MET 3431 Statistikk
Forelesning 1: Introduksjon til Statistikk

Eivind Eriksen
BI, Institutt for Samfunnsøkonomi

10. januar 2012

Litteratur

- Mario Triola
  Essentials of statistics, 4. utg. 2011 - **Obligatorisk litteratur**
  Pensumliste og forslag til oppgaver fra læreboken finnes i It's Learning
- MyLab
  Online tilleggsmateriale til læreboken - **Kan være nyttig**
  Selvtester, e-bok, videoer mm. Informasjon om hvordan man får tilgang finnes i It's Learning
- Dag Einar Sommervoll
  Mattespettboken, 1. utg. 2009 - **Kan være nyttig**
  Mest for de som har matematikkfobi

Dataverktøy

SAS JMP

- Lastes ned fra BI's hjemmesider
- Vi vil gjennomgå JMP allerede i Forelesning 2 - ta med laptop og last helst ned programmet og installer det før forelesningen
- Brukes i flervalgsprøver
- Utskrift fra JMP vil bli gitt på eksamen

Arbeidskrav

Flervalgsprøver

- Det vil bli gitt 8 flervalgsprøver som besvares i It's Learning
- Minst 5 av 8 flervalgsprøver må være bestått for å kunne gå opp til eksamen
- Maks 3 forsøk på hver flervalgsprøve - Minst 30% riktig for å bestå
- En ukes frist på hver flervalgsprøve fra publiseringsdato

Eksamen

Avsluttende eksamen

- Individuell skriftlig 5 timers eksamen teller 100% i dette kurset
- Eksamen blir 30. mai 2012 kl 09.00 - 14.00
- Før eksamen
  - Vær godt forberedt i pensum!
  - Ha gjort ukeoppgavene fra læreboken!!
  - Ha jobbet seriøst med statistikkprogrammet SAS JMP!!
  - Ha fått godkjent minst 5 av 8 arbeidskrav!!!

Forelesninger

Forelesningsplan

- Gjør deg kjent med forelesningsplanen, som inneholder en oversikt over alle forelesninger (emner, aktuelle seksjoner i lærebok, oppgaver mm). Den finnes på It's Learning og oppdateres fortløpende.
- Før hver forelesning:
  - Sjekk forelesningsplanen, og les gjennom aktuelle deler av pensum.
- Etter hver forelesning:
  - Finn forelesningsnotatene på forelesningsplanen, og sjekk at du har forstått det som har blitt gjennomgått.
  - Finn oppgavene fra læreboken som er anbefalt i forelesningsplanen, og jobb med disse oppgavene før neste forelesning.

Målsetninger

Mål med kurset

- Lære enkel statistisk analyse
- Kunne hente informasjon ut i fra data
- Lære kritisk tenkning

Begreper

Viktige begreper i dag

- Populasjon
- Utvalg/stikkprøve
- Parameter og observator
- Kvantitative og kvalitative data
- Diskrete og kontinuerlige data
- Målenivå
- Noen metoder for utvelgning:
  - Tilfeldig
  - Stratifisert
  - Klynge
  - Bekvemmelighet

Hva er statistikk?

Statistikk

Statistikk handler om hvordan vi best

- Skaffer oss et godt utvalg
- Analyserer data fra utvalget for å finne sammenhenger
  - Fornuftig generalisering fra et begrenset antall observasjoner
  - **Skille det *generelle* fra det *tilfeldige***

Transylvania-effekten: Blir flere innlagt på mentalsykehus når det er fullmåne?

Figur: Gj.snittlig antall innleggelser pr dag

Måned	Før fullmåne	Under fullm.	Etter fullm.
Aug	6.4	5.0	5.8
Sept	7.1	13.0	9.2
Okt	6.5	14.0	7.9
Nov	8.6	12 .0	7.7
Des	8.1	6.0	11.0
Jan	10.4	9.0	12.9
Feb	11.5	13.0	13.5
Mars	13.8	16.0	13.1
Apr	15.4	25.0	15.8
Mai	15.7	13.0	13.3
Juni	11.7	14.0	12.8
Juli	15.8	20.0	14.5
Gj.snitt	10.9	13.3	11.4

- Tallene viser at i gjennomsnitt ble flere innlagt under fullmåne
- Er det tilfeldig?

