Notater
Documents
2016/23

Statistics Norway
Statistisk sentralbyrå

Jan F. Bjørnstad

En introduksjon i statistiske metoder for offisiell statistikk

Notater 23/2016

Jan F. Bjørnstad

En introduksjon i statistiske metoder for offisiell statistikk

Statistisk sentralbyrå • Statistics Norway
Oslo–Kongsvinger

Notater I denne serien publiseres dokumentasjon, metodebeskrivelser, modellbeskrivelser og standarder.

Publisert september 2016

ISBN 978-82-537-9389-4 (elektronisk)
ISSN 1891-5906
Emne: Virksomheter, foretak og regnskap

Trykk: Statistisk sentralbyrå

Standardtegn i tabeller	**Symbol**
Tall kan ikke forekomme	.
Oppgave mangler	..
Oppgave mangler foreløpig	…
Tall kan ikke offentliggjøres	:
Null	-
Mindre enn 0,5 av den brukte enheten	0
Mindre enn 0,05 av den brukte enheten	0,0
Foreløpig tall	*
Brudd i den loddrette serien	—
Brudd i den vannrette serien	|
Desimaltegn	,

Forord

Dette kompendiet er utarbeidet for SSB-kurset KLAR 311 Introduksjonskurs i statistiske metoder. Det gir en innføring i

- Planlegging av utvalgsundersøkelser, både for person/husholdnings- og bedriftsundersøkelser
- Basis statistikkbegreper og de viktigste estimeringsmetodene i design-basert tilnærming i utvalgsundersøkelser
- Teoretiske og vitenskapelige betraktninger rundt modell-basert og design-basert tilnærming
- Modellbaserte estimeringsmetoder
- Tre forskjellige variansmål
- Behandling av frafall ved vekting og imputering
- Multippel imputering for frafall

Matematiske utledninger og formelbruk er holdt til et minimum, men noe formelbruk er uunngåelig for å tilegne seg en viss basis forståelse av sannsynlighet, viktige statistikkbegreper og det statistiske språket. Kompendiet inkluderer noe mer, delvis avansert, materiale som ikke er med i kurset. Disse temaene er stjernemerket. Kapittel 7 om økonomiske undersøkelser er basert på SSB-kurset av Tora Löfgren og Svetlana Badina. Appendikset om funksjoner i R er skrevet av Melike Ogus Alper.

Statistisk sentralbyrå, 16. 09. 2016

Bjørnar Gundersen

Innhold

1. Innledning

Hovedtemaer i dette kompendiet er:

- Basisbegreper og basisestimatorer i offisiell statistikk
- Stratifisering
- Frafall
- Økonomisk statistikk
- Modellbasert statistisk inferens

Begrepet "statistikk" – hva betyr det? En enkel beskrivelse er å si at det er vitenskapen om analyse og tolkning av data. Mer detaljert, kan vi si

- Vitenskapen for planlegging av undersøkelser, innsamling og presentasjon av tallmateriale, og *metoder for analyse og beslutninger ut fra innsamlede data*.

Data kan f.eks. være et utvalg fra en populasjon av personer, bedrifter eller andre enheter, eller observasjoner av fysiske fenomener.

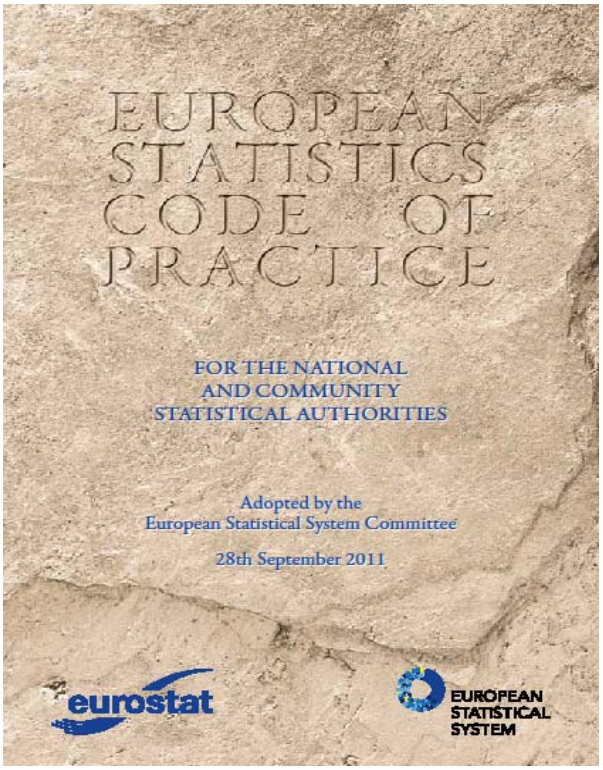
Ordet statistikk brukes også om de innsamlede og analyserte dataene. Opprinnelig ble statistikk brukt om beskrivelser av stats- eller samfunnsforhold, første gang i 1662.

1.1. European Statistics Code of Practice og SSBs Virksomhetsmodell

Statistiske metoder for offisiell statistikk er innenfor rammen av europeiske prinsipper for kvalitet i alle aspekter når det gjelder statistikk.

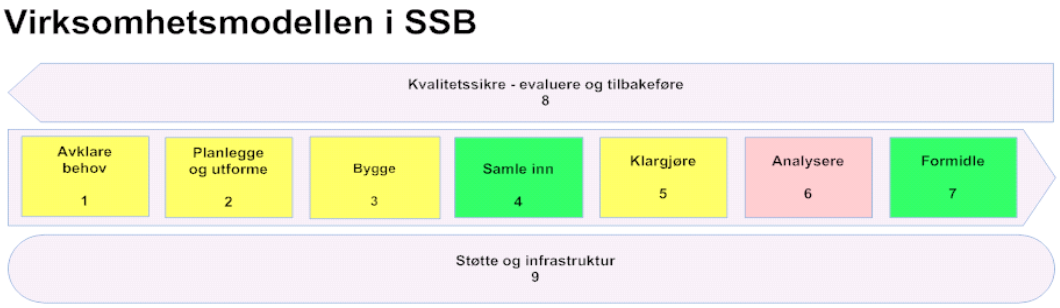
European Statistics Code of Practice er europeiske retningslinjer (kvalitetsprinsipper) for offisiell statistikk. Den er basert på 15 prinsipper som danner en ramme for kvalitet i statistikken, delt inn i tre hovedtemaer.

- Institusjonelle forhold
  - Faglig uavhengighet
  - Mandat for datainnsamling
  - Tilstrekkelige ressurser
  - Kvalitetsbevissthet
  - Konfidensialitet
  - Upartiskhet og objektivitet
- Statistiske prosesser
  - Gode metoder
  - Egnede statistiske prosedyrer
  - Rimelige krav til oppgavegiverne
  - Kostnadseffektivitet
- Statistiske produkter
  - Relevans
  - Nøyaktighet og pålitelighet
  - Aktualitet og punktlighet
  - Sammenheng og sammenlignbarhet
  - Tilgjengelighet og klarhet



Det siste hovedtema er for kvalitet i selve de publiserte offisielle statistikkene. Her er statistiske metoder sentrale for andre og fjerde punkt.

Virksomhetsmodellen i SSB er en detaljert oversikt over hva som inngår i statistiske prosesser:



Virksomhetsmodellen er basert på en internasjonal standard og det er en referanse for:

- Dokumentasjon
- Systemer og metoder: Standardisering!
- Arbeidsrutiner
- Ressursbruk
- Risikovurderinger (f.eks. for feil)

Modellen ligger internt i SSB på http://www.byranettet.ssb.no/Tema/faglig

Den europeiske Code of Practice utgjør sammen med virksomhetsmodellen et kvalitetsrammeverk for SSB og andre europeiske statistikkbyråer.

Tonivå versjonen av virksomhetsmodellen indikerer hva hovednivåene inneholder. Fargene indikerer hvor bra punktene ovenfor er oppfyllt i SSB pr. dags dato.

1.2. Statistiske metoder i SSB

Offisiell statistikk er statistikk som publiseres for allmenheten av SSB eller annet statlig organ. Statistiske metoder er sentral og nødvendig for

- forståelse av statistikken
- kvalitetssikring: nøyaktighet og pålitelighet
- effektivisering av statistikkproduksjonen

I SSBs strategi presiseres det at de beste statistiske metoder skal benyttes for å sikre effektivitet og kvalitet. I denne sammhengen er det viktig med en basis forståelse av statistiske begrep og statistiske metoder i

statistikkproduksjons- seksjonene. Hensikten med dette kompendiet er å gi en innføring i statistiske prinsipper og metoder i statistikkproduksjonen for å:

- lære basisbegreper og tolkninger i den statistiske vitenskapen
- få en bedre forståelse av det statistiske språket

Temaer hvor den statistiske vitenskapen er mest sentral innen statistikkproduksjon av offisiell statistikk er:

- Teoretisk utdyping innen utvalgsundersøkelser
- Spesielt gjelder det
    - planlegging og de viktigste utvalgsplanene, både innen person/husholdnings- og bedrift/foretaks-undersøkelser
    - basis statistikkbegreper og de viktigste estimeringsmetodene i vanlig tradisjonell designbasert tilnærming til utvalgsundersøkelser
    - modellbaserte estimeringsmetoder
    - frafall og statistiske metoder for å rette opp skjevheter på grunn av frafall
    - statistisk inferens: generelt om analyse av en populasjon basert på et utvalg

1.3. Eksempel på tolkning og presentasjon av statistikk, og enkel bruk av statistisk metode

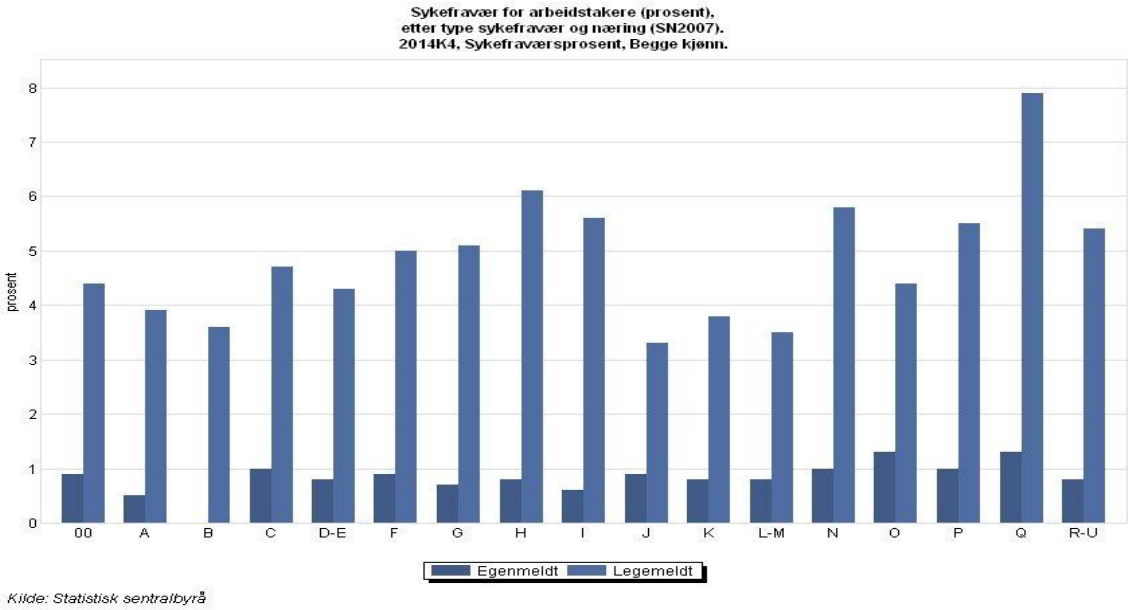
Vi skal se på SSBs sykefraværsstatistikk for 4.kvartal 2014. Den publiseres etter kjønn, alder, bostedskommune og næring. Sykefraværet er basert på både egenmeldinger (utvalgsundersøkelse) og legemeldinger (register). Sykefraværet angis som antall arbeidsdager som er tapt eller antall syke en bestemt arbeidsdag.

Tabell 1.1 Tapte arbeidsdager. Fordelt på kvinner og menn etter meldingstype, i prosent

Kjønn	Totalt sykefravær	Egenmeldt	Legemeldt
Begge	6,4	0,9	5,5
Kvinner	8,2	1,1	7,1
Menn	4,9	0,8	4,1

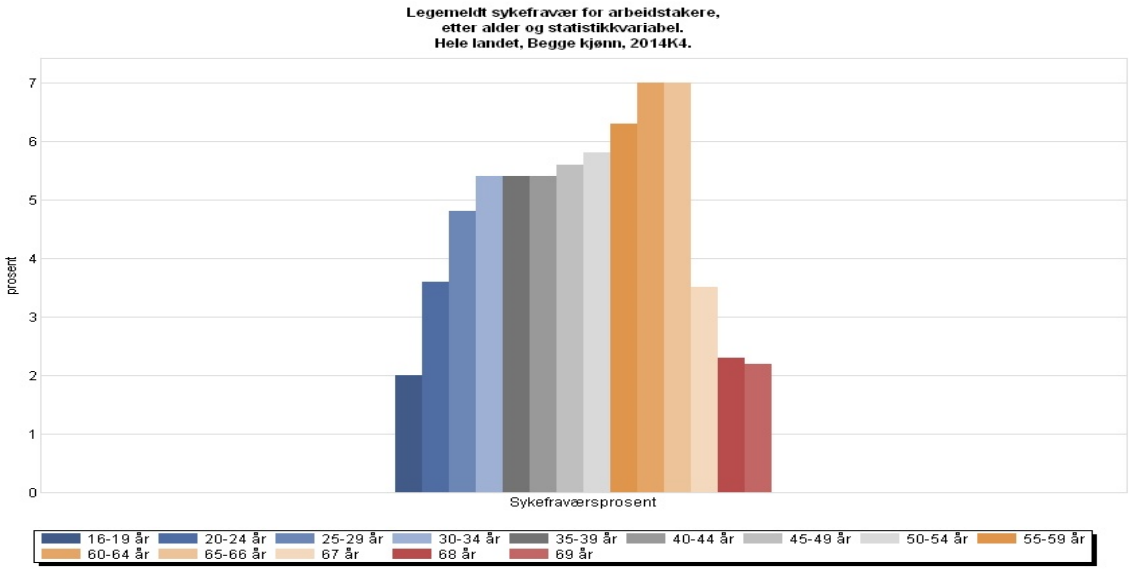
Legemeldt sykefravær utgjør hoveddelen av totalt sykefravær. Neste figur viser sykefraværet etter forskjellige yrkesgrupper.

Figur 1.1 Tapte arbeidsdager. Egenmeldt og legemeldt fravær etter standard for næringsgruppering. For eksempel, A= jordbruk, skogbruk og fiske, P=undervisning, Q=helse og sosialtjeneste



Neste figur viser legemeldt sykefravær etter alder.

