Innhentning (produksjon) av data

Hittil har vi tatt utgangspunkt i analyse av et gitt datasett
uten å spørre om hvordan dataene er innhentet.

Ofte er situasjonen at man ønsker å kunne generalisere resultatene
til en større sammenheng. Da må man ta stilling til:

- Hvordan dataene er innhentet
- Kvalitetene på datene

Datakilder

Det er mange kilder til data. Noen har mer preg av nyheter, og vekten ligger da mer på å få fram spesielle trekk enn på å gi grunnlag for systematisk informasjon.

Anekdotiske data: Dataene representere spesielle situasjoner og verdien ligger i at de er slående uten at de behøver å være representative for noe.

Eksempel Oppslag om tarmkreft i et lokalsamfunn

Men det kan være stor verdi i data selv om de er innhentet for formål som ikke direkte springer ut av det spesielle spørsmålet som man ønsker å belyse.

Tilgjengelige data: Data som er innhentet for andre formål, men som kan være nyttige i andre sammenhenger.

Innhenting av data kan være kostbart, og det er viktig å utnytte informasjon som allerede er tilgjengelig.

Eksempel: Administrative registre, arbeidstaker- arbeidsgiver registeret i NAV er primært til hjelp for å administrere innbetaling og utbetaling av sykepenger. Men det kan også benyttes til forskningsformål som for eksempel å undersøke hvordan langtidsfravær varierer mellom ulike sykdommer.

For innhenting av data er det to systematiske metoder som er viktige, **utvalgsundersøkelser** og **eksperimenter**.

Utvalgsundersøkelser brukes for å kartlegge trekk ved en større populasjon.

For eksempel kan man være interessert i norske stemmerettsberettigedes holdning til EU. Da er alle norske stemmerettsberettigede populasjonen.

For store populasjoner er en **totaltelling** kostbart hvis dataene skal være til å stole på. I stedet nøyer man seg med å undersøke et «omhyggelig» utplukket utvalg.

Utvalgsundersøkelser er et eksempel på observasjonsstudie.

Viktig skille:

Observasjonsstudie: Her observeres enhetene og variable av interesse måles, men man prøver ikke å influere på responsene.

Eksperiment: Her utsettes enhetete med hensikt for en behandling og responsene måles.

Ofte er det av interesse å si noe om hva som hender hvis noe endres. Hva skjer når skatten endres? Blir de frigjorte midlene reinvestert eller brukt til forbruk? Slike ting er vanskelig å svare på uten faktisk å forta endringen eller intervensjonen.

For å belyse årsaksforhold er data fra eksperimenter mest overbevisende. Vi skal se at man da - under noen forutsetninger - kan håndtere underliggende (lurkende) variable.

Forsøksplanlegging

Her skal vi se på prinsippene som ligger bak planleggingen av eksperimenter. Det er tre forhold som er viktige:

1. Sammenligning av ulike behandlinger for grupper av enheter.

2. Hvordan fordele enhetene på gruppene?

3. Ta hensyn til ekstra informasjon.

Noen viktige begreper:

Behandlinger: De ulike betingelsene som enhetene som inngår i eksperimentet utsettes for.

Resultat av eksperimentet: Verdien av variablene som måles for å sammenligne behandlingene.

Hensikten med et eksperiment er å måle responsen i en variabel som resultat av endring i en forklaringsvariabel, spesielt behandlingen. Det er altså (også her) et skille mellom respons- og forklaringsvarible.

I eksperimenter kalles forklaringsvariabelene **faktorer** og verdiene de antar **nivåer** («levels»). Som vi skal se er det også mulig (vanlig) å kombinere flere faktorer

Eksempel: Betydning av klassestørrelse

n= 6385 elever observert i fire år.

Gruppe 1: Regulær, 22-25 elever, 1 lærer
Gruppe 2: Regulær, 22- 25 elever, 1 Lærer + 1 assistent
Gruppe 3: Liten klasse, 13-17 elever.

Målingen er skårer på standardiserte tester etter forsøksperioden.

Faktoren her klassetype som kan anta 3 nivåer.

Mulige underliggende variable er resurser til skolen og familiebakgrunn.

Eksempel: TV-reklame i løpet av et 40 minutters program.

Faktor 1: lengde, 30 sek., 90 sek.
Faktor 2: hyppighet, 1,2 og 5 ganger i løpet av programmet.

Samspill (interaction): Hver faktor kan ha effekt, men sammen kan de forsterke eller nøytralisere hverandre.

		Factor B Repetitions		
		1 time	3 times	5 times
Factor A Length	30 seconds	1	2	3
	90 seconds	4	5	6

I eksperimenter er det mulig **velge** behandlingen (eksponeringen) de ulike enhetene får og så sammenligne grupper etter ulike behandlinger mht responsen.

I observasjonsstudier må man bruke den observerte eksponeringen som kan være sammenblandet med lurkende variable. Om effekten skyldes disse eller behandlingen er det ikke mulig å fastslå uten videre.

I medisin er den såkalte **placebo** effekten vanlig, dvs. at midler uten noen reell virkning gir positiv effekt.
Selve deltagelsen i undersøkelsen være en underliggende variabel.

I laboratorieundersøkelser vil man kanskje gi samme behandling til alle enheter. Da kan man også oppleve placebo-lignende effekter.

For å redusere betydningen av underliggende variable bør man benytte **komparative eksperimenter** der man sammenligner en **kontrollgruppe** og en **behandlingsgrupper** (eller evt. flere behandlingsgrupper).

Individene/enhetene bør helst ikke vite hvilken gruppe de tilhører. Dette kalles et **blindt** eksperiment.

I et **dobbeltblindt** eksperiment vet heller ikke de som utfører eksperimentet hvem som tilhører kontroll- eller behandlingsgruppene.

Det er ikke alltid mulig å lage blinde eksperimenter siden det er umulig å kamuflere behandlingen, f. eks i studier av sentralstimulerende stoffer.