Stikkprøver og statistisk analyse

Fase 1

Det begynner med at man lurer på noe:

- Kjenner mer enn 80% av norske forbrukere til merkevaren Norwegian?
- Er det en sammenheng mellom kjønn og type mobiltelefon?
- Er sannsynligheten for at aksjekursen går opp i morgen avhengig av om den gikk opp i dag?
- Appelerer Skandiabanken til ungdom, mens Postbanken appelerer til eldre?

Fase 2

For å finne svar skaffer man først data.

- Den delen av virkeligheten man er interessert i kalles *populasjonen*. Populasjonen kan være:
  - Alle norske bankkunder
  - Alle kvinner og menn med mobiltelefon
  - Alle velgerne i Hå kommune
- Ofte er populasjonen for stor. Det motiverer studiet av et *utvalg* i stedet for hele populasjonen:
  - 670 norske bankkunder
  - 43 kvinner og menn med mobiltelefon
  - 300 velgerne i Hå kommune

Fase 3

Statistisk analyse av stikkprøven. Hvilken metode vi bruker avhenger av kjennetegnene til data

Data

I spørreundersøkelser og markedsanalyser samler vi inn data for å studere sammenhenger:

- Hva er den mest hensiktsmessige måten å samle inn data?
- Hvordan beskriver vi data på en hensiktsmessig og enkel måte?

Data

Som regel er dette observasjonene (svar på spørreskjema, kjønn, målinger av lønn, høyde) som vi har samlet inn

Statistikk

er metodene for å samle inn og beskrive data, og metodene som benyttes til å generalisere fra utvalg til populasjon (inferens).

- Deskriptiv statistikk - Kap. 1, 2 og 3
- Statistisk inferens - Kap. 4 til 11

Populasjon

hele samlingen av folk, enheter eller objekter som du er interessert i å studere

Utvalg eller stikkprøve

en **subpopulasjon** eller **undergruppe** av populasjonen

Som regel så ønsker vi å sette sammen utvalget slik at det gir et godt bilde av populasjonens sammensetning. For at dette skal være tilfelle, så bør utvalgets enheter (helst) velges tilfeldig fra populasjonen

Søppel inn, søppel ut

Dersom stikkprøven ikke er samlet inn på en ordentlig måte, så er all verdens statistiske triks nytteløse

Viktige begreper

Parameter

et kjennetegn (et tall) ved populasjonen (f.eks. gjennomsnittet)

Som regel ønsker vi å estimere parametrene til populasjonen, og statistiske metoder brukes til å best mulig estimere parametrene

Observator (*statistic* på engelsk)

et tall som beskriver et kjennetegn ved utvalget/stikkprøven (f.eks. utvalgsgjennomsnittet)

Viktige begreper

En måte å klassifisere data er ved å skille mellom kvalitative og kvantitative data

Kvalitative (eller kategoriske) data

data som *kun* kan deles inn i forskjellige kategorier

Eksempler: Studieretningene på BI, kjønn, nasjonalitet, osv. Mao. det gir ikke mening å rangere kategoriene til kvalitative data, eller å si noe om differansen til kategoriene

Kvantitative (eller numeriske) data

tall som representerer tellinger eller målinger

Eksempel: Lønn til ansatte i kommunen, alderen til BI studenter, osv.

Kvantitative data

Kvantitative data kan videre deles opp i *diskrete* og *kontinuerlige*

Diskrete data

når de tillatte verdiene er endelig mange, eller tellbart mange. Eksempel: Antall treff på en nettside.

Kontinuerlige data

når de tillatte verdiene er uendelig mange og uten gap

Eksempel: Tid, Strømforbruket til en bedrift, temperatur, osv.