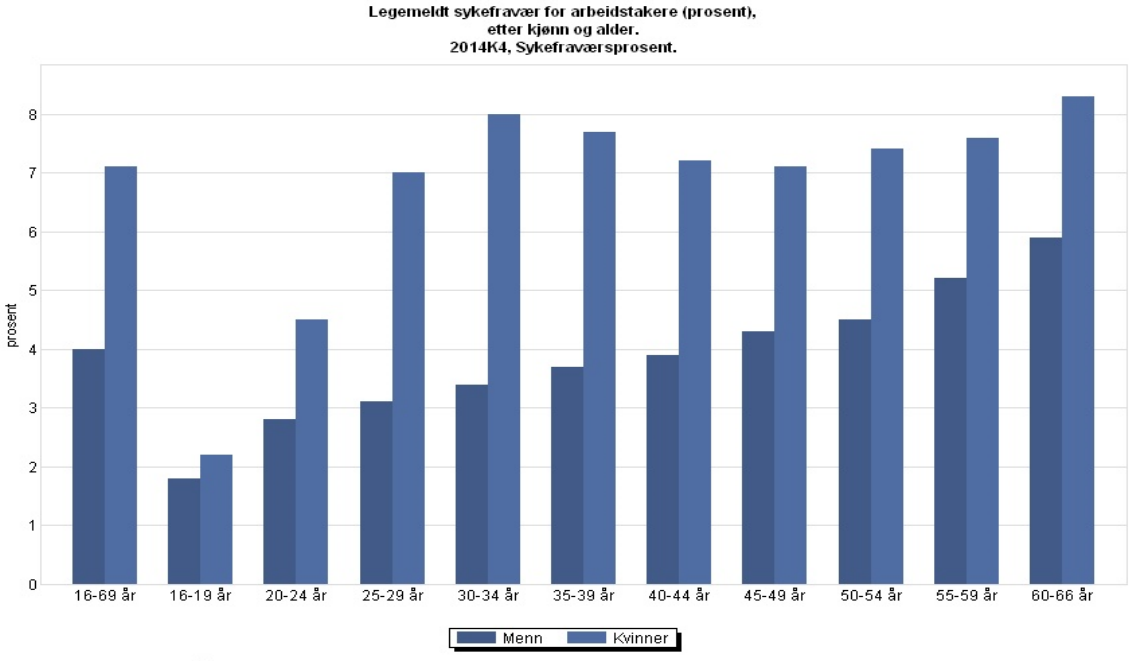
Figur 1.2 Tapte arbeidsdager. Legemeldt fravær etter alder



Kilde: Statistisk sentralbyrå

Neste figur viser legemeldt sykefravær etter kjønn og alder.

Figur 1.3 Tapte arbeidsdager. Legemeldt sykefravær etter kjønn og alder



Kilde: Statistisk sentralbyrå

Tabellen nedenfor viser hvordan sykefraværet fordeler seg på fylkene.

Tabell 1.2 Tapte arbeidsdager-fylkesvis

Fylke	Sykefravær	Fylke	Sykefravær
Østfold	6,6	Rogaland	4,5
Akershus	5,1	Hordaland	5,6
Oslo	4,6	Sogn og Fjordane	4,9
Hedmark	5,6	Møre og Romsdal	5,4
Oppland	5,9	Sør-Trøndelag	5,5
Buskerud	5,7	Nord-Trøndelag	6,2
Vestfold	5,4	Nordland	6,2
Telemark	5,7	Troms	6,3
Aust-Agder	5,8	Finmark	6,8
Vest-Agder	5,3		

Etter denne gjennomgangen vet vi følgende:

- De fleste sykefraværene er legemeldte, og kvinner har høyere sykefravær enn menn
- Sykefraværet varierer med næring (yrke). Høyest i helse og sosialtjenester
- Sykefraværet varierer med alder
  - Ungdom og de over 67 har lavest sykefravær
  - Sykefraværet er klart høyest fra 60 til 66 år
- Sykefraværet varierer fylkesvis
  - Lavest i Rogaland
  - Høyest i Nord-Norge og Østfold,

  Det er ikke tilfeldig, 95 % konfidensintervall for Finmark er 6,77-6,83.

2. Innføring i basisbegreper i utvalgsundersøkelser

I dette kapitlet behandles følgende tema:

- Populasjon, utvalg, utvalgsplan
- Estimering
- Feilkilder
- To SSB eksempler på utvalg og utvalgsplaner
  - Egenmeldt sykefravær og levekår Helse 2012
- Hvorfor utvalg, representativitet
- Estimering ved enkelt tilfeldig utvalg
  - estimator, estimat
  - forventning og forventningsretthet, mål for skjevhet
  - varians, standardfeil
  - konfidensintervall

2.1. Populasjon og utvalg

Populasjonen, også kalt målpopulasjon, er alle enhetene av interesse for en gitt *statistikk*. Den betegnes, med størrelse N, $U = \{1, 2, ..., N\}$. U for "univers". Alle enhetene kan identifiseres og merkes. Noen eksempler er

- politisk meningsmåling – alle voksne som har stemmerett
- arbeidsledighet i Norge – alle personer i Norge, 15 år og eldre
- forbruksundersøkelsen: enhet = husholdning

Utvalget betegnes med s (for engelsk "sample") og er, de enhetene som trekkes ut, den delen av populasjonen som skal "observeres". For eksempel, $s = \{3,17, 55,70\}$ hvis enhetene 3, 17, 55, 70 er trukket ut. Utvalget bør være "representativt" for populasjonen. Det betyr forskjellig for personutvalg og bedriftsutvalg.

Utvalgsplanen beskriver hvordan utvalget velges. Utvalget er et *sannsynlighetsutvalg* hvis alle enhetene i utvalget trekkes med visse sannsynligheter, og slik at hver enhet i populasjonen har en positiv sannsynlighet for å trekkes ut til utvalget. *Sannsynlighet* for en begivenhet er lik *andel* ganger begivenheten inntreffer hvis vi trekker utvalget «uendelig» mange ganger, dvs. sannsynlighet er langtidsfrekvenen for begivenheten. Vi skal kun betrakte sannsynlighetsutvalg. Den enkleste utvalgsplanen er:

- *Enkelt tilfeldig utvalg* (*ETU*). La n være utvalgsstørrelsen. Da har alle *utvalg* med n enheter samme sjanse å bli trukket ut. Det medfører at alle enhetene i populasjonen samme trekkesannsynlighet n/N.

Eksempel. Anta N=4 slik at $U = \{1, 2, 3, 4\}$. La $n = 2$. Da er det 6 mulige utvalg: {1,2}, {1,3}, {1,4}, {2,3}, {2,4} og {3,4}, som alle har samme sannsynlighet 1/6 for å bli trukket.

Noen grunner til å ta et utvalg fra populasjonen er:

- Et utvalg reduserer kostnader for akseptabelt nivå på nøyaktighet (penger, arbeidskraft, tid til bearbeidelse...)
- Kan samle inn mer informasjon for hver person i utvalget
- Vi får resultatene mye raskere.

Et naturlig kvalitetskrav til personundersøkelser er representativitet på viktige demografiske variable, for eksempel, balanse på kjønn og alder:

- Andel kvinner i utvalget er lik andelen i populasjonen
- Andeler i aldersgrupper i utvalget er lik andelene i populasjonen

Et *ideelt* representativt utvalg er en miniatyr versjon av populasjonen og impliserer at hver enhet i utvalget representerer egenskaper/trekk til et kjent antall enheter i populasjonen. Passende sannsynlighetsutvalg sikrer et representativt utvalg "gjennomsnittlig"

Det grunnleggende statistiske problem er *estimering* som gis en kort innledning i neste seksjon.

2.2. Estimering

En typisk undersøkelse har mange variabler av interesse. Formålet med en undersøkelse er vanligvis å få informasjon om totaler og gjennomsnitt for disse variablene for hele populasjonen. Et eksempel:

- Arbeidsledighet i Norge– Ønsker å estimere det totale antall arbeidsledige t.

For hver person i (minst 15 år gammel) i Norge så kan vi definere følgende binære variabel:
$y_i = 1$ hvis person i er arbeidsledig, og 0 ellers. Da er det totale antall arbeidsledige lik

$$t = y_1 + y_2 + ... + y_N = \sum_{i=1}^{N} y_i .$$

Generelt, variabel av interesse betegnes y med y_i lik verdien til y for enhet i i populasjonen, og totalen betegnes med

$$t = \sum_{i=1}^{N} y_i .$$

Det typiske problemet er å estimere t eller populasjonsgjennomsnittet t/N . Noen ganger er vi også interessert i å estimere forholdet mellom to totaler.

- Eksempel - *estimering av andel arbeidsledige.*
  I tillegg til y variabelen som indikerer om en person er arbeidsledig så trenger vi følgende variabel:
  $x_i = 1$ hvis person i er i arbeidsstyrken, og 0 ellers.
  Arbeidsstyrken = alle sysselsatte + arbeidssøkere (ledige). La totalene for de to variablene betegnes med t_y, t_x. Arbeidsledighetsandel blir t_y/t_x.

2.3. Feilkilder i utvalgsundersøkelser

Grovt sett kan vi dele opp feilkildene i fire grupper.

1. Målpopulasjon U mot Registerpopulasjon U_F

Tilgang til populasjonen er via en liste av enheter – et register U_F. U and U_F kan være forskjellige, tre mulige feil i U_F (spesielt i bedriftsundersøkelser) er:

- Underdekning: Noen enheter i U er ikke i U_F
- Overdekning: Noen enheter i U_F er ikke i U
- Dubletter: en enhet i U er listet mer enn en gang i U_F

U_F kalles av og til utvalgsrammen (sampling frame). I dette kompendiet så antas at $U = U_F$

2. *Frafall - manglende data*
   - Noen personer kan ikke bli kontaktet
   - Noen nekter å delta i undersøkelsen
   - Noen kan være syke og ute av stand til å svare
   - I postale surveys: Kan være så mye som 70 % frafall
   - I telefon surveys: 50 % frafall er ikke uvanlig
   - Mulige konsekvenser:
     - Utvalgsskjevhet, ikke lenger representativt for populasjonen.
     - Estimering blir mer unøyaktig

3. *Målefeil*– måler ikke korrekt verdi av y_i
   - Vanligst i bedriftsundersøkelser: f.eks.1000-feil (oppgir i gal måleenhet)
   - I intervju-undersøkelser:
     - Intervjuereffekt: folk kan si hva de tror intervjueren ønsker å høre- underrapportering av alkoholbruk, tobakkbruk
     - Misforstår spørsmålet, husker ikke riktig

(1) *Utvalgs«feil»* (Utvalgstoleranse)

- Feilen(usikkerhet, avvik) forårsaket av at vi observerer et utvalg og ikke hele populasjonen. Vi bruker begrepet utvalgsfeil fordi det er en vanlig betegnelse, om enn noe misvisende.
- For å anslå denne feilen måler vi feilmarginen: Den måler variasjonen fra utvalg til utvalg hvis vi trekker utvalget mange ganger. Ett slikt mål er 95 % konfidensintervall
- Sannsynlighetsutvalg medfører at vi kan estimere utvalgsfeil og beregne konfidensintervall.
- De første tre feilene kalles ikkesampling-feil, og kan være mer betydelige enn utvalgsfeilen
- I dette kompendiet behandles kun frafall av ikkesampling-feil.

2.4. To SSB eksempler på utvalg og utvalgsplaner

Vi skal se på Egenmeldt Sykefravær for 4. kvartal 2014 og Levekår Helse for 2012.

Utvalgsplanen og utvalg for egenmeldt sykefravær

- Et tilfeldig utvalg av10 000 bedrifter velges ut, stratifisert etter næring og størrelse
- Postal undersøkelse, spørreskjema sendes til bedriftene i utvalget siste uke i hvert kvartal
- Oppgaveplikt, svarprosent er over 90
- Alle bedrifter innenfor samme nærings-og størrelsestratum har samme sannsynlighet for å bli trukket ut
- Trekksannsynligheten øker med størrelsen
  - Ingen små bedrifter (3 eller færre ansatte) skal trekkes ut
  - Alle store bedrifter (flere enn ca. 150 ansatte) blir trukket ut
- Undersøkelsen dekker 36 prosent av ansatte ved å trekke ut 5,5 prosent av bedriftene
- Mer om utvalgsplaner for bedriftsundersøkelser i kapittel 6 (??)

Utvalgsplanen for Levekår Helse 2012

- Det ble trukket et tilfeldig utvalg på 10 000 personer i alderen 16 år og eldre, bosatt i private husholdninger
- Stratifisert (representativt) etter kjønn, aldersgrupper og landsdel
- Intervju-undersøkelse, telefon (99,5 %) og besøk

Tabell 2.1 Utvalget for Levekår Helse 2012

	Antall	Prosent
Utvalget	10 000	
Avgang (døde, bosatt i utlandet/institusjon)	229	
Bruttoutvalg	9 771	100
Frafall	4 111	42
Nettoutvalg (personer oppnådd intervju med)	5 660	58
Besøksandel	29	0,5
Intervjutid: 33 minutter		

2.5. Sannsynlighet – en kort innføring

Formålet med å samle inn data er å trekke konklusjoner om populasjonen som data er observert fra. Fundamentet for å kunne gjøre dette er sannsynlighetsteorien, som er en teori om *mekanismen* som genererer data. I (design-basert) utvalgsundersøkelser er det trekking av utvalget som genererer data.

Sannsynlighetspråket er det matematiske verktøy vi trenger for å utføre statistisk analyse på data, dvs., statistisk inferens.

Sannsynlighetsbegrepet er knyttet til sjansen for at en (usikker) begivenhet inntreffer, for eksempel at gjennomsnittlig personinntekt fra et enkelt tilfeldig utvalg er større enn en spesiell verdi. Hvis vi generelt lar A være en begivenhet, så er sannsynligheten for begivenheten A, betegnet med $P(A)$, definert som grenseverdien til andel (det relative antall) ganger A inntreffer ved gjentatte trekkinger av utvalget. Eller mer generelt, hvis A er en begivenhet som kan inntreffe i et stokastisk forsøk, dvs. vi kan ikke på forhånd si hva utfallet av forsøket blir, så er $P(A)$ det relative antall ganger A inntreffer i det lange løp ved gjentatte forsøk.

Som et enkelt eksempel, kan vi betrakte kast med mynt. Når vi sier at $P(\text{Kron}) = ½$, så mener vi at i gjentatte forsøk med å kaste mynten så vil Kron inntreffe i 50 % av kastene i det lange løp.