I eksperimenter som ikke har kontrollgruppe eller ikke er blindet kan forhold som organisering av forsøket, utvelgelsen av enhetene og placeboeffekter dominere resultatet.

Forventningsskjeve eksperimenter: Opplegg eller design som systematisk begunstiger spesielle verdier for responsvariabelen.

Det neste spørsmålet er hvordan enhetene fordeles på kontroll- og behandlingsgruppene.

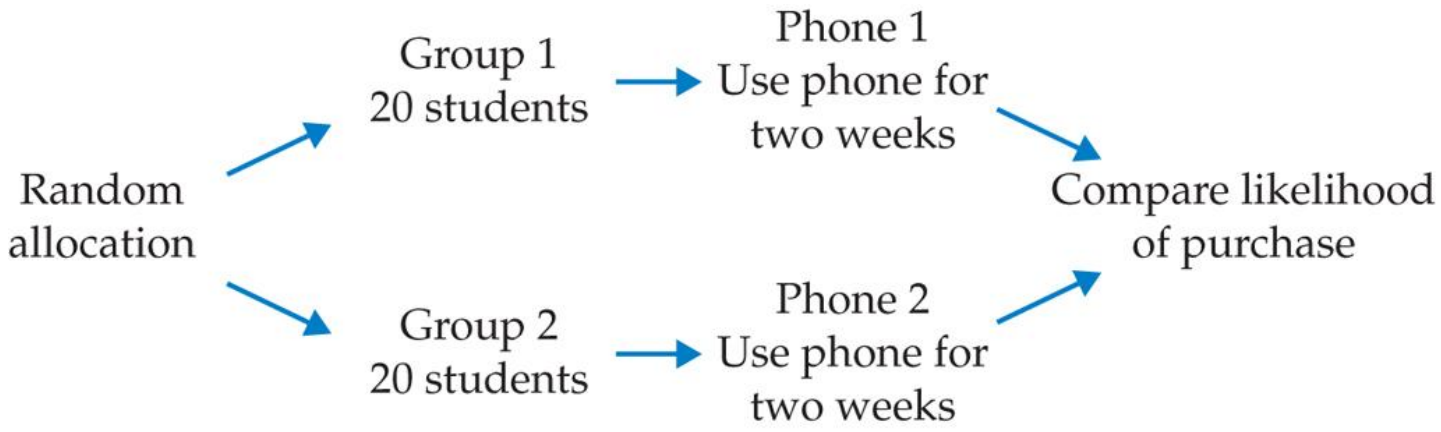
Skal sammenligning av behandlingene ha noen mening må enhetene i de enkelte gruppene i utgangspunktet være noenlunde like.

Matching (tilordning), dvs. å bruke ulike bakgrunnsvariable for å lage homogene grupper, kan være nyttig, men er ikke tilstrekkelig.

Randomisering vil si å fordele enhetene **tilfeldig** på de ulike gruppene og dermed uavhengig av egenskapene til enhetene og vurderingene til de som administrerer eksperimentet.

Eksempel To prototyper av en smartphone skal sammenlignes. 40 studenter deltar, deles tilfeldig inn i to grupper på 20, og det trekkes lodd om hvilken gruppe som får hvilken telefon.

Forsøk med en faktor, telefontype med to nivåer.



Ved bruk av randomisering sikrer man at forskjeller forklares ved

- effekt av behandling

Alternativt kan man uttrykke dette som at

- randomiseringen har resultert i to grupper som systematisk avviker bare på grunn av behandlingen.

Man kan riktignok ha vært uheldig med randomiseringen og fått at en bakgrunnsvariabel tilfeldigvis samsvarer med behandlingen, men sannsynligheten for dette reduseres raskt når antall enheter per behandling øker.

Prinsippene for forsøksplanlegging består derfor i:

1. Sammenlign to eller flere behandlinger med en kontrollgruppen.

2. Randomiser.

3. Flere enheter i hver gruppe bidrar til å redusere sjansen for at eventuelle forskjeller skyldes den tilfeldige mekanismen.

Randomisering kan gjøres ved loddtrekning, dvs hvis de 40 studentene skal deles i to grupper kan man gå fram slik:

1. Studentene gis tall 1-40.
2. 40 lapper med numrene 1-40 legges i en urne.
3. De 20 først utrukne tall bestemmer gruppe 1, resten utgjør den andre gruppa.

Dette er prinsippet bak randomisering, men man kan erstatte lappene/ urnen med en tabell over tilfeldige tall (tabell B i læreboka) eller en rutine i en statistikkpakke.

En tabell over **tilfeldige tall** har egenskapene:

1. Alle tall 0,1,2,...,9 har samme sjanse for å bli trukket ut.
2. Verdien i en posisjon er uavhengig av verdien i en annen posisjon.

Det betyr at alle par (0,0,),(0,1),...,(9,9) har like stor sjanse for å forekomme, og at alle trippler (0,0,0),(0,0,1), (0,0,2),...,(9,9,9) har like stor sjanse for å forekomme.

For å fordele 40 studenter på to grupper kan man derfor:

1. Tildele hver student et tall fra 0 til 39.

2. Starte et vilkårlig sted i tabellen, og plukk ut de 20 studentene som svarer til de 20 første parene av type (0,0),(0,1),...,(3,9).

Ved bruk av en rutine i et statistikkprogram vil hver enhet tilordnes et tall mellom 0 og 1. Tildel enhetene et tall 1,...,n.
For hver enhet trekkes et tilfeldig tall. Sorter disse, og la de enhetene som svarer til de første utgjøre behanlingsgruppa.

Randomisering av 10 enheter.