De 4 Målenivåer

En annen måte å klassifisere data på er ved å skille mellom fire målenivåer:

Nominalnivå: Det gir *kun* mening å skille mellom verdiene til data. Eksempel: Studieretninger på BI, kjønn, nasjonalitet, osv.

Ordinalnivå: I tillegg til å skille mellom verdiene til data, så gir det også mening å rangere verdiene. Eksempler: Ønskene på en ønskeliste, preferanser, svar på holdningsspørsmål (enig, delvis enig, uenig), osv.

Intervallnivå: I tillegg til å skille og rangordne verdiene til data, så gir det også mening å sammenligne distanser mellom verdiene. Eksempler: Årstall, temperatur, osv. NB: Data på intervallnivå har ikke et naturlig nullpunkt

Forholdstallnivå: I tillegg til å skille, rangordne og sammenligne distansen mellom verdier, så gir det også mening å si noe om forhold mellom verdier. Eksempler: Priser, alder, osv.

De 4 Målenivåer - oppsummering

1. Nominal - Bare kategorier
2. Ordinal - Kategorier som kan rangeres eller ordnes i rekkefølge
3. Interval - Differanser er OK, men ikke naturlig nullpunkt
4. Forholdstall (*ratio* på engelsk) - Differanser og et naturlig nullpunkt

Oppsummering

Vi har til nå sett på

- Definisjoner og begreper knyttet til data
- Parameter vs. observator
- Typer av data (kvantitative og kvalitative)
- Målenivå til data

Sunn fornuft

Tenkning

Det er viktig at du hele tiden tenker fornuftig i statistikk. Matematikk er viktig, men du må også tenke i tillegg!

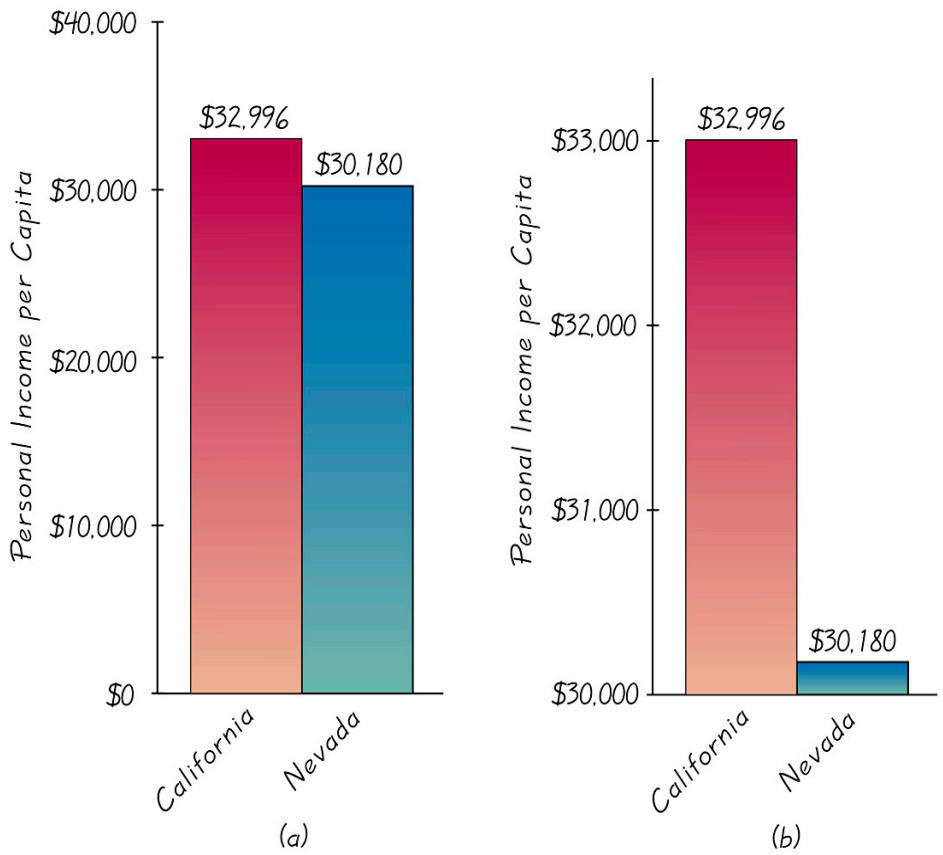
Vanlige svakheter ved utvalg:

- Selvseleksjon (*voluntary response* på engelsk). Respondentene bestemmer selv om de vil være med i spørreundersøkelsen. Da vil konklusjonene dine bare gjelde den gruppen som ønsker å svare på spørreskjema
- For små utvalg. Jo større, jo bedre

Statistisk signifikant *vs.* praktisk relevans

Dårlige grafer

Se på tallene i grafen slik at du ikke blir lurt av formen. Hvilken graf kan du bli lurt av?



Typer studier.

Observasjonell studie

Her måler og observerer du folk eller objekter uten å gripe inn. Du bare måler det som er der. Eksempel: Meningsmålinger, antall klikk på internettreklame, osv.

Eksperimentell studie

Her manipulerer du folk eller objekter. Du sammenligner hva som skjer med en gruppe som har fått behandling med en gruppe som ikke har fått behandling. Så sammenligner du gruppene. Eksempel: Blindtest av Cola vs. Pepsi

Noen metoder for seleksjon

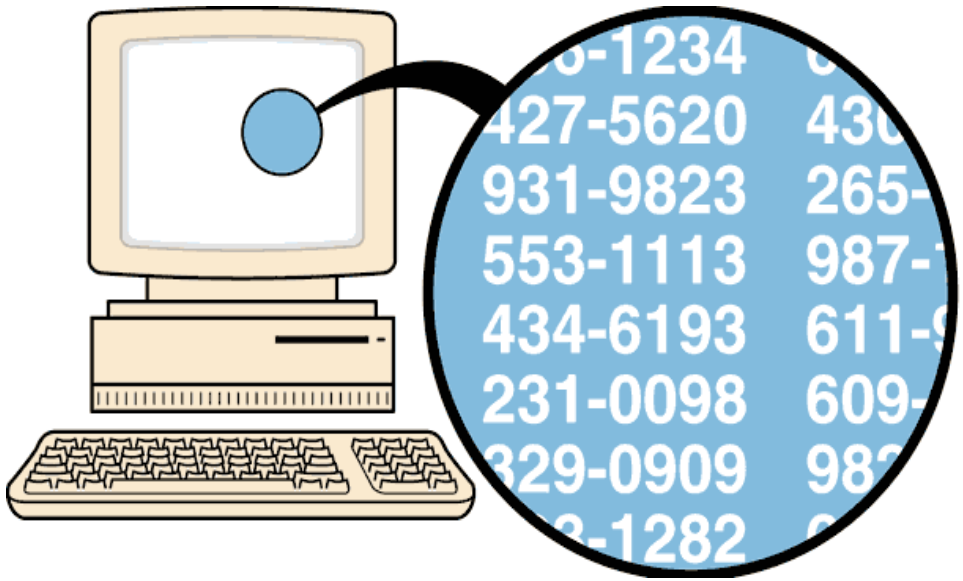
- Tilfeldig
- Systematisk
- Bekvemmelighet
- Stratifisert
- Klynge

Tilfeldige utvalg

- Tilfeldig utvelging: Man velger fra populasjonen slik at hvert individ velges med samme sjanse
- Enkel tilfeldig utvelging: Man velger slik at alle utvalg av størrelse n velges med samme sjanse