Et annet eksempel er kast med terning hvor vi noterer X = antall øyne. Hvis terningen er «rettferdig» så vil alle verdier 1-6 ha sannsynlighet 1/6, $P(X= x) =1/6$ for $x = 1,\ldots, 6$. Nå kan vi snakke om *forventningen* til X, som ikke er det samme som forventet verdi, men heller lik forventet verdi av *gjennomsnittet* av X ved gjentatte kast av terningen. Det betyr at forventningen til X er lik *summen av verdier·sannsynlighet,* fordi sannsynlighet for verdi x angir andel ganger x inntreffer i det lange løp. Presist kan det beskrives på følgende måte:

- Anta m gjentatte kast med terning med X-verdier: $x_1,\ldots, x_m$
- m_x = antall ganger verdien x inntreffer, $x = 1, \ldots, 6$
- $\sum_{i=1}^{m} x_i = x_1+ \ldots+x_m = 1\cdot m_1+2\cdot m_2+\ldots+6\cdot m_6= \sum_{x=1}^{6} x\cdot m_x$

- Gjennomsnittet: $\bar{x} = \frac{1}{m}\sum_{x=1}^{6} x \cdot m_x = \sum_{x=1}^{6} x \cdot \frac{m_x}{m} \to \sum_{x=1}^{6} x \cdot P(X = x)$ når $m \to \infty$

Forventningen betegnes med $E(X)$ (engelsk: expected value eller expectation). Vi har altså at

$$E(X) = 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + \ldots + 6 \cdot P(X = 6) = (1+2+..+6)/6 = 3{,}5.$$

2.6. To eksempler på bruk av sannsynlighet*

Spill med tre dører

Et TV-show dreier seg om å tippe bak hvilken av tre dører premien («bil») er. Hver gang en deltaker tipper en dør, la oss si nr.2, så vil TV-verten (som vet hvor bilen er) åpne en av dørene som gjenstår som ikke inneholder premien, f.eks. nr. 3. Deretter får deltakeren et valg mellom å beholde sitt første tips eller bytte til den gjenværende døren, her nr. 1.

Spørsmålet er: Bør deltakeren bytte dør, spiller det ingen rolle eller bør man ikke bytte dør?

Dette problemet skapte en stor debatt i amerikanske aviser på 90-tallet. Mange matematikere og statistikere tok feil!

La oss se på et *intuitivt svar*:
Anta spillet ble gjentatt 90 ganger. Hver gang ble bilen plassert tilfeldig bak en av dørene, slik at bilen er bak dør 1 30 ganger, bak dør 2 30 gange og bak dør 3 30 ganger. Anta det er to deltakere:

Deltaker A valgte dør 1 hver gang og beholdt dette valget etter at TV-verten har åpnet en dør.
Deltaker B valgte dør 1 først og byttet deretter til den døren som sto igjen etter åpningen av en dør. Dvs., hvis dør 2 åpnes så velger B dør 3, og hvis dør 3 åpnes velger B dør 2.

Vi ser da:
Deltaker A vinner 30 av 90 ganger: *A har 1/3 sjanse* for å vinne.
Deltaker B vinner hver gang A taper, dvs. 60 av 90 ganger (hver gang bilen er bak dør 2 eller dør 3):
B har 2/3 sjanse for å vinne.

Konklusjon: Det lønner seg å bytte dør. Sjansen til å vinne blir dobbelt så stor!

Du kan sjekke dette ved å spille et lignende spill for to personer med tre kort. La ess = «bil» og velg to vilkårlige kort (ikke ess), f.eks. to jokere. Den ene personen spiller, mens den andre er «TV-vert». Ved hvert spill legges kortene i tilfeldig rekkefølge, og spilleren velger et av kortene. «TV-verten» snur et kort som ikke er ess, og spilleren bytter deretter til det kortet som ikke først ble valgt. Spill dette et par hundre ganger og se hva som skjer. Et matematisk bevis er gitt i Appendiks A.

Finaler i fotball

Finalen i EM 2016 var Frankrike mot Portugal (som Portugal vant 1 – 0). Her ble Frankrike regnet som ganske stor favoritt. Mye av interessen i en slik finale er at i *en* enkel kamp er det absolutt ikke sikkert at det beste laget vinner. Alt kan skje, slik at det svakeste laget kan vinne.

La oss nå prøve å lage et opplegg som sikrer at det beste laget blir europamester. Det betyr at EM-finalen må bestå av flere kamper, og Europamester blir laget som vinner majoriteten av disse kampene. (Hver kamp avgjøres med, om nødvendig, ekstraomganger og straffe). For eksempel hvis det bestemmes at 5 kamper skal spilles så må Frankrike vinne tre av disse for å bli europamester.

Det skal spilles så mange kamper at vi er "95 % sikker" på at det beste laget blir EM-mester, dvs. sannsynligheten for at det beste laget vinner skal være 0,95.

Spørsmålet er nå: Hvor mange kamper må spilles?
Svar: Det avhenger av styrkeforholdet mellom lagene.

La **n** betegne antall kamper som må spilles. La oss si at Frankrike ville slått Portugal i 60 % av gjentatte møter. Dvs., Frankrike ville vunnet 6 av 10 kamper mot Portugal. Hva er nå **n**? Svaret er 71**.** Hvis Frankrike antas å slå Portugal i 3 av 4 kamper så er **n** = 11**.** En utledning er gitt i Appendiks A. Fra tabell A2 har vi:

Frankrike styrke	**n**
55 %	279
60 %	71

65 %	31
70 %	17
75 %	11
80 %	7

For eksempel, hvis Frankrike og Portugal er ganske jevnbyrdige, la oss si at Frankrike antas å vinne 55 % av kampene så er **n** = 279.

Et alternativt opplegg**:** Tillat uavgjorte kamper. Må da vinne majoriteten av de kampene som ikke ender uavgjort. Det er også da mulig å beregne **n**. To eksempler:

- Hvis fordelingen av seire, uavgjort og tap for Frankrike antas å være: 60, 20, 20 prosent: **n** = 14**.**
- Hvis fordelingen av seire, uavgjort og tap for Frankrike antas å være: 70, 10, 20 prosent: **n** = 9.

2.7. Estimeringsteori – enkelt tilfeldig utvalg (ETU)

Som nevnt i Seksjon 2.1, hvert utvalg *s* med størrelse *n* har samme sannsynlighet for å bli trukket. I prinsippet kan trekkingen utføres ved å trekke en og en enhet tilfeldig uten tilbakelegging . La oss se på estimering av populasjonsmiddelverdien av variabelen *y*:

$$\mu = \sum_{i=1}^{N} y_i / N.$$

En naturlig *estimator* er gjennomsnittet i utvalget:

$$\bar{y}_s = (\text{summen av alle } y - \text{verdiene i utvalget } s) / n = \sum_{i \in s} y_i / n.$$

Estimatet er den beregnede verdien av estimatoren når utvalget er observert, slik at en estimator er *selve* funksjonen av data. For å beskrive egenskaper til en estimator trenger vi å beregne forventningen:

> *Forventningen til en estimator er den gjennomsnittlige verdien av estimatoren ved (uendelig mange) gjentatte observasjoner av estimatoren.*

Forventningen betegnes med $E(\hat{\mu})$ for en estimator $\hat{\mu}$. Denne fortolkningen av forventningsbegrepet illustreres i øvelse «2». Formelt er forventningen lik summen av estimatverdi·sannsynlighet (for verdien). For eksempel, hvis $\hat{\mu}$ kan ta verdiene 1, 2, 3 med sannsynlighetene 0,5, 0,3 og 0,1 henholdsvis, så er $E(\hat{\mu})$ =1·0,5 + 2·0,3 + 3·0,1 = 1,4.

En estimator er forventningsrett (engelsk: unbiased) hvis $E(\hat{\mu}) = \mu.$ Skjevheten (engelsk: bias) til en estimator er $E(\hat{\mu}) - \mu.$ Det kan vises at $\bar{y}_s$ er forventningsrett for μ i ETU design. Det betyr altså at hvis utvalgstrekkingen gjentas mange ganger (på samme tid, dvs. hypotetiske gjentakelser) så vil gjennomsnittsverdien av estimatoren bli lik μ.

Usikkerheten til en forventningsrett estimator måles med den *estimerte* utvalgsvariansen eller den estimerte standardfeilen (*SE* for engelsk: standard error).

$$Var(\hat{\mu}) = E(\hat{\mu} - \mu)^2, \text{ hvis } E(\hat{\mu}) = \mu.$$

Dvs., $Var(\hat{\mu})$ er gjennomsnittlig verdi ved hypotetiske gjentakelser av $(\hat{\mu} - \mu)^2$. Standardfeilen er da $\sqrt{Var(\hat{\mu})}$.

Hvis $\hat{\mu}$ ikke er forventningsrett så defineres variansen til å være $Var(\hat{\mu}) = E(\hat{\mu} - E(\hat{\mu}))^2$. La $\hat{V}(\hat{\mu})$ være et (helst forventningsrett) estimat av $Var(\hat{\mu})$. Den estimerte standardfeilen er da $SE(\hat{\mu}) = \sqrt{\hat{V}(\hat{\mu})}$. For enkelthets skyld bruker vi kun betegnelsen *standardfeil* for den *estimerte* standardfeilen. Noen resultater for enkelt tilfeldig utvalg:

(1) La π_i være sannsynligheten for at enhet *i* er i utvalget, trekkesannsynligheten. Da er $\pi_i = n/N$, utvalgsandelen.

(2) $E(\bar{y}_s) = \mu$.

(3) La σ^2 være populasjonsvariansen, $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$.

Her er $\sum_{i=1}^{N}(y_i - \mu)^2$ er summen av alle $(y_i - \mu)^2$ i populasjonen.

Da er $Var(\bar{y}_s) = \frac{\sigma^2}{n}(1-\frac{n}{N})$. Faktoren (1- *n*/*N*) kalles endelig populasjonskorreksjon.

σ^2 er et mål på hvor stor variasjon det er i populasjonen, dvs. hvor mye y_i varierer mellom enhetene, f.eks., hvor forskjellig egenmeldt sykefravær er for de forskjellige bedriftene.

Populasjonsvariansen σ^2 estimeres ved utvalgsvariansen

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y}_s)^2.$$

Hvor $\sum_{i\in s}(y_i - \bar{y}_s)^2$ summerer, for alle enheter i utvalget *s*, $(y_i - \bar{y}_s)^2$.

Estimert varians: $\hat{V}(\bar{y}_s) = \frac{\hat{\sigma}^2}{n}\left(1-\frac{n}{N}\right)$.

Vanligvis rapporterer vi standardfeilen til estimatet: $SE(\bar{y}_s) = \sqrt{\hat{V}(\bar{y}_s)}$.

For å illustrere beregningen av variansestimat og standardfeil, anta *y* er egenmeldt sykefravær i prosent og at populasjonen består av 10 bedrifter. Utvalget er på 4 bedrifter med *s* = (1,4,7,8) med *y*-verdiene 4,0 – 6,7 – 9,0 – 3,5. Da er $\bar{y}_s = (y_1 + y_4 + y_7 + y_8)/4 = 23{,}2/4 = 5{,}8$ og

$\sum_{i\in s}(y_i - \bar{y}_s)^2 = (y_1 - \bar{y}_s)^2 + (y_4 - \bar{y}_s)^2 + (y_7 - \bar{y}_s)^2 + (y_8 - \bar{y}_s)^2 = 1{,}8^2 + 0{,}9^2 + 3{,}2^2 + 2{,}3^2 = 19{,}58.$

Det gir at $\hat{\sigma}^2 = 19{,}58/3 = 6{,}527$ og estimert varians blir $\hat{V}(\bar{y}_s) = 6{,}527\cdot(1-4/10)/4 = 0{,}979$ og

$SE(\bar{y}_s) = \sqrt{0{,}979} = 0{,}989.$

Feilmarginen er definert som $2\cdot SE(\bar{y}_s)$, som kan forklares ved begrepet "konfidensintervall":

- Et konfidensintervall er et intervall som med stor sikkerhet inneholder den størrelsen vi ønsker å estimere.
- Det mest vanlige er å beregne et 95 % konfidensintervall: Da er vi 95 % "sikker" på at intervallet inkluderer den sanne verdien.
- Konkret tolkning av begrepet "sikker": Hvis vi trekker utvalget 100 ganger så vil det beregnede intervallet inneholde den sanne verdien 95 ganger.

Konfidensintervallet for μ ved ETU er basert på sentralgrenseteoremet: For store *n*, *N* – *n* så er $\bar{y}_s$ (tilnærmet) normalfordelt. Av dette får vi at 95 % konfidensintervall for μ er gitt ved:

$$\bar{y}_s - 1{,}96\cdot SE(\bar{y}_s), \bar{y}_s + 1{,}96\cdot SE(\bar{y}_s) = \bar{y}_s \pm 1{,}96\cdot SE(\bar{y}_s).$$

Derav ser vi hvorfor $2\cdot SE(\bar{y}_s)$ betegnes som feilmarginen.

2.8. Eksempel – Kvalitetsindeks i California skoler

- Academic Performance Index (API) for alle California skoler
- Basert på standardisert testing av elevene
- Data fra alle skoler med minst100 elever
- Enhet i populasjon = skole (Grunnskole/Ungdomsskole/Videregående)
- Populasjonen består av *N* = 6194 observasjoner
- Ser på variabelen: *y* = api00 = API i 2000
- Middel(*y*) = 664.7 med min(*y*) =346 og max(*y*) =969
- Datasett i R: apipop og y= apipop$api00

For ett utvalg av størrelse *n* = 100 fikk vi følgende resultater: $\bar{y}_s = 654{,}5$ og $SE(\bar{y}_s) = 12{,}6$. Et tilnærmet 95 % konfidensintervall blir: $654{,}5 \pm 1{,}96\cdot 12{,}6 = 654{,}5 \pm 24{,}7 = (629{,}8 - 679{,}2)$.

R-kode (engelsk tegnsetting) som ble brukt:

```
s=sample(1:6194,100)
ybar=mean(y[s])
se=sqrt(var(y[s])*(6194-100)/(6194*100))
ybar
[1] 654.47
var(y[s])
[1] 16179.28
se
[1] 12.61668
```

Her er var(y[s])= $\hat{\sigma}^2$.
Verdien av utvalgsfeilen er lite informativ hvis den ikke er relatert til selve estimatet. For eksempel, *SE* = 2 er liten hvis estimatet er 1000, men meget stor hvis estimatet er 3. Variasjonskoeffisienten for estimatet er et mål på den relative variasjonen til estimatet og er definert ved:

$$CV(\bar{y}_s) = SE(\bar{y}_s) / \bar{y}_s.$$

I dette eksemplet så er $CV(\bar{y}_s) = 12{,}6 / 654{,}5 = 0{,}019 = 1{,}9\%$.
CV er uavhengig av måleenhet og mer stabil over gjentatte undersøkelser. *CV* kan brukes planlegging, for eksempel til å bestemme utvalgsstørrelsen. Den er spesielt meningsfull ved estimering av andeler.