Excel

	A	B
1	1	
2	2	
3	3	
4	4	
5	5	
6	6	
7	7	
8	8	
9	9	
10	10	
11		

(a)

Excel

	A	B
1	1	0.925672
2	2	0.893959
3	3	0.548247
4	4	0.349591
5	5	0.118440
6	6	0.390180
7	7	0.760262
8	8	0.077044
9	9	0.348467
10	10	0.601167

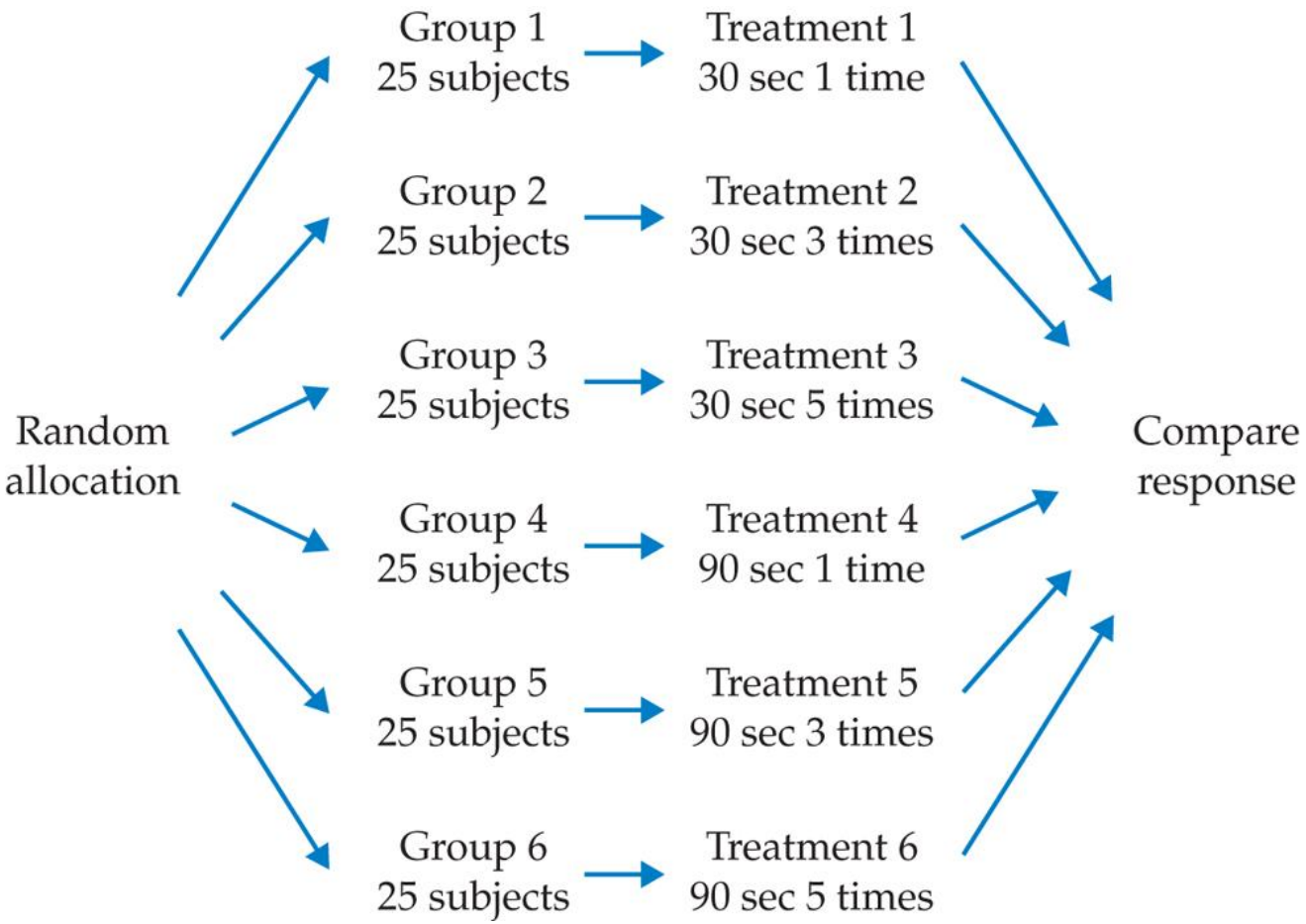
(b)

Excel

	A	B
1	8	0.077044
2	5	0.118440
3	9	0.348467
4	4	0.349591
5	6	0.390180
6	3	0.548247
7	10	0.601167
8	7	0.760262
9	2	0.893959
10	1	0.925672
11		

(c)

Opplegg for komplett randomisert forsøk for sammenligning av seks behandlinger.



Vi har sett på den ideelle metoden for å gjennomføre et forsøk. Men i praksis vil man ofte støte på problemer. I noen situasjoner er det umulig å gjennomføre blinde eksperimenter, som vi har sett.

Mangel på realisme er en annen faktor. Det kan være vanskelig å gjenskape kunstig den situasjonen man vil studere.

Eksempel Studie av risikoaversjon. For å sammenligne reaksjonen på ulike strategier ved aksjehandel gir man en gruppe studenter et beløp og sammenligner reaksjonene. Poenget er at beløpene ikke vil være i nærheten av de beløp som er involvert i de situasjonene man er interessert i.