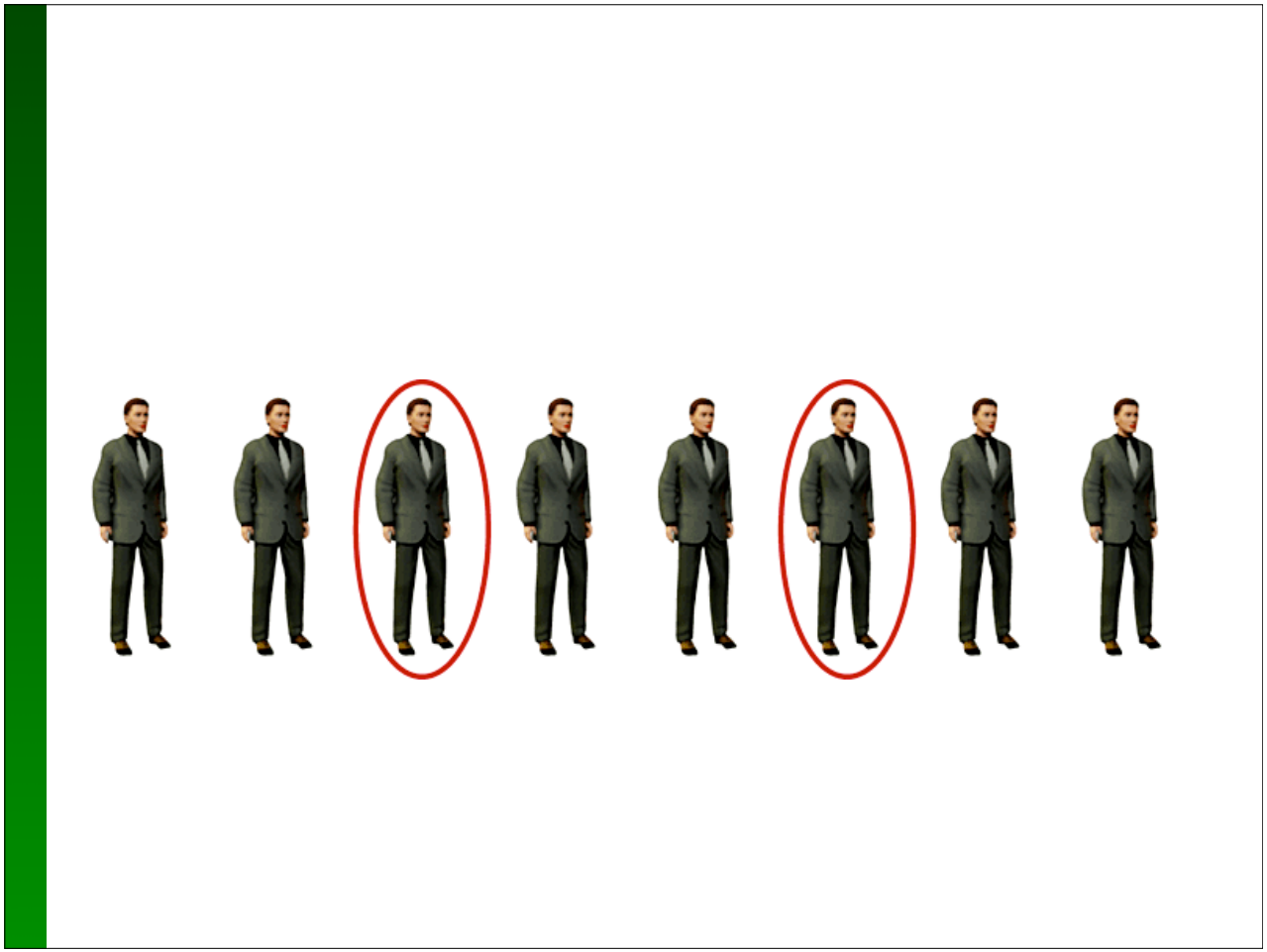
Tilfeldige utvalg.

Hvert individ har lik sjanse til å bli valgt ut. Datamaskiner blir ofte brukt til å generere tilfeldige telefonnumre.