Vi gjentok trekkingen 10 ganger til. Resultatene er vist i tabell 2.2.

Tabell 2.2 Ti konfidensintervall fra ti enkle tilfeldige utvalg på n = 100

95 % konfidensintervall	Inkluderer sann verdi 664,7
1. 644,9 – 692,2	ja
2. 668,9 – 716,2	nei
3. 616,7 – 670,3	ja
4. 671,1 – 721,9	nei
5. 650,2 – 702,9	ja
6. 623,3 – 667,2	ja
7. 651,3 – 699,0	ja
8. 629,0 – 675,5	ja
9. 615,6 – 669,8	ja
10. 631,2 - 680,6	ja

2.9. Estimering av populasjonsandel *p* med en viss egenskap/kjennemerke A

La *p* = (*antall enheter i populasjonen med A*)/*N*. Definer variabelen *y* ved $y_i = 1$ hvis enheten *i* har kjennemerke A, 0 ellers. Da er *p* populasjonsgjennomsnittet av y_i'ene. La *X* være antall enheter i utvalget med kjennemerke A. Da kan utvalgsgjennomsnittet uttrykkes som

$$\hat{p} = \bar{y}_s = X / n.$$

Med enkelt tilfeldig utvalg så har vi at $E(\hat{p}) = p$, og estimatet av variansen til estimatoren blir

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}(1-\frac{n}{N}).$$

Eksempel: I en politisk meningsmåling med et tilfeldig utvalg på 1000 stemme-berettigede personer i Norge, sier 280 de vil stemme på AP. Den estimerte andel av AP stemmer i Norge er gitt ved:

$$\hat{p} = 280/1000 = 0{,}28.$$

Standardfeilen er $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}(1-\frac{n}{N})} = \sqrt{\frac{0{,}28 \cdot 0{,}72}{999}} = 0{,}0142$, og 95 % konfidensintervall:

$$\hat{p} \pm 1{,}96 \cdot SE(\hat{p}) = 0{,}280 \pm 0{,}028 = (0{,}252 - 0{,}308).$$

3. Estimeringsmetoder i utvalgsundersøkelser

Dette kapitlet omfatter følgende:

- Planlegging av utvalgsstørrelse
- Basisestimator 1 for utvalgsundersøkelser: Rate-estimatoren
- Basisestimator 2 for utvalgsundersøkelser. Horvitz-Thompson estimatoren
- Modifisert Horvitz-Thompson estimator
- Ikke-eksistens av optimale estimatorer

3.1. Bestemme utvalgsstørrelse basert på konfidensintervall, for populasjonsandel

Utvalgsstørrelsen har avgjørende effekt på undersøkelsens kostnad og tidsbruk. Hvor stor n bør være avhenger av formålet med undersøkelsen. I en meningsmåling for å estimere partipreferanse så er n = 1000 typisk nok. I kvartalsvis AKU så er n = 24000, spesielt på grunn av ønsket pålitelighet for endringstall.

Det er hovedsakelig tre faktorer som bør betraktes/vurderes:

1. Ønsket nøyaktighet på estimater for mange variabler. Fokusere på en eller to variabler av primær interesse
2. Homogenitet i populasjonen. Behøver mindre utvalg hvis liten variasjon i populasjonen
3. Estimering for delgrupper, «domener», i populasjonen.

Det er ofte faktor 3 som setter det høyeste kravet på undersøkelsen. Det bør da tas et *stratifisert utvalg*, et utvalg fra hvert domene (stratum).
Anta problemet er å estimere en populasjonsandel p for et visst stratum, og vi bruker utvalgsandelen fra stratumet til å estimere p. La n være utvalgsstørrelsen for dette stratumet, og anta at n/N er ubetydelig. La oss si at ønsket nøyaktighet for dette stratumet er at 95 % KI for p skal være ±5 %. Vi har da

$$\text{tilnærmet 95 \% KI for } p: \hat{p} \pm 1{,}96\sqrt{\hat{p}(1-\hat{p})/n}$$

slik at nøyaktighetskrav blir nå:

$$1{,}96\sqrt{\hat{p}(1-\hat{p})/n} = 0{,}05 = 1/20 \Rightarrow n = 1{,}96^2 \cdot 20^2\, \hat{p}(1-\hat{p}) \le 3{,}84 \cdot 400 \cdot 0{,}5 \cdot 0{,}5 = 384. \qquad (1)$$

Dette kommer av at $\hat{p}(1-\hat{p}) \le 0{,}5 \cdot 0{,}5$ for alle verdier av $\hat{p}$.
Estimatet er ukjent i planleggingsfasen. Vi kan bruke den konservative størrelsen 384 eller en planningsverdi p_0 med n = 1536 p_0(1- p_0). For eksempel, med p_0 = 0,2 så blir n = 246.

3.2. Bestemme utvalgsstørrelse basert på variasjonskoeffisienten, for populasjonsandel*

Et alternativt mål på nøyaktighet er å bruke variasjonskoeffisienten CV, $CV(\hat{p}) = c$. Det betyr at

$$SE(\hat{p})/\hat{p} = c \Leftrightarrow n \approx \frac{1}{c^2} \cdot \frac{1-\hat{p}}{\hat{p}}.$$

Det følger av at $SE(\hat{p})/\hat{p} \approx \frac{1}{\sqrt{n}}\sqrt{(1-\hat{p})/\hat{p}}$ slik at $SE(\hat{p})/\hat{p} = c \Leftrightarrow \sqrt{n} \approx \frac{1}{c} \cdot \sqrt{(1-\hat{p})/\hat{p}}$.

Med planleggingsverdi p_0: $n = \frac{1}{c^2} \cdot \frac{1-p_0}{p_0}$.

For en gitt planleggingsverdi p_0 og $CV = c$, så er $SE = c \cdot p_0$. Med c = 0,1 så blir utvalgsstørrelsen og tilhørende konfidensintervall:

Med p_0 = 0,5: n = 100 og tilnærmet 95 % konfidensintervall = $\hat{p} \pm 2 \cdot SE(\hat{p}) = \hat{p} \pm 2 \cdot 0{,}1 \cdot p_0 = \hat{p} \pm 0{,}10$
Med p_0 = 0,1: n = 900 og tilnærmet 95 % konfidensintervall = $\hat{p} \pm 2 \cdot SE(\hat{p}) = \hat{p} \pm 2 \cdot 0{,}1 \cdot p_0 = \hat{p} \pm 0{,}02$

Eksempel: Månedlig arbeidsledighet
Det er viktig å oppdage endringer i arbeidsledighet fra måned til måned, La oss bruke en planleggings-verdi p_0 = 0,05. La d være ønsket nøyaktighet på feilmarginen. Da har vi, fra (1):

$$1{,}96 \cdot SE(\hat{p}) = d \Leftrightarrow 1{,}96\sqrt{p_0(1-p_0)/n} = d \Leftrightarrow 1{,}96^2\, p_0(1-p_0)/n = d^2$$

$$\Rightarrow n = 3{,}84 \cdot p_0(1-p_0)/d^2 = 0{,}1824/d^2$$

Noen utvalgte verdier av *d*:

d = 0,001 (feilmargin = 0,1 %) gir *n* = 182400
d = 0,002 : *n* = 45600
d = 0,005: *n* = 7300

Merk at $d = 0{,}005 \Leftrightarrow SE(\hat{p}) = d/1{,}96 = 0{,}00255$ og $CV(\hat{p}) = 0{,}00255/0{,}05 = 0{,}051 = 5{,}1\%$.

3.3. Bestemme utvalgsstørrelse basert på variasjonskoeffisienten, generelt

Generelt, hvis vi skal estimere et populasjonsmiddel μ, så vil *n* avhenge av hvor stor *y*-variasjonen σ er i populasjonen. Hvis vi bruker utvalgsmiddel som estimat så er variasjonskoeffisienten i populasjonen, $CV = \frac{\sigma/\sqrt{n}}{\mu}$, (*N* er så stor at endelig populasjonskorreksjon kan neglisjeres). Med *CV* som mål på nøyaktighet så blir

$$\sqrt{n} = \frac{\sigma/\mu}{CV}$$

Tabellen viser hvordan *n* varierer med σ/μ for et gitt krav *CV*.

Tabell 3.1 Utvalgsstørrelse som funksjon av CV og populasjonens relative variasjon

CV	σ/μ		
	0,1	0,25	0,5
0,025	16	100	400
0,05	4	25	100
0,10	1	7	25

I de neste kapitlene 3.2 og 3.3 skal vi se på to basis estimatorer i utvalgsundersøkelser, rate-estimatoren og Horvitz-Thompson estimatoren. Vi skal betrakte rate-estimatoren for enkelt tilfeldig utvalg. Horvitz-Thompson estimatoren er utviklet for generelle sannsynlighetsutvalg hvor trekksannsynlighetene kan være ulike.

3.4. Rate-estimatoren

Anta vi har kjent tilleggsinformasjon for hele populasjonen, $\mathbf{x} = (x_1, x_2, ..., x_N)$. La $X_\circ = \sum_{i=1}^{N} x_i$. For eksempel, **x** kan være:

- I personundersøkelser: alder, kjønn, inntekt
- I økonomiske undersøkelser: omsetning, antall ansatte til en bedrift/virksomhet

Rate-estimatoren er vanligst for bedriftsundersøkelser. Hvis målet er å estimere en populasjonstotal *t* for en variabel *y,* så er rate-estimatoren er definert ved

$$\hat{t}_R = X_\circ \frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i} = X_\circ \frac{\bar{y}_s}{\bar{x}_s}.$$

Vi kan uttrykke rate-estimatoren på følgende form:

$$\hat{t}_R = \frac{X_\circ}{N\bar{x}_s}(N\bar{y}_s).$$

Den vanlige estimatoren, $\hat{t}_e = N\bar{y}_s$, kalles ekspansjonsestimatoren. Vi ser at rate-estimatoren justerer ekspansjonsestimatoren i de tilfeller hvor *x*-verdiene i utvalget er for små eller for store. Dette er rimelig hvis det er en positiv korrelasjon mellom *x* og *y*.

En modellbegrunnelse for rate-estimatoren
Hvis det er en proporsjonal sammenheng mellom *x* og *y*, for eksempel forbruk i forhold til inntekt, så kan vi uttrykke det på følgende måte:

$$y_i \approx \beta x_i \Rightarrow \sum_{i=1}^{N} y_i \approx \beta \sum_{i=1}^{N} x_i$$

$$\text{dvs.}, t \approx \beta \cdot X_\circ \text{ og } \beta \approx \sum_{i=1}^{N} y_i / \sum_{i=1}^{N} x_i = R.$$

Hvis *R* hadde vært kjent så kunne vi estimert *t* med $R \cdot X_{\circ}$. I enkelt tilfeldig utvalg kan vi bruke rateforholdet i utvalget til å estimere *R*:

$$\hat{R} = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} = \frac{\bar{y}_s}{\bar{x}_s} \text{ og } \hat{t} = \hat{R} \cdot X_{\circ} = \hat{t}_R.$$

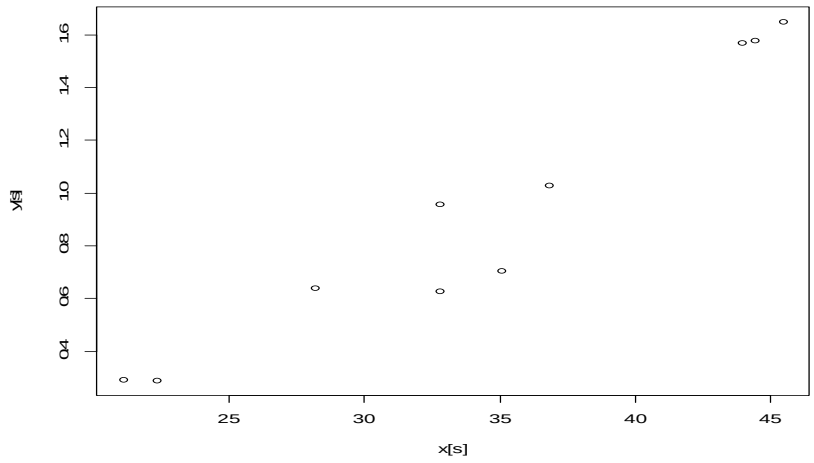
Eksempel*:* datasettet «trees» i R.
Populasjonen består av 31 trær (sorte kirsebær trær), og det er foretatt målinger av:
diameter (cm), høyde (m) og volum (m^3).
Det er vanskelig å måle volum så vi skal estimere totalt volum for de 31 trær ved å trekke et tilfeldig utvalg på 10 trær. Dvs., vi skal estimere

$$t = \sum_{i=1}^{31} y_i$$

hvor y_i er volum til tre *i*.

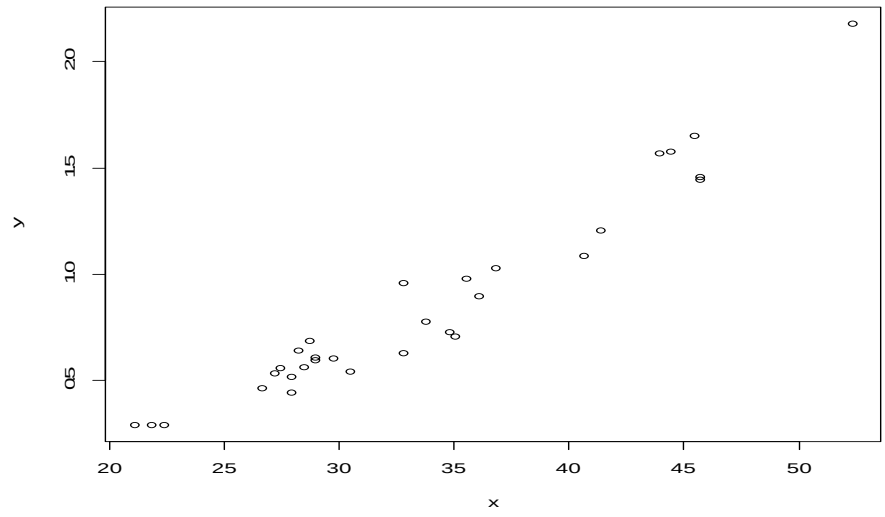
ET enkelt tilfeldig utvalg på 10 trær ga følgende observasjoner som vist i spredningsplottet nedenfor.

Figur 3.1 Spredningsplott for diameter mot volum for et enkelt tilfeldig utvalg på 10 trær



Tilsvarende har vi spredningsplott for hele populasjonen.