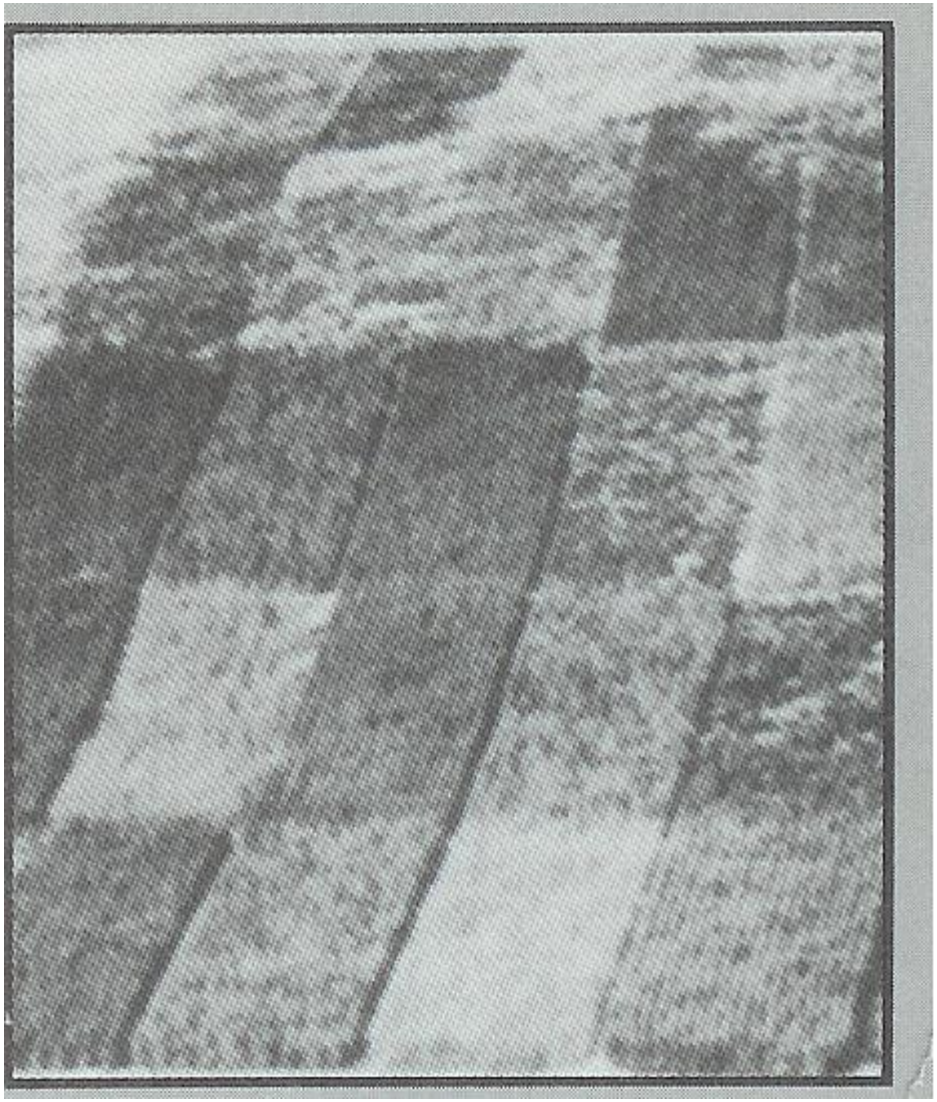
Slik **mangel på realisme** anser MMcC for å være den største svakheten ved eksperimenter.

Vi skal nå se på det tredje viktige punktet ved forsøksplanlegging: Hvordan ekstra informasjon kan tas hensyn til for å forbedre forsøksplanen.

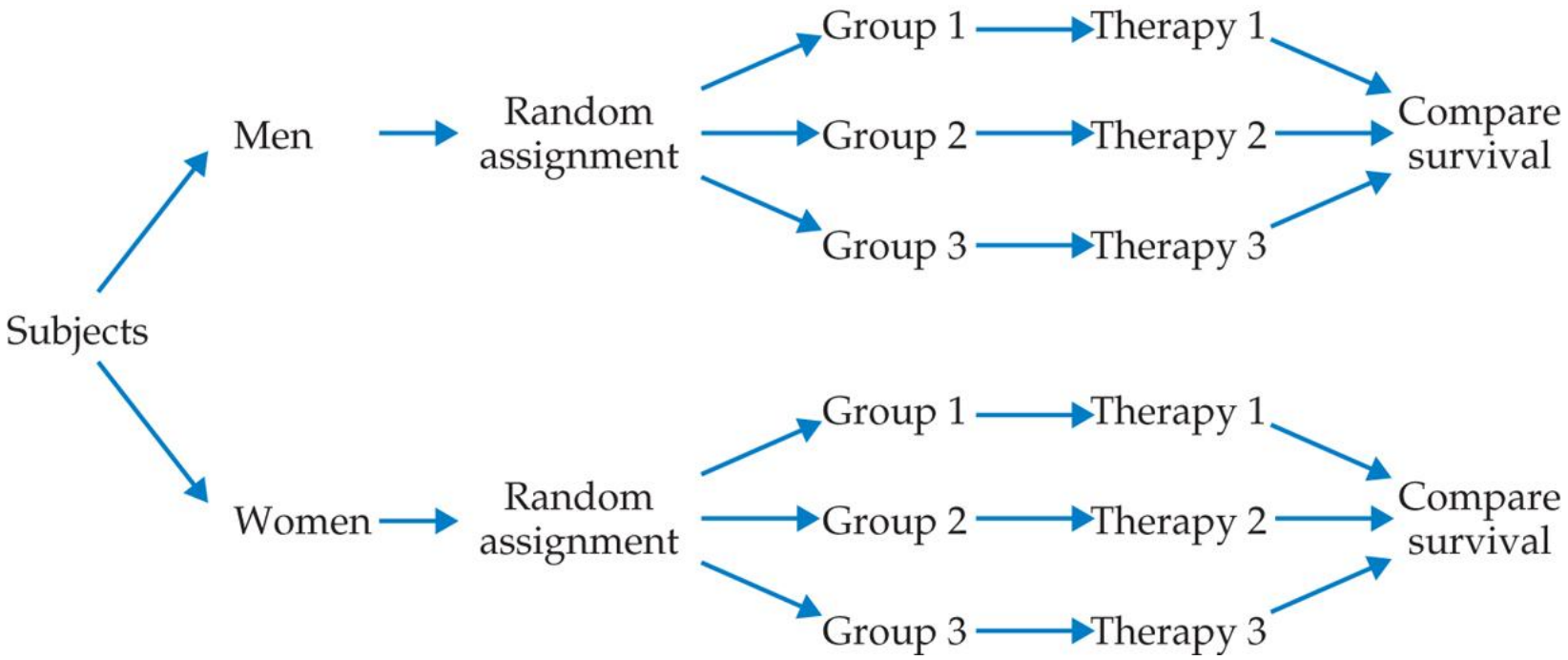
Blokk: Dette er en gruppe enheter som man før eksperimentet vet er like etter noen variable som man vet påvirker responsen på behandlingen.
Et spesialtilfelle av blokker er **matched par**: Her består hver blokk av to like enheter.

