenforliggende variabelen var at det er flere menn enn kvinner i Stortinget.

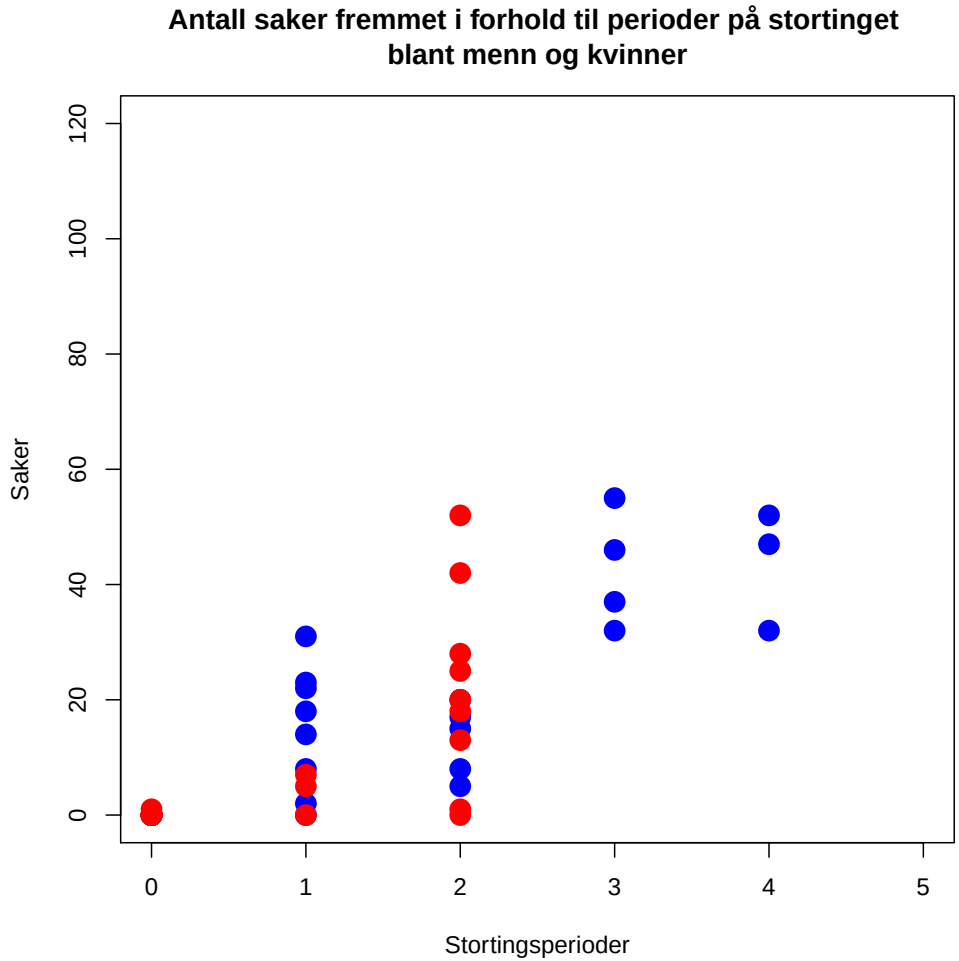
	Mann	Kvinne
Gj. representantforslag	16	9
Prosent representantforslag	63	37

Table : Gjennomsnittlig antall representantforslag

Likevel ser vi fortsatt at kvinner skriver færre forslag enn menn. Kan det forklares med at de har vært kortere på Stortinget?

Trivariate analyser – spredningsdiagram

Når vi har fjernet uteliggerne, og fargelagt enhetene etter kjønn (verdien på den tredje variabelen), finner vi at kvinner (rødt) har generelt vært kortere i Stortinget enn menn (blått).



Trivariate analyser – trivariat tabell

Færre kvinner fremmer forslag - selv kontrollert for tid på Stortinget

	Kvinne, forslag	Kvinne, ingen forslag	Mann, forslag	Mann, ingen forslag
0-1 perioder	21	79	37	63
2-5 perioder	90	10	100	0

Table : Prosent kvinner og menn som fremmer forslag kontrollert for tid på Stortinget