Figur 3.2 Spredningsplott for diameter mot volum for alle 31 trær



Siden vi kjenner volumet til alle trær i populasjonen, så kan den sanne verdien beregnes, og verdien er $t = 26,48$**.** Rateestimatet blir:

$$\hat{t}_R = 28,40.$$

Hvis vi ikke hadde tilleggsinformasjonen om diameter så hadde vi estimert med

$$\hat{t}_e = N \cdot \bar{y}_s = 28,91.$$

For å illustrere variasjonen i estimatene har vi trukket 5 utvalg til. Resultatene er gitt i tabell 3.2.

Tabell 3.2 Resultater fra fem enkle tilfeldige utvalg. Sann t = 26,48

Utvalg nr	Rate-estimat	Ekspansjons-estimat
2	25,01	24,90
3	20,77	18,25
4	25,52	26,31
5	22,04	20,43
6	30,41	34,37

Merk at med utvalg 4 får vi et bedre estimat ved kun å bruke gjennomsnittet.
Noen egenskaper for rate-estimatoren:

$E(\hat{t}_R) \approx t$, for store n, tilnærmet forventningsrett.

Variansen er gitt ved

$$Var(\hat{t}_R) \approx N^2 \cdot (1 - \frac{n}{N}) \cdot \frac{1}{n} \cdot \frac{1}{N-1} \sum_{i=1}^{N} (y_i - Rx_i)^2.$$

Vi merker oss følgende viktige egenskaper:

- Rate-estimatoren er meget presis når populasjons-punktene (y_i, x_i) ligger nær en rett linje gjennom origo.
- Rate-estimatoren er mer nøyaktig enn ekspansjonsestimatoren hvis Rx_i predikerer y_i bedre enn hva μ_y gjør:

$$Var(\hat{t}_R) < Var(N\bar{y}_s) \Leftrightarrow \sum_{i=1}^{N} (y_i - Rx_i)^2 < \sum_{i=1}^{N} (y_i - \mu_y)^2.$$

- I økonomiske bedriftsundersøkelser er det ganske vanlig å bruke en rate-estimator, med omsetning eller antall ansatte som tilleggsvariabel.

Estimert varians for rate-estimatoren:

$\sum_{i=1}^{N} (y_i - Rx_i)^2 / (N-1)$ er estimert ved $\sum_{i \in s} (y_i - \hat{R}x_i)^2 / (n-1)$.

Variansestimatet blir:

$$\hat{V}(\hat{t}_R) = \left(\frac{\mu_x}{\bar{x}_s} \right)^2 \cdot N^2 \cdot \frac{1-f}{n} \cdot \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{R}x_i)^2.$$

Merk at hvis $\bar{x}_s$ er veldig liten, da er $\hat{R}$ mer usikker og variansestimatet blir større for å gjenspeile det.

Standardfeilen til rateestimatoren: $SE(\hat{t}_R) = \sqrt{\hat{V}(\hat{t}_R)}$ og 95 % konfidensintervall er gitt ved

$$\hat{t}_R \pm 1,96 \cdot SE(\hat{t}_R).$$

I dette sjette utvalget med rateestimat lik 30,41 så er $SE(\hat{t}_R) = 2,14$ og konfidensintervallet blir (26,22 – 34,60). Ekspansjonsestimatet har standardfeil lik 4,27.

3.5. Horvitz-Thompson estimator – ulike trekkesannsynligheter

Vi ser på (lineære) estimater på formen $\hat{t} = \sum_{i \in s} w_i y_i$ hvor w_i ikke avhenger av s. Det kan vises at $\hat{t}$ er forventningsrett for alle verdier av y_i hvis og bare hvis $w_i = 1/\pi_i$.

$$\hat{t}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}.$$

Den eneste vekten som gir forventningsretthet er $1/\pi_i$, den inverse trekksannsynligheten. I enkelt tilfelidig utvalg så er $\pi_i = n/N$ og dermed $\hat{t}_{HT} = \sum_{i \in s} \frac{N}{n} y_i = N\overline{y}_s$.

Horvitz-Thompson estimatoren er en vanlig brukt estimator i offisiell statistikk. Variansen er liten hvis trekkesannsynlighetene bestemmes slik at y_i/π_i er tilnærmet like, dvs., π_i øker med økende y_i. Vi kjenner selvsagt ikke verdien til y_i når vi planlegger en survey, så vi bruker isteden kjent tilleggs-informasjon x_i og velger

$$\pi_i \propto x_i \Rightarrow \pi_i = nx_i / X$$

siden summen av alle π_i er lik n. I øvelse 2 illustreres det som er hovedproblemet med en generell anvendelse av Horvitz-Thompson estimatoren, nemlig at variansen kan bli så stor at estimatoren blir uinteressant og kan ikke brukes.

3.6. En modifisert H-T estimator*

Betrakt først estimering av populasjonsmiddel $\overline{y} = t/N$. Et opplagt valg av estimator er $\hat{\overline{y}}_{HT} = \hat{t}_{HT} / N$.
Alternativt s å kan vi også estimere N, uansett om N er kjent eller ikke.

$$\hat{N} = \sum_{i \in s} \frac{1}{\pi_i} \text{ (her er } y_i = 1 \text{ for alle } i)$$

For enkelt tilfeldig utvalg, $\pi_i = n/N \Rightarrow \hat{N} = \sum_{i \in s} \frac{N}{n} = N$.

Den modifiserte HT-estimatoren er da

$$\hat{\overline{y}}_w = \hat{t}_{HT} / \hat{N} = \frac{\sum_{i \in s} y_i / \pi_i}{\sum_{i \in s} 1/\pi_i} \Rightarrow \hat{t}_w = N\hat{\overline{y}}_w$$

$\hat{t}_w$ er ofte bedre enn $\hat{t}_{HT}$, og tilnærmet forventningsrett. Den har vanligvis mindre varians. Så $\hat{t}_w$ er vanligvis estimatoren som bør brukes, uansett om N er kjent eller ikke. Vi ser at den er en type "rate-estimator". Hvis utvalgsstørrelsen varierer så vil "rate-estimatoren" fungere bedre enn H-T estimatoren, raten er mer stabil enn telleren.

Illustrasjon
$y_i = c$, for $i = 1,\ldots, N$. Utvalgsplan er Bernoulli sampling; hver enhet i populasjonen velges med sannsynlighet π, en etter en. Da er utvalgsstørrelsen n en stokastisk variabel og har en binomisk (N,π) fordeling med $E(n) = N\pi$. De to estimatorene blir nå:

$$\hat{t}_{HT} = \frac{n}{\pi} \cdot c$$

$$\hat{t}_w = N \frac{nc/\pi}{n/\pi} = Nc = t$$

H-T estimatoren varierer siden n varierer, mens den modifiserte H-T er perfekt stabil.

3.7. Ikke-eksistens av optimale estimatorer

I vanlige statistiske modeller så finnes det optimale estimatorer, forventningsrette med minst varians blant alle forventningsrette estimatorer. Det gjelder for eksempel i lineær regresjonsanalyse. Der er de estimerte regresjonskoeffisientene optimale i denne forstand, blant alle lineære estimatorer hvis det ikke antas noen fordeling på residualene, og blant alle estimatorer hvis residualene antas normalfordelte.

En særegenhet ved design-basert inferens i utvalgsundersøkelser er at det ikke finnes slike «beste» estimatorer. «Vanlige» grunnleggende estimatorer har ikke samme egenskaper i design-basert utvalgsteori som de har i vanlige statistiske modeller.

Vi har faktisk et mye sterkere resultat (som også medfører at uansett hvor liten populasjon og utvalg er, så nytter det ikke å lete etter en forventingsrett estimator med minst varians):

Teorem: Anta en hvilken som helst utvalgsplan. Anta hver y_i kan ha minst to verdier. Da eksisterer det ingen uniformt best (minimum varians) design-forventningsrett estimator for totalen t.

4. Stratifisering og flertrinnsutvalg

Dette kapitlet tar oppfølgende temaer:

- Utvalgsplan med begrunnelse
- Estimering av populasjonstotaler og andeler i stratifiserte utvalg
- Fordeling av utvalg mellom strataene. Proporsjonal og optimal allokering
- Andre utvalgsplaner
  - Klyngeutvalg
  - 2-trinnsutvalg

4.1. Stratifiserte utvalgsplaner

Gunnleggende idé er å dele opp populasjonen U i H delpopulasjoner , kalt strata. Størrelsen på stratum h betegnes med N_h og antas kjent. Fra hvert stratum trekkes et separat utvalg s_h av størrelse n_h, uavhengig mellom strata. Stratifiserte utvalgsplaner krever at man har tilgang på god register-informasjon. I person-undersøkelser er det vanlig å stratifisere etter geografiske regioner, aldersgrupper, kjønn, mens i bedriftsundersøkelser så er det vanlig å stratifisere ved å bruke næring og antall sysselsatte som "stratifiseringsvariable".

For eksempel, i SSBs Levekår Helse 2012 er det stratifisert etter

- kjønn
- 5 aldersgrupper
- 7 landsdeler

slik at det totale antall strata = 2x5x7 = 70.

Noen begrunnelser for stratifisering er:

1. at strata danner domener av interesse hvor separate estimater av gitt presisjon er ønsket. For eksempel, strata = geografiske regioner.
2. å "spre" utvalget over hele populasjonen. Det blir lettere å få et representativt utvalg.
3. å få mer nøyaktige estimater av populasjonstotaler, dvs., redusere utvalgsvarians.
4. at det kan brukes forskjellige datainnsamlingsmetoder i forskjellige strata, for eksempel telefon i noen strata og besøksintervjuer i andre.

4.2. Estimering i stratifisert enkel tilfeldig utvalg

Dette er den mest vanlige stratifiserte utvalgsplan. Fra hvert stratum trekkes et enkelt tilfeldig utvalg. Vi trenger litt notasjon:

- Fra stratum h: utvalg s_h av størrelse n_h
- Total utvalgsstørrelse $n = \sum_{h=1}^{H} n_h$
- Gjennomsnittet i s_h: $\bar{y}_h$
- Utvalgsanden i stratum h : n_h/N_h

Populasjonstotalen er $t = \sum_{h=1}^{H} t_h$, hvor $t_h = y$-total for stratum h. Vi ser på det tilfelle at vi har ingen tilleggsinformasjon utenom stratifiseringsvariablene og estimerer t_h med $\hat{t}_h = N_h \bar{y}_h$. Med ekstra tilleggsinformasjon kunne vi brukt en rate-type estimator for t_h. Den stratifiserte estimatoren av t er da summen av t_h – estimatorene

$$\hat{t}_{st} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h.$$

Estimering av populasjonsmiddel t/N er stratifisert middelverdi:

$$\bar{y}_{st} = \hat{t}_{st} / N = \sum_{h=1}^{H} (N_h / N) \bar{y}_h.$$

Vi ser at denne estimatoren er et vektet gjennomsnitt av utvalgsmiddel-verdiene. Egenskaper til den stratifiserte estimatoren følger fra egenskaper til ETU-estimatorer. Vi innfører følgende notasjon:

Populasjonens middelverdi i stratum h er μ_h og stratumvarians betegnes med σ_h^2. Vi har da følgende

$$E(\hat{t}_{st}) = t, \ \hat{t}_{st} \text{ er forventningsrett}$$

$$Var(\hat{t}_{st}) = \sum_{h=1}^{H} Var(\hat{t}_h) = \sum_{h=1}^{H} N_h^2 \frac{\sigma_h^2}{n_h}(1 - \frac{n_h}{N_h})$$

Estimert varians oppnås ved estimering av stratum varians med stratum utvalgsvarians,

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (y_i - \bar{y}_h)^2$$

$$\hat{V}(\hat{t}_{st}) = \sum_{h=1}^{H} N_h^2 \frac{\hat{\sigma}_h^2}{n_h}(1 - \frac{n_h}{N_h})$$

Og 95 % konfidensintervall blir: $\hat{t}_{st} \pm 1{,}96\sqrt{\hat{V}(\hat{t}_{st})}$.

Dette forutsetter at estimatoren er tilnærmet normalfordelt, dvs., utvalgsstørrelsene n_h kan ikke være for små.

Estimering av populasjonsandel i stratifisert enkelt tilfeldig utvalg

Andel i stratum *h* med et visst kjennemerke A betegnes med p_h, estimert ved $\hat{p}_h = \bar{y}_h$. Her er $y_i = 1$ hvis enhet *i* i stratum *h* har kjennemerke A, og 0 ellers. Populasjonens middelverdi er

$$p = t / N = \sum_{h=1}^{H} N_h p_h / N$$

slik at stratum middelestimator er

$$\hat{p}_{st} = \bar{y}_{st} = \sum_{h=1}^{H} (N_h / N)\hat{p}_h$$

4.3. Allokering (fordeling) av utvalgsenhetene

Det er viktig å ha gode kriterier for å bestemme størrelsene på stratumutvalgene, gitt totalt utvalg på *n* enheter og gitt stratainndelingen. Dvs., hvordan vi vil *allokere* utvalgsenhetene til strataene. Den mest vanlige allokeringen er *proporsjonal allokering*:

- Et representativt utvalg bør speile populasjonen
- Strata andeler: $W_h = N_h/N$
- Strata utvalgsandeler bør være de samme: $n_h/n = W_h$
- *Proporsjonal allokering:*

$$n_h = n\frac{N_h}{N} \Leftrightarrow \frac{n_h}{N_h} = \frac{n}{N} \text{ for alle } h.$$

Trekkesannsynlighetene i stratum *h* er $\pi_i = n_h / N_h = n / N$. Dvs., lik for alle enheter i populasjonen, men det er ikke enkelt tilfeldig utvalg. Den stratifiserte estimatoren blir da

$$\hat{t}_{st} = N \cdot \bar{y}_s .$$

Det ses på følgende måte:

$$\hat{t}_{st} = \sum_h N_h \bar{y}_h = \sum_h \frac{N_h}{n_h} \sum_{i \in s_h} y_i = \frac{N}{n} \sum_h \sum_{i \in s_h} y_i = N \cdot \bar{y}_s .$$

Et like-vektet utvalgsmiddel, vi sier at utvalget er *selv-veiende*: Hver enhet i utvalget representerer det samme antall enheter i populasjonen, *N*/*n*.

La oss nå sammenligne denne estimatoren med estimatoren i enkelt tilfeldig utvalg, $\hat{t}_{ETU} = N\bar{y}_s$.