Blokk design: Her blir randomiseringen til de ulike gruppene gjort innnen hver blokk.

Eksempel Sammenligning av ulike kornsorter.
Her vil blokkene utgjøres av delene av åkrene som har det samme jordsmonnet.



Eksempel Blocking i en kreftstudie. Menn og kvinner er blokkene. En faktor, tre nivåer.



Poenget med bruk av blokker er å ta hensyn til de viktigste kildene til variasjon . Randomisering vil nøytralisere effekten av de øvrige slik at eventuelle effekter kan tilbakeføres til de ulike behandlingene.

Spesielt for **«mathed pair design»** kan man se på differansen i respons mellom den behandlede og kontrollenheten.

Man har en viss frihet i å bestemme hva «par» er som man ser av følgende alternative opplegg for sammenligning av de to prototypene av smartphone.

Eksempel Smartphone prototyper.

Her lar vi hver student bruke begge telefonene i perioder på 14 dager hver og sammenligner hvor tilfredse de er.
Hvilken telefon som skal brukes først bestemmes ved randomisering.

Her er «parene» de to periodene, og eventuelle variasjoner mellom studentene tas hensyn til ved at man sammenligner forskjellen i tilfredshet for hver enkelt student.

Utvalgsundersøkelser

Hensikten her er å finne andelen eller størrelsen på et kjennetegn i en stor populasjon. Å undersøke hele populasjonen er dyrt og også kan resultere i dårlige data. Derfor undersøkes bare en del.

Populasjon: Gruppen av enheter man vil ha informasjon om.
Utvalg: De enhetene som plukkes ut for nærmere undersøkelse.

Et trekkeregister er en liste over enheter som man bruker for å lage utvalget.

Eksempel Populasjon: Alle lærere i grunn- og videregående skole
Trekkeregister: Medlemslister i de fire lærerorganisasjonene. Hva med de som ikke er medlemmer?

Som man ser behøver ikke populasjon og trekkeregister være samsvarende. Dette bør rapporteres sammen med resultatene.

Et annen viktig opplysning er **responsraten**, dvs. andelen av de som blir trukket ut som faktisk svarer. Den er ofte svært lav i spørreundersøkelser.

Det er svært viktig å lage utvalgene slik at de gjenspeiler populasjonen.

Eksempel Presidentvalg mellom Roosevelt og Landon i 1936. Bruk av galt trekkeregister (Literary Digest abonnement).

Frivillig utvalg: Her består utvalget av enheter som på egen hånd melder sin interesse.

Eksempel Innringing etter en TV-debatt, «likes» på facebook.

Galt konstruerte utvalg gir resultater som ikke gjenspeiler de tilsvarende i populasjonen. Det kalles **skjevhet.**

Tilfeldige utvalg er en metode for å redusere og kontrollere skjevhet. Der fjernes innflytelsen både av den som lager utvalget og de som deltar.

Det enkleste tilfeldige utvalg er:

Enkelt tilfeldig utvalg: Det består av n enheter fra poulasjonen/trekkeregisteret som er valgt slik at alle utvalg av størrelse n har like stor sjanse for å bli valgt.

Merk: I enkelt tilfeldige utvalg har alle enheter samme sjanse til å bli med.

Enkle tilfeldige utvalg lages som følger:

1. Lag en liste over enhetene i populasjonen/trekkeregisteret.
2. Trekk et tilfeldig tall for hver enhet.
3. Sorter de tilfeldige tallene.
4. Et enkelt tilfeldig utvalg av størrelse n består i å velge de n enhetene som svarer til de n minste tilfeldige tallene.

Generelt kalles utvalg der man kjenner sannsynligheten for å velge hvert mulig utvalg for **sannsynlighetsutvalg**.

Enkle tilfeldige utvalg er sannsynlighetsutvalg, men det finnes også andre.

Ofte er populasjonen delt inn i grupper, f. eks, kvinner og menn. Slike grupper kalles i utvalgsteori for **strata.**

Stratifisert tilfeldig utvalg. Her deles først populasjonen i grupper av enheter som er like. Deretter trekkes enkle tilfeldige utvalg innen hvert strata og disse kombineres til et helt utvalg.

Strataene lages før utvalget trekkes og baseres på bakgrunnsinformasjon. Det er m.a.o. en måte å trekke slike opplysninger inn i trekkingen av utvalg.

Strata i utvalgsundersøkelser har samme rolle som blokker i forsøksplanlegging.

Ved enkle tilfeldige utvalg kan man risikere å ende opp med «sære» utvalg, også disse har den samme sannsynlighetet for å bli valgt. Stratifisering reduserer dette problemet.

Flertrinns tilfeldige utvalg brukes der det er manuell interviewing. Bruken er mindre etter at telefoninterview ble vanlig.
Hensikten er ofte å spare reisekostnader for interviewerne.
Et enkelt tilfeldig utvalg av den norske befolkning ville resultere i individer spredt ut over hele Norge. Det blir dyrt i lengden med personlig oppmøte. I stedet kan man dele landet i geografiske områder, trekke ut noen av disse og så benytte stratifiserte utvalg innen de utvalgte områdene.

Utvalgsundersøkelser er ikke uten problemer.

1. Underdekning fordi populasjon og trekkeregister spriker.
2. Lav responsrate.

I tillegg kommer målefeil i spørreundersøkelser:
Hvordan spørsmålene er formulert kan ha betydning, og
ved telefonintervjuer kan intervjuerne «legge svar i munnen på» respondentene. I tillegg kan enkelt typer spørsmål være mer følsomme enn andre.