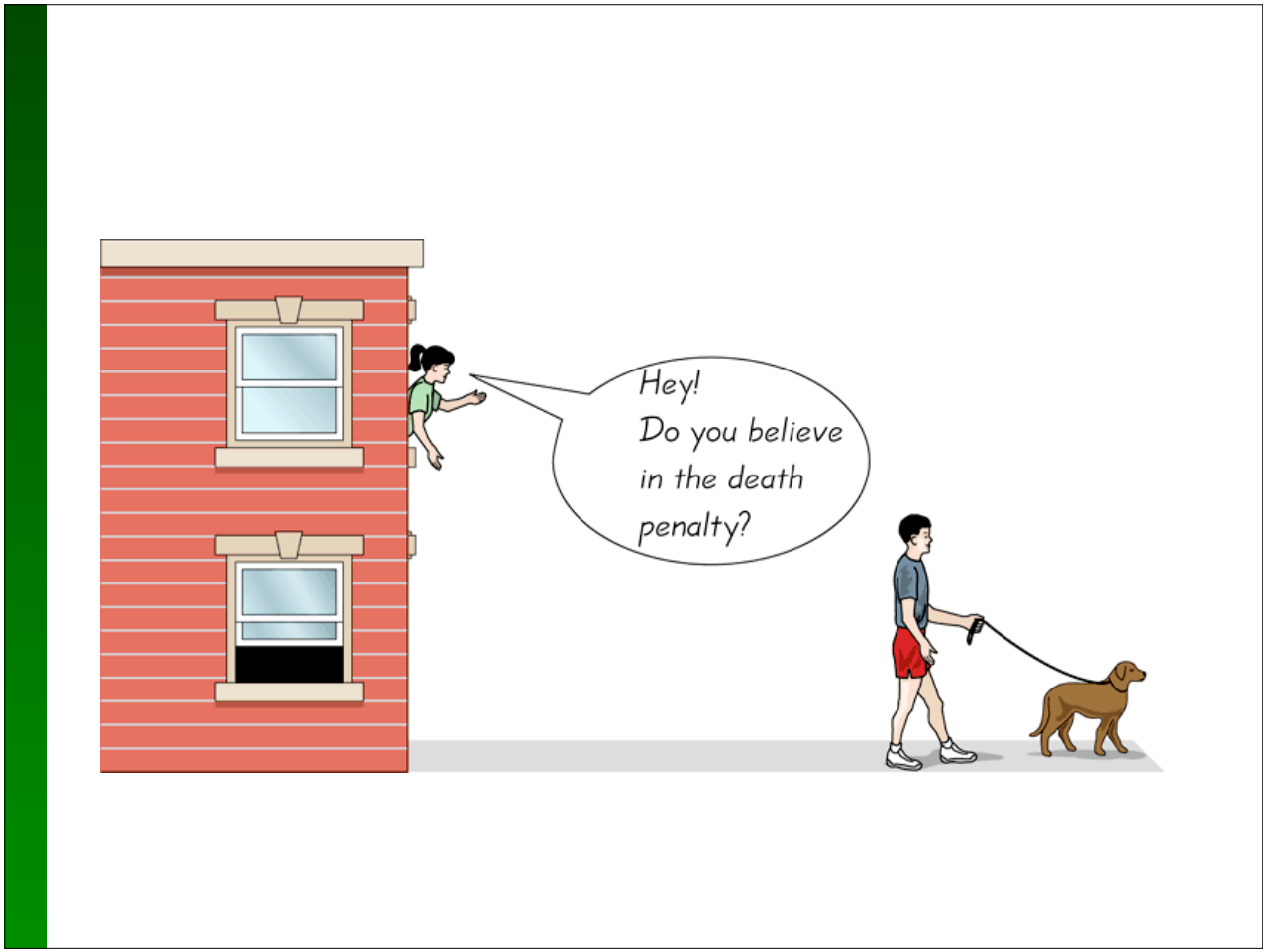
Systematiske utvalg.

Man velger ut f.eks. hvert tredje element i populasjonen.