Under proporsjonal allokering, $\hat{t}_{st} = \hat{t}_{ETU}$, samme estimator, men v*ariansene* er forskjellige:

Under enkelt tilfeldig utvalg: $Var_{ETU}(\hat{t}_{ETU}) = N^2 \cdot \frac{1}{n} \cdot (1 - \frac{n}{N})\sigma^2$

Under proporsjonal allokering: $Var(\hat{t}_{st}) = N^2 \cdot \frac{1}{n} \cdot (1 - \frac{n}{N}) \sum_{h=1}^{H} W_h \sigma_h^2 .$

Vi har følgende uttrykk for den totale populasjonsvariansen:

$$\sigma^2 = \sum_{h=1}^{H} W_h \sigma_h^2 + \sum_{h=1}^{H} W_h (\mu_h - \mu)^2 .$$

Total varians = varians *innen* strata + varians *mellom* strata.

Noen implikasjoner er:

1. Uansett stratifiseringsopplegg : Proporsjonal allokering gir mer nøyaktige estimater for populasjonstotalen enn enkelt tilfeldig utvalg.
2. Velg strata med liten variasjon, mindre strata varianser. Da vil *strata middelverdiene variere mer* og "mellomvariansen" blir større og presisjonen til estimatene øker sammenlignet med enkelt tilfeldig.

Eksempel – California skolenes kvalitetsindikator fra delkapittel 2.8

Ser på estimering av gjennomsnittlig API i 2000. Stratifiseringsvariabel er «schooltype» . Det blir da tre strata:

Stratum1: Elementary schools $N_1 = 4421$
Stratum 2: Middle schools $N_2 = 1018$
Stratum 3: High schools $N_3 = 755$

Et 5 % stratifisert utvalg, $n = 310$, med proporsjonal allokering gir følgende utvalgsstørrelser:

$n_1 = 221$
$n_2 = 51$
$n_3 = 38$

Estimering av populasjonsmiddel t/N ved stratifisert middelverdi ga følgende resultater:

$$\bar{y}_{st} = \hat{t}_{st} / N = \frac{1}{N}\sum_{h=1}^{3} N_h \bar{y}_h = 661,9$$

$$SE = 7,25$$

95 % konfidensintervall: 647,7 – 676,1 (sann verdi er 664,7).
Til sammenligning, et enkelt tilfeldig utvalg på 310 skoler ga som resultat et estimat lik 651,2 med $SE = 7,42$.
I en vanlig undersøkelse er selvsagt populasjonssnittet ukjent, og vi må bruke standardfeilen for å si noe om usikkerheten på estimatet.

4.4. Optimal allokering

Hvis det eneste vi ønsker å betrakte er estimering av en populasjonstotal t så vil naturligvis:

- velge n_h slik at variansen til stratifisert estimator er minimum

Det viser seg at

- løsningen avhenger av ukjente stratum varianser
- hvis stratum variansene er omtrent like, så vil proporsjonal allokering minimere variansen til stratifisert estimator

Det kan vises at optimal allokering er gitt ved:

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{k=1}^{H} N_k \sigma_k} = n \cdot \frac{N_h \sigma_h}{N_1\sigma_1 + N_2\sigma_2 + ... + N_H\sigma_H}.$$

Den kalles Neyman allokering (ble vist av Neyman i en artikkel fra 1934). Fortolkning av resultatet er:

- Ta mange observasjoner i stratumet hvis
  - stratum utgjør en stor del av populasjonen
  - stratumvarians er stor
- Hvis stratumvariansene er like så er dette proporsjonal allokering

Problemet, selvsagt, er at stratumvariansene er ukjente. En måte få informasjon om stratumvariansene er å ta et lite preliminært utvalg (pilot). Samtidig er variansen til den stratifiserte estimatoren ikke veldig følsom for avvik fra optimal allokering. Vi trenger derfor bare grove tilnærminger til stratum variansene.

Noen andre temaer ved allokering:

- Vanligvis er det mange studievariabler i en undersøkelse, og variablene leder til forskjellige optimale løsninger. Man kan da velge en eller to nøkkelvariabler eller bruke proporsjonal allokering som et kompromiss.
- Det er mulig å trekke inn kostnader for forskjellige typer innsamlinger som telefon, besøk, web.
- Hovedinteressen er noen ganger estimering av stratumtotaler og mindre interesse i presisjonen til estimatet for populasjonstotalen. Da bør n_h bestemmes for å oppnå ønsket nøyaktighet for estimatet av t_h, som nevnt tidligere. Hvis vi har bestemt oss for proporsjonal allokering i utgangpunktet, så kan det bety i små strata (små N_h) at utvalgsstørrelsen n_h må økes.

Vi skal nå kort beskrive noen andre typer av utvalgsplaner.

4.5. Klyngeutvalg og flertrinnsutvalg

Utvalgsplanene vi så langt har nevnt er direkte trekking av enheter i ett trinn. Av økonomiske og praktiske hensyn så kan det være nødvendig å modifisere disse utvalgsplanene. For eksempel,

- Det eksisterer ikke noe populasjonsregister, og det er umulig eller veldig kostbart å produsere et slikt register.
- Populasjonsenheter er spredt over et stort område og et direkte utvalg vil også være veldig spredt. Hvis det skal foretas besøksintervjuer så vil reisekostnader bli høye og det vil ikke være mulig å besøke alle enhetene i utvalget.

En modifisert utvalgstrekking kan gjøres ved å:

1. velge utvalget indirekte i grupper, kalt klynger (engelsk:clusters); *klyngeutvalg*
    - Populasjonen er gruppert i *klynger*
    - Utvalget består av *et utvalg av klynger* og *alle* enheter i utvalget av klynger

For eksempel, i AKU er klynger = husholdninger og enheter = personer.

2. velge utvalget i flere trinn.

Eksempler

1. *Klyngeutvalg.* Man skal ha et utvalg av elever i videregående skole (vg) i et visst område, for å undersøke røyking og alkoholbruk. Hvis en liste av vg klasser er tilgjengelig så kan vi velge et utvalg av vg klasser og gi spørreskjema til hver elev i de valgte klassene. Dette er et klyngeutvalg med vg klasser som klynger.
2. *Totrinns klyngeutvalg.* Hvis en liste av klasser ikke er tilgjengelig så kan vi først velge vg- skoler, deretter klasser og tilslutt alle elvene i de utvalgte klassene. Da har vi et totrinns klyngeutvalg med
    - PU = videregående skole
    - SU = klasser
    - Enheter = elever

Enkelt tilfeldig klyngeutvalg

Estimatoren for populasjonens middelverdi er her $\bar{y}_s$. Følgende egenskaper er viktige å merke seg:

- Estimatorens varians er sterkt påvirket av hvordan klyngene er konstruert. Variansen blir mindre jo mer y – variasjon det er i klyngene, slik at det meste av y – variasjonen ligger i klyngene. Det betyr at middelverdiene i klyngene blir liknende.
- Merk at det er motsatt i stratifiserte utvalg.
- Typisk så dannes klyngene av "nærliggende enheter' som husholdninger, skoler, sykehus på grunn av økonomiske og praktiske grunner, med liten variasjon innen klyngene:
    - Enkelt tilfeldig klyngeutvalg vil derfor medføre mye mindre presise estimater sammenlignet med vanlig enkelt tilfeldig utvalg, men til gjengjeld så får vi store kostnadsreduksjoner.

To-trinnsutvalg

En begrunnelse er at med homogene klynger og et gitt budsjett så er det ikke nødvendig å samle inn informasjon fra alle enheter i klyngene - kan isteden velge flere klynger.

- Populasjonen delt inn i N primære utvalgsenheter (PU)
- Trinn 1: Velg et utvalg s_I of PU, ofte geografiske regioner
- Trinn 2: For hver valgte PU i i s_I: Velg et utvalg s_i av enheter (sekundære utvalgsenheter, SU)
- Klyngetotaler t_i må estimeres fra utvalget.

5. Frafall i person- og husholdningsundersøkelser

Med frafall menes manglende opplysninger fra enkelte enheter i utvalget. For person- og husholdningsundersøkelser er frafall den mest alvorlige feilkilde sammen med utvalgsfeilen, ja frafall kan ofte medføre større usikkerhet enn utvalgsfeilen. Vi skal derfor gå gjennom dette problemet relativt grundig. Temaer i dette kapitlet er følgende:

- Årsaker til frafall
- Frafallsmekanismer
- Eksempler på effekten av frafall
- Effekt av frafall, en enkel analyse
- Etterstratifisering
- Eksempel på korrigering av informativt frafall
  - Modellering av informativt frafall
  - Sammenligning med etterstratifisering
- Justeringsceller og inverse svarsannsynligheter
- Kalibrering

5.1. Innledning

Frafall inntreffer i alle undersøkelser, selv i en oppgavepliktig kvartalsvis undersøkelse som Arbeidskraftundersøkelsen (AKU) hvor frafallet nå ligger rundt 20 prosent. Frafallet har vært økende gjennom mange år, men i senere tid har det stabilisert seg på grunn av bedre telefonsporing. Det er viktig å ta frafall alvorlig på grunn av to sentrale konsekvenser:

- Potensiell *skjevhet* (nesten alltid), svarutvalget er ikke representativt for populasjonen.
- Økt usikkerhet i estimater, på grunn av redusert utvalgsstørrelse.

Frafall er et sentralt problem i person/husholdningsundersøkelser, hvor det ofte er minst 50 prosent. I bedriftsundersøkelser er frafallet mye mindre, vanligvis ca. 5-10 prosent, på grunn av oppgaveplikt.
Noen begrepspresiseringer:

- *Enhetsfrafall*: enhet (person, husholdning) i utvalget gir ikke noe svar
  - Kan være meget stort, for eksempel 70 % in postale undersøkelser
  - 30 % er ikke uvanlig i telefonundersøkelser
  - 48 % i Forbruksundersøkelsen 2012, opp fra ca. 30 % for 20 år siden
- *Partielt frafall*: observasjoner mangler på noen, men ikke alle, variable for en enhet i utvalget

De vanligste *statistiske opprettingsmetoder* er *vekting* for enhetsfrafall og *imputering* for partielt frafall. Imputering betyr å predikere de manglende verdiene ved hjelp av det vi vet om den enheten som har partielt frafall. Formålet med estimeringsmetodene er å

- redusere *effekten* av frafall, ved å redusere skjevheten og korrigere den opprinnelige estimatoren beregnet for et fullt utvalg.

Metoden er vanligvis basert på en frafallsmodell, fra studier av data, og hvis frafallsmodellen stemmer noenlunde med virkeligheten så kan mye av skjevheten rettes opp.
Basis idéen ved vekting er:

- noen deler av populasjonen er underrepresentert i svarutvalget
- veie opp disse delene for å kompensere for underrepresentasjonen

En vanlig vektingsmetode: *Etterstratifisering*

5.2. Årsaker til frafall

Noen av årsakene til enhetsfrafall er:

- Ikke-kontakt, får ikke koblet til telefonnummer eller personen svarer ikke på oppringing
- Nekting: enhet ønsker ikke å delta
- Ikke i stand til å svare: for eksempel på grunn av dårlig helse eller språkproblemer

Vanlige tiltak for å øke svarprosenten:

- Rekontakt, ringer opp flere ganger eller sender purringer på andre måter

- Forandre tidspunkt for å oppnå kontakt
- Spørreskjemadesign i webundersøkelser
- Opplæring og valg av intervjuere
- Incitament, gavekort eller pengesum

Noen av årsakene til partielt frafall:

- Respondent:
  - svar ukjent'
  - nekter (følsomt ellet irrelevant spørsmål)
- Intervjuer:
  - stiller ikke spørsmålet
  - glemmer å notere svar
- Prosessering
  - svaret forkastes ved editering
  - skanningsproblemer

For mange variable er partielt frafall kun på 1-2 prosent. Det er ofte høyest for økonomiske variabler, for eksempel, total husholdningsinntekt kan ha 20% manglende data.

5.3. Frafallsmekanismer

Vi bruker sannsynlighetsspråket for å beskrive selve mekanismen som leder til frafall. Det er den vanlige statistiske tilnærmingen for en variabel (svar/frafall) vi ikke kjenner resultatet av på forhånd.
Hovedspørsmålet når det gjelder frafallsmekanismen (responsmekanismen):

- Er sannsynligheten for frafall avhengig av studievariabelen eller ikke?

Forskjellige modeller for analyse er basert på forskjellige antakelser om frafallsmekanismen. La y være studievariabelen og x verdiene av tilleggsvariable, kjent fra register for hele populasjonen.
La for en enhet (person eller husholdning) i populasjonen:

- $R = 1$ hvis enhet svarer når den er med i utvalget, og $R = 0$ hvis frafall.

Anta enhetene responderer uavhengig av hverandre. Analyse av frafallsskjevhet avhenger av antagelser om frafallsmekanismen, som vi sorterer i tre typer.
Tre typer av frafallsmekanismer:

- *Tilfeldig frafall.* Sannsynligheten for frafall er uavhengig av y og x
  - $P(R = 0| y, x) = P(R = 0)$
  - De observerte verdiene til y danner et rent tilfeldig delutvalg av det opprinnelige utvalget
- *Stratifisert tilfeldig frafall.* Sannsynligheten for frafall avhenger av x, men ikke y.
  - $P(R = 0| y, x) = P(R = 0|x)$
  - De observerte verdiene of y danner tilfeldige utvalg innen delklasser definert ved x
- *Informativt frafall.* Sannsynligheten for frafall avhenger av y og muligens x også. I dette tilfellet er frafallsmekanismen *ikke-ignorerbart.* For eksempel, de som ikke svarer i helseundersøkelsen har dårligere helse enn de som svarer.

I mange samfunnsvitenskapelige undersøkelser så er selve svarandelen den mest rapporterte kvalitetsindikator, men den trenger ikke ha noen sammenheng med hvor stor skjevheten er. Vi skal nå se på tre eksempler for å illustrere hvordan frafall kan lede til sterkt villedende statistisk analyse, selv når svarandelen er høy. I alle tilfeller så er frafallsmekanismen er informativ (på engelsk: MNAR, missing not at random). I to av eksemplene ser vi på hvordan vi kan korrigere for frafall.

5.4. Tre frafallseksempler

Frafallseksempel 1. Et klassisk eksempel, med svarandeler 81-85 prosent

Politisk meningsmåling før det amerikanske presidentvalget i 1948 av instituttet Roper gjaldt hovedkandidatene Truman for Demokratene og Dewey for Republikanerne. Det var meningsmålinger i juli, august, september og oktober. Valget var i november. Resultatene ble:

Tabell 5.1 Meningsmålinger før det amerikanske presidentvalget i 1948

	Juli	August	Sept	Okt	Valg
Truman	37,8	37,0	35,2	40,4	**49**
Dewey	55,5	52,4	57,0	53,4	**45**
Andre	6,7	10,5	7,7	6,2	6
utvalgsstørrelse	3011	3490	3490	3500	
Antall svar	2510	2951	2936	2841	
Frafall	501	539	554	659	
(Prosent)	(18,6)	(15,4)	(15,9)	(18,8)	

Vi ser her at på alle meningsmålingene er Dewey langt foran Truman. Vi legger også merke til at det er høye svarandeler etter nåtidens standard. Det ble brukt såkalt kvotesampling:

- Hver intervjuer fikk tildelt et visst antall personer å intervjue. Intervjueren skulle intervjue et visst antall personer i forskjellige kategorier, basert på bosted, kjønn, alder, rase, økonomisk status og andre variable. Hensikten med kvotesampling er å sikre at utvalget representerer populasjonen i alle viktige aspekter. Hver intervjuer kunne fritt bestemme hvilke personer som skulle intervjues. Det er tydelig at denne friheten skapte en seleksjonseffekt i favør av Dewey.