Mot statistisk inferens

Vi ønsker å kartlegge norske stemmeberettiges holdning til EU.
Anta at andelen mot norsk medlemskap i EU er lik p.

Tenk at vi trekker 10 enkle utvalg på 1000 personer.
Da kan vi beregne de empiriske andelene $\hat{p}_1,\ldots,\hat{p}_{10}$.

Trolig er alle forskjellige, men i nærheten av p.

Å si noe om hvordan $\hat{p}_1,\ldots,\hat{p}_{10}$, eller $\hat{p}_1$ hvis vi bare tar ett utvalg, forholder seg til p er et eksempel på det som kalles **statistisk inferens**.

Følgende skille er fundamentalt i denne sammenhengen:

Parameter: Detter er en teoretisk størrelse som beskriver populasjonen. I eksemplet foran var p parameteren.

Statistikk: Dette er en empirisk størrelse som kan beregnes i utvalget. (I *norsk* statistisk ordbruk er det vanlig å bruke uttrykket observator for statistikk). I eksemplet var statistikken $\hat{p}$.

At $\hat{p}_1,\ldots,\hat{p}_{10}$ varierer skyldes den tilfeldige mekanismen utvalget er konstruert i henhold til.

Tilfeldige utvalg har **to** fordeler.

For **det første**, som vi har sett, elimneres innflytelse av de som står for undersøkelsen og deltagerne når utvalget konstrueres.

Den **andre** store gevinsten er at trekker man gjentatte tilfeldige utvalg (enkle tilfeldige, stratifiserte, to-trinns osv.) vil variasjonen fra utvalg til utvalg følge et bestemt mønster.

Det betyr at om man lager et histogram over verdiene til $\hat{p}_1,\ldots,\hat{p}_s$ og lar s vokse så vil histogrammene nærme seg et "grensehistogram".

Videre kan histogrammene brukes til å angi andelen som ligger i ulike intervaller. Dessuten kan man beregne størrelser som senter og standardavvik i denne fordelingen.

Et grunnleggende sitat fra læreboka:

”All statistical inference is based on one idea: to see how trustworthy a procedure is, ask what would happen if we repeated it many time”

En måte å få en ide om hvordan «grensehistogramet» ser ut er ved bruk av tilfeldige tall og simuleringer på datamaskin.

Et enkelt tilfeldig utvalg kan simuleres på følgende måte.

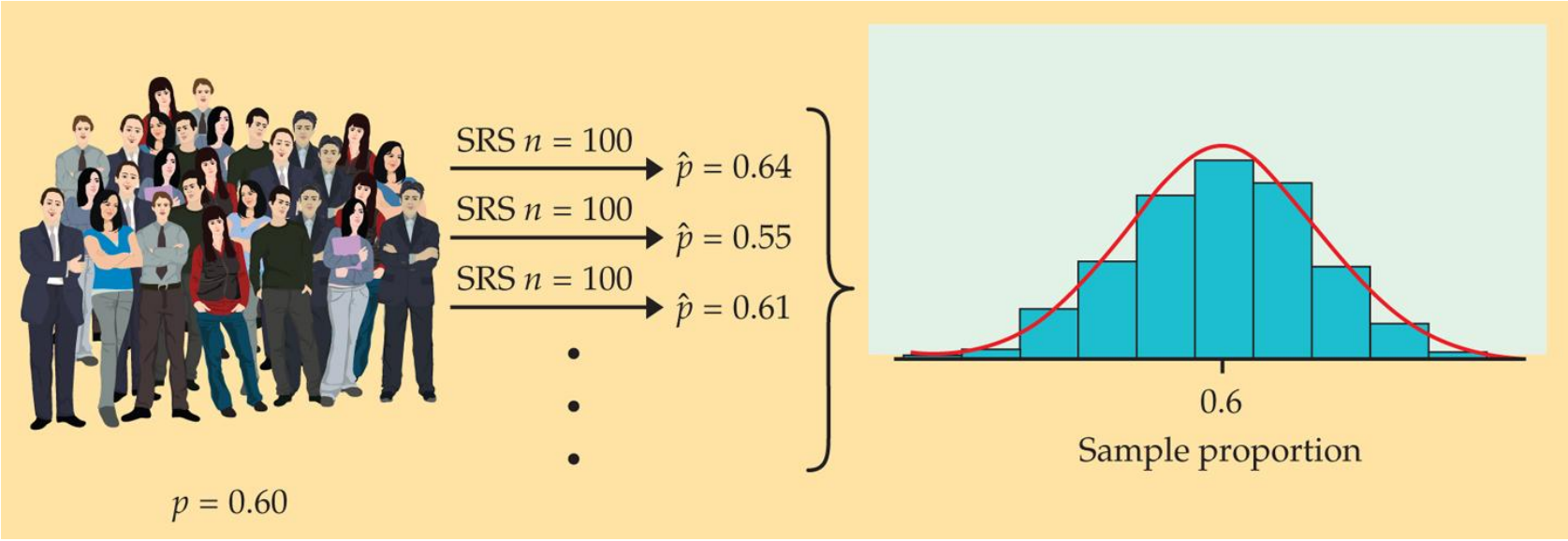
Hvis andelen som har et kjennetegn i populasjonen er, f. eks 0.6, lar vi siffrene 0,1,2,3,4,5 i en tabell over tilfeldige tall svare til at enheten har kjennetegnet.

Tilsvarende vil siffrene 6,7,8,9 svare til at enheten ikke har kjennegnet.

Andelen av 100 tilfeldige tall som har sifrene 0,1,2,3,4,5 vil nå svare til andelen i et enkelt tilfeldig utvalg som har kjennetegnet.