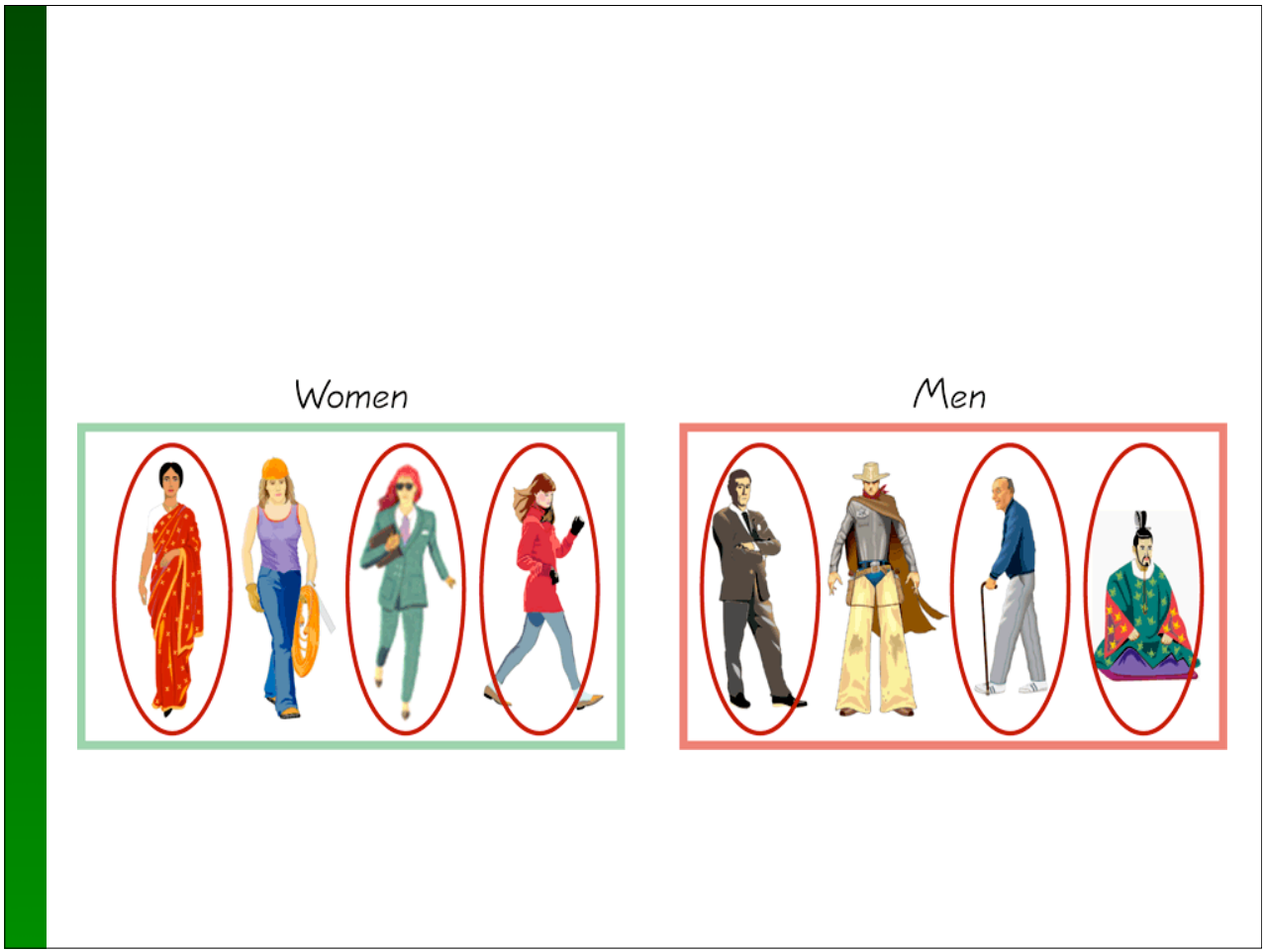
Bekvemmelighetsutvalg.

Man velger ut de personene som er lettest å få tak i.



Stratifiserte utvalg

Del populasjonen inn i minst to undergrupper (stratum), og trekk et utvalg for hvert strata.



Klynge utvalg.

Del populasjonen inn i seksjoner/klynger. Velg tilfeldig noen av klyngene og velg **alle** medlemmene i hver klynge.