Det var en klar mistanke om at skjevheten skyldes større frafall blant økonomisk svake grupper. En analyse publisert i *Journal of the American Statistical Association* i 1988 viser imidlertid at det er ikke nok til å rette opp frafallsskjevheten å bare ta hensyn til sosio-økonomisk (gruppert med hensyn på utdanning og økonomiske forhold) tilhørighet. For å kompensere for frafall må det antas en informativ frafallsmodell:

- Sannsynligheten for svar avhenger av hvilken kandidat personen vil stemme på, innen hver sosio-økonomisk gruppe.
- Dette gir Truman et estimat på 51 prosent oppslutning.
- Metoden som brukes er imputering, og det anslås at 93-99% stemte for Truman i frafallsgruppen
- Hvis man bruker en såkalt etterstratifisert modell etter sosio-økonomisk gruppe gir det estimat på kun 41% for Truman. Vi skal behandle etterstratifisering i delkapitttel 5.6

Frafallseksempel 2. Valgundersøkelsen i Norge 2009

Vi skal bruke undersøkelsen til å estimere en størrelse som er kjent i populasjonen, nemlig valgdeltakelsen som var 76,4 %. Det er da mulig å måle effekten av frafall på estimeringsskjevheten. Kort oppsummert:

- Utvalget var på 2944 personer
- Antall svar: 1782, slik at frafallet er 1162 som tilsvarer 39,5 %
- Av de 1782 personene som deltok i undersøkelsen, sa 1506 at de stemte i Stortingsvalget, dvs. estimatet på valdeltakelsen er 1506/1782 = 0,845 = 84,5 %.
- Feilmarginen blir 1,7 %:

$$2 \cdot SE = 2\sqrt{\frac{0{,}845 \cdot 0{,}155}{1782}} = 2 \cdot 0{,}00857 = 0{,}017.$$

- Estimatet 84,5 % er skjevt på grunn av høyere frafall blant "ikke-velgere". Frafallsmekanismen er informativ, slik at svarutvalget er ikke representativt for frafallsgruppen (typisk tilfelle)

Frafallseksempel 3. Estimering av en-person husholdninger i Norge i 1992

Basert på data fra Forbruksundersøkelsen i 1992. Populasjonen er alle personer 15 år og eldre. Utvalget bestod av 1698 personer og var selvveiende, dvs., alle personer har samme trekke-sannsynlighet. Vi skal se på estimering av antall en-person husholdninger. Norge har et familieregister. Det betyr at familiestørrelsen til hver person i Norge er kjent.

Det er viktig å ha korrekte tall på husholdninger av forskjellige størrelse og type for økonomiske prognoser og kommunal planlegging av boliger, skoler og barnehager. Noen forskjeller mellom familie og husholdningsbegrepet er at

(i) samboere regnes som to familier
(ii) studenter som ikke bor hjemme regnes med i familien med foreldrene.

Resultatene angående husholdningsstørrelse og familiestørrelse er gitt i neste tabell.

Tabell 5.2 Husholdningsstørrelse og frafall etter familiestørrelse. Fra Forbruksundersøkelsen 1992

Familie-størrelse	Husholdningsstørrelse					Total	Frafall	% frafall
	1	2	3	4	5+			
1	83	48	29	9	2	162	153	**48,6**
2	9	177	37	4	3	230	160	**41,0**
3	10	25	131	40	6	212	91	**30,0**
4	2	13	37	231	17	300	123	**29,1**
5+	1	4	4	17	181	207	60	**22,5**
Total	**105**	267	229	301	209	1111	587	**34,6**

For eksempel, det er 48 personer som er registrert som en-person familier og samtidig tilhørende to- person husholdninger (antakelig samboere). Vi ser at frafall avtar kraftig med familiestørrelse.

Populasjonsstørrelsen pr. 1.1.93 var $N = 4\ 131\ 874$. Standard estimat for antall en-person husholdninger blir, uten å ta hensyn til frafall:

$$\frac{105}{1111} \cdot N = 390\,501.$$

En kvalitetsundersøkelse av FoB1990 anslo at antall en-person husholdninger var 626 000. Dvs., standardestimatet *underestimerer «enormt»* antall en-person husholdninger. Hvis tallet 390 500 hadde blitt publisert offentlig så ville SSB miste mye av troverdigheten som offisiell statistikk leverandør. Underestimeringen skjer blant annet fordi frafall blant en-person familier er mye høyere enn for større familier. Dette kan rettes ved modellering av svarsannsynligheter, som vi skal se på i delkapittel 5.6.
Nå først en enkel analyse av effekten av frafall som vil vise hva etterstratifisering kan korrigere.

5.5. Effekt av frafall, en enkel analyse

Vi antar en forenklet deterministisk populasjonsmodell for frafall:

U_R = svarpopulasjonen, med størrelse N_R
U_M = frafallspopulasjonen, med størrelse N_M ; bruker indeks M for «missing».
$q_R = N_R/N$ = forventet svarandel.

Vi betrakter nå er enkelt tilfeldig utvalg s av størrelse n. Svarutvalget s_r er den delen av s som er fra U_R med størrelse n_r. Problemet er å estimere populasjonsmiddel $\mu = \overline{Y} = \sum_{i=1}^{N} y_i / N$. Vi skal se på skjevheten til standard estimatoren som er det observerte gjennomsnittet i svarutvalget, $\overline{y}_r = \sum_{i \in s_r} y_i / n_r$.

Populasjonsmiddelverdiene i U_R og U_M er $\overline{Y}_R$ og $\overline{Y}_M$.

Gitt at vi har n_r observasjoner i svarutvalget, så er s_r et enkelt tilfeldig utvalg fra U_R. Det betyr at skjevheten blir

$$E(\overline{y}_r) - \overline{Y} = \overline{Y}_R - \overline{Y} = \overline{Y}_R - \frac{N_R\overline{Y}_R + N_M\overline{Y}_M}{N} = \frac{(N-N_R)\overline{Y}_R - (N-N_R)\overline{Y}_M}{N}$$

$$= (1-q_R)(\overline{Y}_R - \overline{Y}_M).$$

Herav ser vi følgende mulige konsekvenser av frafall:

1. Skjevheten er uavhengig av n, kan ikke reduseres ved å øke n
2. Skjevheten øker med økende frafallsandel (1 - q_R)
3. Skjevheten øker når $|\overline{Y}_R - \overline{Y}_M|$ øker
4. Hvis $\overline{Y}_R = \overline{Y}_M$ så er det ingen skjevhet. Vi har da *ignorerbar* frafallsmekanisme. Det betyr at frafallet har ingen sammenheng med variabelen y.

Det er urealistisk å anta at $\overline{Y}_R = \overline{Y}_M$, men innen mindre delpopulasjoner er det ikke så urimelig. Det gjelder spesielt hvis variablene som brukes til å dele opp populasjonen er høyt korrelerte med y. Dette kalles *etterstratifisering*, som er et mye brukt verktøy for å korrigere for frafall når stratifisert tilfeldig frafall er en rimelig modell for frafallsmekanismen.

5.6. Etterstratifisering

Vi kan beskrive etterstratifiseringsopplegget på følgende måte:

1. Stratifiserer *etter* at data er samlet inn

2. Stratifiserer ved å bruke tilleggsvariable som deler opp populasjonen i homogene grupper
3. Stratifiserer etter varierende svarandeler

Vi trenger litt notasjon La H være antall etterstrata. For etterstratum h, U_{Rh} er svardelen og U_{Mh} er frafallsdelen. Videre,

q_h = svarandel i etterstratum h
$W_h = N_h/N$, hvor N_h er populasjonsstørrelsen til etterstratum h
$\bar{Y}_{Rh}$ = middelverdi i svarstratum h i populasjonen
$\bar{Y}_{Mh}$ = middelverdi i frafallsstratum h i populasjonen.

Man bør velge etterstrata slik at q_h varierer så mye som mulig og $\bar{Y}_{Rh} \approx \bar{Y}_{Mh}$. Dvs., velg etterstratifiseringsvariable som er høyt korrelerte med y, som nevnt tidligere.

La $\bar{y}_h$ være observert middel fra etterstratum h. Etterstratifiseringsestimatoren er gitt ved

$$\hat{\bar{y}}_{est} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h, \text{ og for totalen, } \hat{t}_{est} = \sum_{h=1}^{H} N_h \bar{y}_h.$$

Det kan vises at skjevheten til etterstratifiseringestimatoren blir

$$E(\hat{\bar{y}}_{est}) - \bar{Y} = \sum_{h=1}^{H} (1-q_h) W_h (\bar{Y}_{Rh} - \bar{Y}_{Mh}).$$

Etterstratifisering retter opp all frafallsskjevhet hvis svarutvalget i etterstratum h er representativt for frafallsgruppen i stratum h. Det betyr i så fall at vi har stratifisert tilfeldig frafall.

Eksempel. Valgundersøkelsen for Stortingsvalget 1993

Som i frafallseksempel 2, skal vi se på estimering av valgdeltakelsen. I 1993 var valgdeltakelsen på 75,5 %.Vi har omtrent samme skjevhet som i det eksemplet med estimat lik observert andel som stemte. Vi skal se om den estimatoren kan bli forbedret ved å etterstratifisere etter valgdeltakelsen i 1989. Resultatene ble:

- Utvalg: 3000
- Antall svar etter to gjenbesøk (rekontakter): 1403 slik at frafall er 1597 som tilsvarer 53,2 %.
- Av de 1403 sier 1190 at de stemte i valget så estimert valgdeltakelse blir 1190/1403 = 0,848. Dvs., 84,8 % som gir en relativ skjevhet på 12,3 %.

La oss nå prøve å korrigere frafallsskjevheten ved etterstratifisering etter valgdeltakelse i Stortingsvalget 1989:

- Etterstratum 1= deltok i valget 1989: N_1 = 2 510 669
- Etterstratum 2= deltok ikke i 1989: N_2 = 508 288
- Etterstratum 3= nye velger i 1993: N_3 = 241 000

I svarutvalget, med y=1/0 hvis deltok/deltok ikke i 1993 valget:

Tabell 5.3 Valgundersøkelsen for Stortingsvalget i 1993.Valgdeltakelse i 1993 stratifisert etter valgdeltakelse i 1993

Etterstratum	1	2	3
Totalt	1192	115	96
Antall som deltok i 1993	1060	57	73
$\bar{y}_h$	0,889	0,496	0,760

Populasjonsstørrelsen $N = N_1 + N_2 + N_3$ = 3 259 957. Etterstratifiseringsestimatet for total valgdeltakelse i 1993 blir:

$$\hat{t}_{est} = N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3$$
$$= 2510669 \cdot 0{,}889 + 508288 \cdot 0{,}496 + 241000 \cdot 0{,}760$$
$$= 2667256.$$

Etterstratifisert estimat for valgdeltakelsen i 1993 blir dermed:

$$\hat{\bar{y}}_{est} = \hat{t}_{est} / N = 2667256 / 3259957 = 0{,}818 = 81{,}8\ \%.$$

Retter opp kun 32 prosent av skjevheten til observert andel.

Vi ser nå igjen på estimeringen av en-person husholdninger. Resultatene fra undersøkelsen er gjengitt i tabell 5.2 som gir estimerte sannsynligheter for forskjellige husholdningsstørrelser blant de som *svarer*. For eksempel, estimert sannsynlighet for husholdningsstørrelse 1, gitt familiestørrelse 1, blant respondenter er andelen 83/162 = 0,512.

Registeret over familiestørrelse for hele Norges befolkning pr. 1.1.93 gir følgende informasjon.

Tabell 5.4 Familieregisteret over antall familier av forskjellige størrelser

Familie-størrelse, *x*	Antall familier	Antall personer (N_x)
1	793 869	793 839
2	408 440	816 880
3	261 527	784 581
4	266 504	1 066 016
5+	127 653	670 528
Total	1 857 993	4 131 874

Etterstratifisering av antall en-person husholdninger:

Tabell 5.5 Andel med husholdningsstørrelse 1 i etterstrata

Etterstrata: Familiestørrelse *x=h*	1	2	3	4	5+
Observert andel med hush.størrelse 1	0,5123	0,0391	0,0472	0,0067	0,0048

z_i =1 hvis husholdningsstørrelse er lik 1, og 0 ellers. Etterstratifiseringsestimatoren etter familiestørrelse blir:

$\hat{t}_{est} = \sum_{h=1}^{5+} N_h \bar{z}_h$

= 793869·0,5123+816880·0,0391+…+670528·0,0048 = 486 032.

Sammenlignet med det *uvektede estimatet 390 501*, så reduserer etterstratifisering skjevheten med ca. 40 %. Dette tyder på at vi har informativt frafall. En analyse ble foretatt hvor responsmodellen antas å være informativ, spesifikt at sannsynligheten for svar avhenger av *husholdningsstørrelse* og bosted (tettbygd, spredtbygd), med en logistisk regresjonsmodell. Estimater er gitt i tabellen nedenfor.

Tabell 5.6 Standard estimater, etterstratifiseringsestimater og modellbasert estimater

	Standard	Etterstratifisering	Modellbasert
Husholdningsstørrelse =1	391 000	486 000	595 000
Total	1 599 000	1 682 000	1 765 000

Modellbaserte estimater er basert på de estimerte sannsynligheter for at *Y* tar verdien 1 gitt familiestørrelse, $\hat{P}(Y=1 \mid x)$, som vises i neste tabell, men $\hat{P}(Y=1 \mid x, R=1)$ i parentes.

Tabell 5.7 Estimerte sannsynligheter for husholdningsstørrelse 1, i prosent. (I parentes, observert andel)

Fam. størrelse *x*	1	2	3	4	5+
$\hat{P}(Y=1 \mid x)$	60.01 (51,23)	5.27 (3,91)	7.53 (4,72)	1.06 (0,67)	0.84 (0,48)

Det er tydelig at det er ikke-ignorerbart frafall. Sannsynligheten for frafall avhenger av variabelen av interesse, husholdningsstørrelse. Modellbasert estimat blir:

$$\hat{H}_1 = \sum_{x=1}^{5+} N_x \hat{P}(Y=1 \mid x)$$
$$= 793869 \cdot 0,6001 + 816880 \cdot 0,0527 + ... + 670528 \cdot 0,0084$$
$$= 595\ 462$$

Standardfeil og 95 % konfidensintervall basert på etterstratifieringsestimator, under stratifisert tilfeldig frafall

$\hat{t}_{est} = \sum_h N_h \bar{y}_h$, med $\bar{y}_h = \frac{1}{n_{rh}} \sum_{i \in s_{rh}} y_i$, middelverdi i etterstratum *h* i svarutvalget.