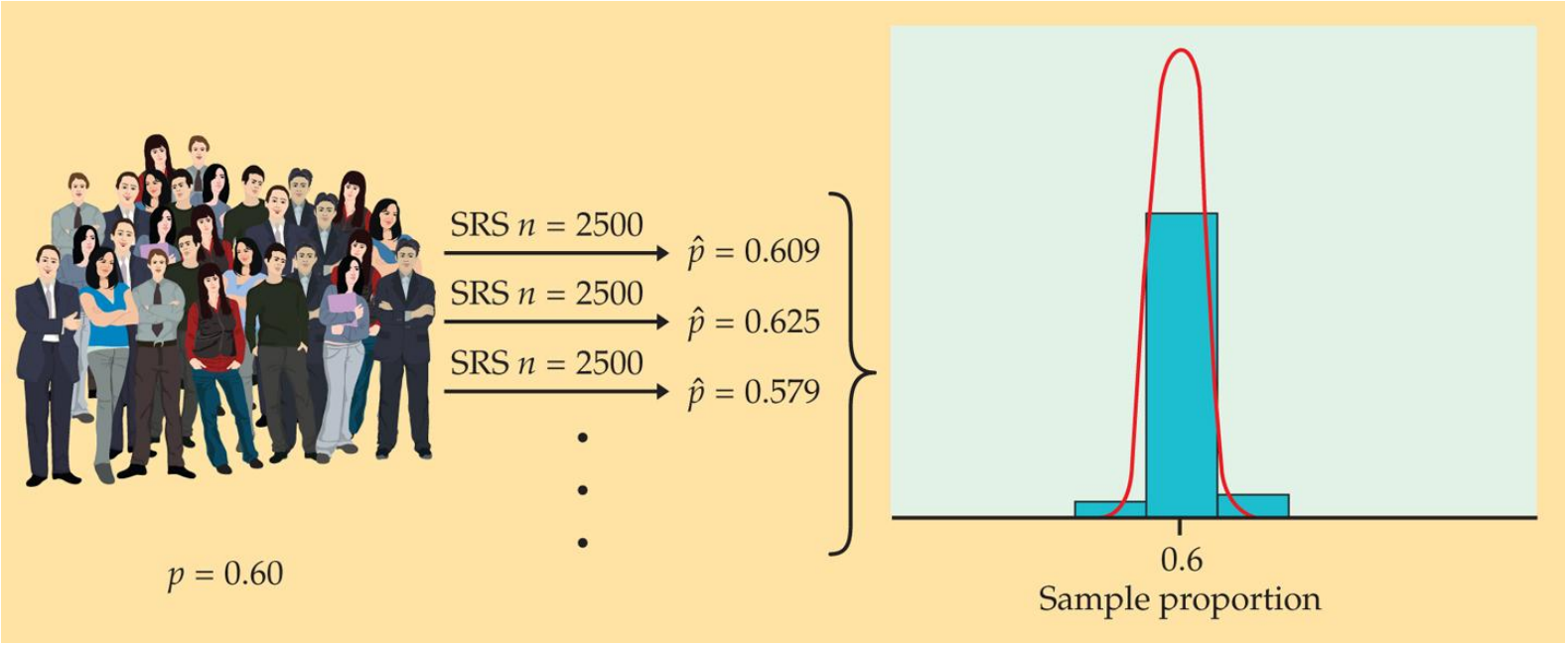
Ved å gjenta denne prosedyren s ganger finner vi $\hat{p}_1,\ldots,\hat{p}_s$.

Eksempel: Anta at p=0.6 og utvalgstørrelsen 100, samt antall gjenntak s=1000. Historgramet er baser på de 1000 gjentakene.



og her er p=0.6 og s=1000 forsatt, men utvalgsstørrelsen økt til 2500 .

Som man ser er spredningen nå mye mindre selv om senteret fortsatt er det samme.



Utvalgsfordeling: Utvalgsfordelingen til en statistikk er fordelingen av verdiene som statistikken antar i alle mulige utvalg av samme størrelse trukke fra samme populasjon.

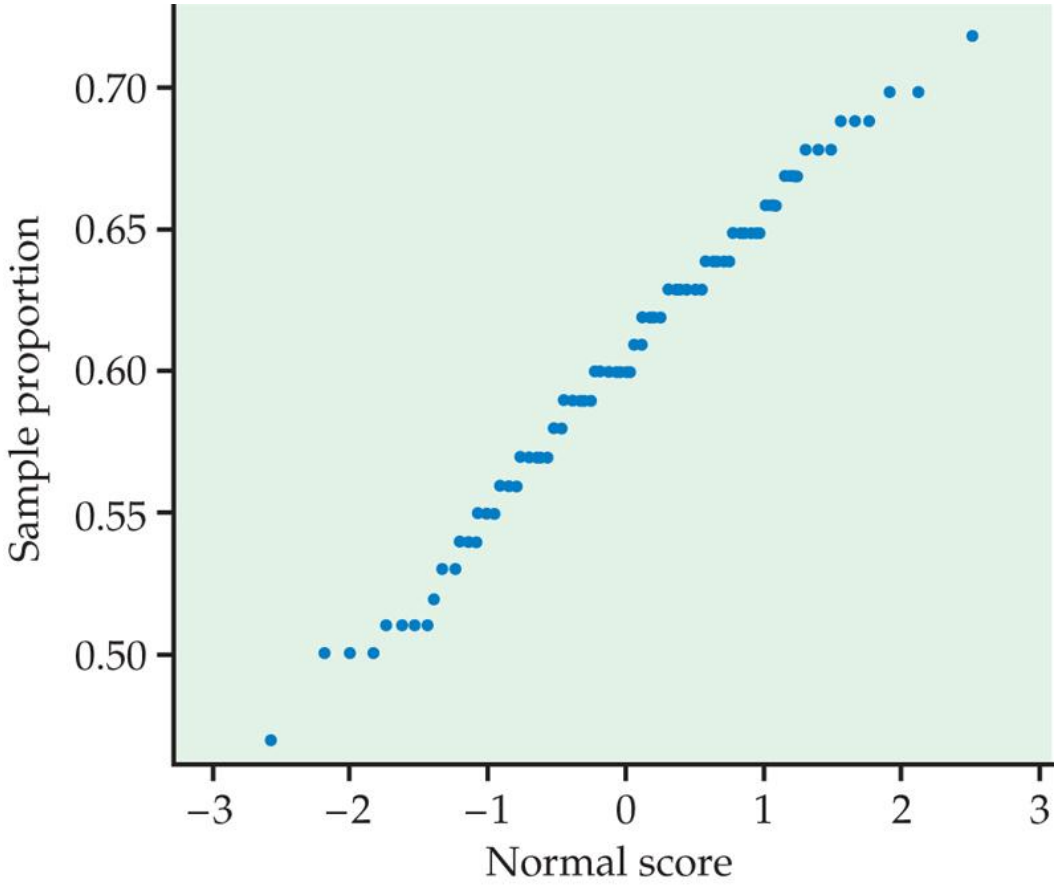
Dette er en ideel (tenkt) størrelse og for store utvalg kan og store populasjoner kan den i praksis ikke beregnes (eller simuleres) eksakt. Men tilnærmelsen kan bli så gode vi ønsker bare ved å la antall gjentak øke.

Man kan nå beskrive fordelingen med den metodene vi har sett på tidligere: senter, spredning, form (symmetrisk/skjev?).

- For både utvalg av størrelse 100 og 2500 er sentret i fordeling til $\hat{p}$ i nærheten av populasjonsverdien p. Det er videre liten eller ingen skjevhet.

- Spredningen er mye mindre i fordelingen til $\hat{p}$ for et utvalg på 2500 enn et utvalg på 100 .

Det er også mulig å sammenligne den tilnærmede utvalgsfordelingen med en normalfordelingskurve, her for utvalgsstørrelse 100 og for 10% av de s=1000 simulerte verdiene.



Her er noen viktige egenskaper og begreper i forbindelse med utvalgsfordelinger.

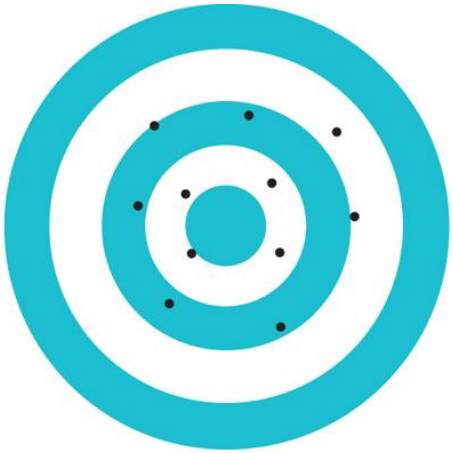
Skjevhet dreier seg om sentret i utvalgsfordelingen. En estimator er **forventningsrett** hvis forventningen i utvalgsfordelingen er lik verdien til parameteren som skal estimeres.

Hvor variabel eller usikker en statistikk er, beskrives av spredningen i utvalgsfordelingen. Spredningen bestemmes av utvalget og størrelsen n på utvalget. Spredningen synker med utvalgsstørrelsen.

Neste figur illustrerer de to egenskapene ved en utvalgsfordeling.

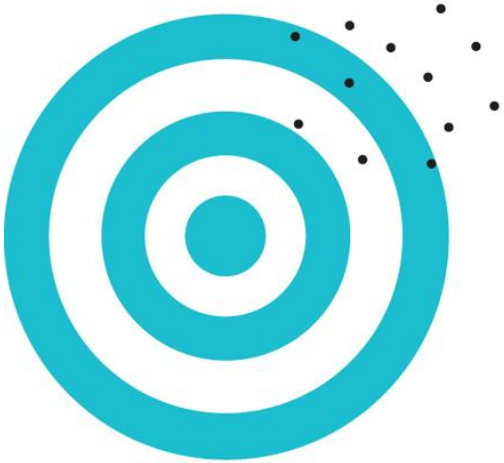
High bias, low variability

(a)



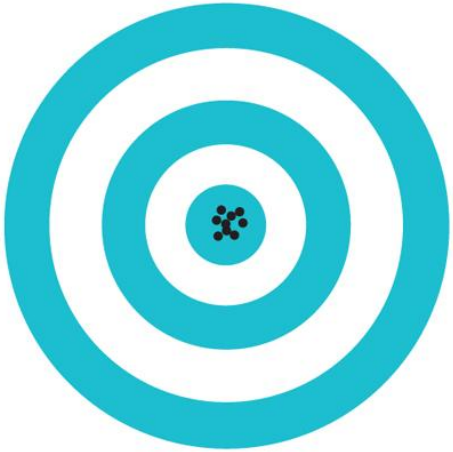
Low bias, high variability

(b)



High bias, high variability

(c)



The ideal: low bias, low variability

(d)

Enkle tilfeldige utvalg der man trekker utvalg av størrelse n fra en populasjon av størrelse N og estimerer andelen i populasjonen som har et kjennetegn med andelen i utvalget som har kjennetegnet , illustrerer noen viktige prinsipper:

- Enkle tilfeldige utvalg gir forventningsrette estimatorer.
- For å redusere spredningen til estimatoren kan man bruke større utvalg.
- Hvor variabel eller usikker statistikken er, avhenger ikke av størrelsen N på populasjonen, så lenge N er mer enn hundre ganger n. En måte å forstå dette på er å tenke på forskjellen mellom **utvalg med og uten tilbakelegging.**

Etikk

Innhenting av data reiser ofte etiske spørsmål.

Dette gjelder spesielt når dataene angår personer,
men også i andre sammenhenger er det også viktig, som i bruk av forsøksdyr. Derfor er det noen kjøreregler,

- Studier som involverer eksperimenter med dyr og menesker skal rutinemessig godkjennes av nevder.

- Deltagerne i undersøkelser skal informeres og samtykke i å delta.

- Dataene skal være konfidensielle og ved offentligjøring skal deltagerne ikke kunne identifiseres. Dette betyr f. eks at tabeller ikke kan være for detaljerte.