Utvalgsvariansen i etterstratum *h*: $\hat{\sigma}_h^2 = \frac{1}{n_{rh}-1} \sum_{i \in s_{rh}} (y_i - \bar{y}_h)^2$.

Betinget på svarutvalgsstørrelsene i etterstrata, n_{rh}, så er standardfeilen til etterstratifiserings-estimatoren lik standardfeilen fra vanlig stratifisert estimator:

$$\hat{V}(\hat{t}_{est}) = \sum_h N_h^2 (1 - \frac{n_{rh}}{N_h}) \frac{\hat{\sigma}_h^2}{n_{rh}} \quad og \quad SE_{est} = SE(\hat{t}_{est}) = \sqrt{\hat{V}(\hat{t}_{est})}$$

Det gir 95 % konfidensintervallet: $\hat{t}_{est} \pm 1,96 \cdot SE_{est}$.

Vi illustrerer dekningsgraden til konfidensintervallet ved å trekke noen enkle tilfeldige utvalg fra API populasjonen av California skoler med forskjellige svarandeler i etterstrata definert ved skoletype, og sammenligner med det vanlige konfidensintervall basert på rent utvalgsmiddel fra hele svarutvalget.

Tabell 5.8 API for Californiaskoler. Etterstrata er 1(E), 2(H), 3(M) med svarandeler r1, r2, r3. Estimert konfidensnivå for 95 % konfidensintervall for etterstratifisering og utvalgsmiddel, basert på 1000 simuleringer av ETU. Frafallsmodell: Stratifisert tilfeldig frafall etter skoletype

n	r1	r2	r3	Konf.nivå etterstrat.	Konf.nivå middel
200	0,5	0,5	0,5	0,9483	0,9486
200	0,3	0,8	0,9	0.9514	**0,8727**
500	0,7	0,2	0,5	0,9450	0,9363
500	0,5	0,5	0,5	0,9436	0,9448
1000	0,3	0,8	0,9	0,9498	**0,7892**
2000	0,3	0,8	0,9	0,9514	**0,6025**
2000	0,6	0,6	0,6	0,9496	0,9517

Etterstratifisert konfidensintervall har korrekt dekningsgrad generelt. Konfidensintervallet basert på utvalgsmiddel fungerer bare når svarandelen er de samme i alle etterstrata. Det betyr at svarutvalget er et enkelt tilfeldig utvalg fra «bruttoutvalget» og da vil svarutvalgets middel være forventningsrett, dvs. ingen skjevhet.

5.7. Justeringsceller og kalibrering

Det kan være at N_h i ønskede etterstrata er ukjente. Vi kan da bruke en lignende estimator såfremt størrelsene n_h på etterstrata i det opprinnelige utvalget er kjente. Etterstataene kalles nå justerings-celler. Vi bruker da $N\cdot(n_h/n)$ som et estimat for N_h . Justeringsestimatoren blir derfor:

$$\hat{t}_{just} = N\sum_{h=1}^{H}\frac{n_h}{n}\bar{y}_h.$$

La $\hat{q}_h = n_{rh}/n_h$ være observert svarandel i etterstratum h, dvs. et estimat for svarsannsynligheten i etterstratum h. Vi kan da uttrykke justeringsestimatoren på følgende form:

$$\hat{t}_{just} = N\sum_{h=1}^{H}\frac{n_h}{n}\cdot\frac{1}{n_{rh}}\sum_{i\in s_{rh}} y_i = N\sum_{h=1}^{H}\sum_{i\in s_{rh}}\frac{1}{n}\cdot\frac{1}{\hat{q}_h}y_i.$$

$\hat{t}_{just}$ justerer vektene til det opprinnelige middelestimat fra $1/n_r$ til $1/(n\cdot\hat{q}_h)$ og derfor kalles strataene for justeringsceller. Vi ser at

- essensielt så vektes observasjonene med estimerte inverse svarsannsynligheter
- vekting opp til «bruttoutvalget» s, i motsetning til populasjonen som i etterstratifisering
- hvis bruttoutvalget gjenspeiler populasjonen så er forskjellen mellom etterstratifisering og justeringsceller liten.

Justeringsceller med inverse svarsannsynligheter som vekter er en av de mest vanlige metoder for å korrigere for frafall.

Vi skal nå kort si litt om en annen vanlig brukt vektingsmetode, kalt *kalibrering*.

Dette er en metode som tilfredsstiller visse *kalibrering restriksjoner.* Det er en design-basert tilnærming hvor vi «starter» med HT-estimatoren $\hat{t}_{HT} = \sum_{i\in s}(1/\pi_i)y_i$. Vektene $d_i = 1/\pi_i$ kalles designvektene. La s_r betegne svarutvalget. Relevant tilleggsinformasjon er kjente totaler av x-variable som er korrelert med studievariabelen:

$$t_{x1} = \sum_{i=1}^{N} x_{1i}, t_{x2} = \sum_{i=1}^{N} x_{2i}, ..., t_{x\mathrm{k}} = \sum_{i=1}^{N} x_{\mathrm{k}i}$$

Endelige utvalgsvekter w_i oppfyller kaliberingsrestriksjonene:

$$\sum_{i\in s_r} w_i x_{1i} = t_{x1}, \sum_{i\in s_r} w_i x_{2i} = t_{x2}, ..., \sum_{i\in s_r} w_i x_{\mathrm{k}i} = t_{x\mathrm{k}}.$$

Kalibrert estimator for y-total: $\hat{t}_{cal} = \sum_{i\in s_r} w_i y_i$.

Vi velger vanligvis de kalibrerte vektene slik at "avstanden" mellom d_i og w_i er så liten som mulig. Det kan vises at etterstratifisering er et eksempel på kalibrering, i den forstand at vektene er kalibrert med hensyn på størrelsene på etterstrata.

Som nevnt tidligere så er det andre hovedtema i frafallsbetraktningene er imputering. Det behandles i neste kapittel.

6. Imputering

Temaer i dette kapittel er:

- Standard imputeringsmetoder i nasjonale statistikkbyråer
- Multippel imputering
- Regresjonsbaserte imputeringsmetoder.

Imputering betyr å fylle inn for hver manglende dataverdi ved å predikere de manglende verdiene. Mest brukt for partielt frafall, men kan også brukes for enhetsfrafall.
Det er viktig å legge merke til at partielt frafall skaper problemer selv når frafallet er rent tilfeldig, fordi det resulterer i få komplette enhetsdata.
Den vanlige *imputeringsbaserte* estimatoren for en gitt variabel *y,* for å estimere populasjon total eller gjennomsnitt, er å bruk estimatoren konstruert for hele utvalget, basert på de observerte og imputerte data.

Det er to viktige aspekter ved imputeringsmetoder:

- Rette variansestimater for den imputeringsbaserte estimatoren
- Produsere komplette data sett som tillater standard statistisk analyse. Det er da viktig at de imputerte verdiene reflekterer den riktige variasjonen i data.

6.1. Standard imputeringsmetoder, mye brukt i statististiske sentralbyråer

La generelt y_i^* betegne den imputerte verdien for en manglende y_i – verdi. Tre vanlige imputeringsmetoder er:

(i) *Middel* imputering: $y_i^* = \bar{y}_r$
Hvis dette gjøres innen etterstrata, så tilsvarer dette at den imputeringsbaserte estimatoren blir lik den etterstratifiserte estimatoren.

(ii) *Hot-deck* imputering (typisk innen etterstrata):
y_i^* trekkes tilfeldig fra de observerte *y* verdiene, med tilbakelegging.

(iii) *Nærmeste nabo* imputering: Finn en *donor* (*giver*) i svarutvalget s_r basert på «nærhet» målt ved tilleggsvariabler. Med denne metoden får man imputert verdiene til alle variablene hvis det er enhetsfrafall.

Vi skal nå se på standard analyse med imputerte verdier og betrakter det enklest mulige tilfelle. Dvs., enkelt tilfeldig utvalg med rent tilfeldig frafall og ingen tilleggsinformasjon. To vanlige imputerings-metoder er middel imputering og hot-deck imputering. Vi merker oss at middel imputering ikke kan brukes hvis det totale datasettet, inkludert imputerte verdier, skal reflektere forventet variasjon i utvalget.

Betrakt nå standard analyse for populasjonsmiddelet $\bar{Y}$ basert på komplette utvalgsdata; observerte og imputerte data. Da er estimatoren gitt ved $\bar{y}_s$, utvalgsmiddel hvis hele *s* observeres, og utvalgsvariansen er lik

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y}_s)^2.$$

For store *n*, *N*-*n* så er standard 95 % konfidensintervall lik:

$$\text{KI: } \bar{y}_s \pm 1{,}96\hat{\sigma}\sqrt{\frac{1}{n}-\frac{1}{N}}$$

Med frafall, så blir dette standard KI, basert på det komplette datasettet med observerte og imputerte verdier,

$$\text{KI}^*\text{: } \bar{y}_s^* \pm 1{,}96\hat{\sigma}_*\sqrt{\frac{1}{n}-\frac{1}{N}}$$

Her er $\bar{y}_s^*, \hat{\sigma}_*^2$ lik $\bar{y}_s, \hat{\sigma}^2$ basert på det komplette datasettet med observerte og imputerte verdier.

6.2. Dekningsgrad for konfidensintervall med middel imputering og hot-deck imputering*

Dekningsgrad med middel imputering

$$W_r = \frac{\bar{y}_r - \bar{Y}}{\hat{\sigma}_r\sqrt{\frac{1}{n_r} - \frac{1}{N}}} \sim N(0,1) \text{ tilnærmet}$$

$$\hat{\sigma}_*^2 = \frac{n_r - 1}{n-1}\hat{\sigma}_r^2 \text{ slik at, med } \hat{r} = n_r / n,$$

$$KI^* \approx \bar{y}_r \pm 1{,}96 \cdot \hat{r}\hat{\sigma}_r\sqrt{\frac{1}{n_r} - \frac{1}{N}}.$$

Konfidensnivået til *KI** blir da:

$$C_* = P(\bar{y}_r - 1{,}96 \cdot \hat{r}\hat{\sigma}_r\sqrt{\tfrac{1}{n_r} - \tfrac{1}{N}} \le \bar{Y} \le \bar{y}_r + 1{,}96 \cdot \hat{r}\hat{\sigma}_r\sqrt{\tfrac{1}{n_r} - \tfrac{1}{N}})$$

$$= P(-1{,}96 \cdot \hat{r}\hat{\sigma}_r\sqrt{\tfrac{1}{n_r} - \tfrac{1}{N}} \le \bar{y}_r - \bar{Y} \le 1{,}96 \cdot \hat{r}\hat{\sigma}_r\sqrt{\tfrac{1}{n_r} - \tfrac{1}{N}})$$

$$= P(-\hat{r} \cdot 1{,}96 \le W_r \le \hat{r} \cdot 1{,}96)$$

Dvs., $C_* = P(\bar{Y} \in KI^*) = P(|W_r| \le \hat{r} \cdot 1{,}96)$, gitt i tabell 6.1 for utvalgte frafallsprosenter.

Tabell 6.1 Konfidensnivå til standard 95 % konfidensinterval med middel imputering

Frafall (%)	0	10	20	30	40	50
Konfidensnivå	0,950	0,922	0,883	0,830	0,760	0,673

Dekningsgrad med hot-deck imputering

Følgende resultater kan vises:

$$E(\bar{y}_s^*) = \bar{Y}$$

$$Var(\bar{y}_s^*) \approx \sigma^2 \frac{1}{n}\left[(1-\hat{r}) + \frac{1}{\hat{r}}\right]$$

$$E(\hat{\sigma}_*^2) \approx \sigma^2$$

$$W_* = \frac{\bar{y}_s^* - \bar{Y}}{\hat{\sigma}_*\sqrt{\frac{1}{n}\{(1-\hat{r}) + \frac{1}{\hat{r}}\}}} \sim N(0,1) \text{ tilnærmet}$$

Konfidensnivå blir:

$$C_* = P(\bar{y}_s^* - 1{,}96 \cdot \hat{\sigma}_*\sqrt{\tfrac{1}{n} - \tfrac{1}{N}} \le \bar{Y} \le \bar{y}_s^* + 1{,}96 \cdot \hat{\sigma}_*\sqrt{\tfrac{1}{n} - \tfrac{1}{N}})$$

$$\approx P(-1{,}96 \cdot \hat{\sigma}_*\sqrt{\tfrac{1}{n}} \le \bar{y}_s^* - \bar{Y} \le 1{,}96 \cdot \hat{\sigma}_*\sqrt{\tfrac{1}{n}})$$

$$= P(-1{,}96 \le \frac{\bar{y}_s^* - \bar{Y}}{\hat{\sigma}_*\sqrt{\frac{1}{n}}} \le 1{,}96) = P(-1{,}96 / \sqrt{1 - \hat{r} + \frac{1}{\hat{r}}} \le W_* \le 1{,}96 / \sqrt{1 - \hat{r} + \frac{1}{\hat{r}}})$$

Dvs., $C_* \approx P(|W_*| \le 1{,}96 / \sqrt{1 + \frac{1}{\hat{r}} - \hat{r}})$. Sammen med tabell 6.1 så har vi dekningsgraden til KI* med hot-deck imputering i tabell 6.2.

Tabell 6.2 Konfidensnivå til standard 95 % konfidensinterval med hot-deck og middel imputering

Frafall (%)	Middel imputering	Hot-deck imputering
0	0,95	0,95
10	0,922	0,925
20	0,883	0,896
30	0,830	0,864
40	0,760	0,826
50	0,673	0,785

6.3. Multippel imputering for variansestimering

Vi ser at selv hot-deck imputering ikke gir tilstrekkelig variasjon i de imputerte dataene. En mulig løsning er *multippel imputering.* Vi foretar da *m* hot-deck imputeringer for hver manglende verdi og oppnår *m* komplette utvalg med tilsvarende middelestimater og variansestimater:

$\bar{y}_s^*(i), \hat{\sigma}_*^2(i)$ for $i = 1,...,m$

som er $\bar{y}_s, \hat{\sigma}^2$ basert på de *m* komplette utvalgene.

Gjennomsnitt av middelestimatene og variansestimatene:

$$\bar{\bar{y}}_s^* = \sum_{i=1}^m \bar{y}_s^*(i)/m \text{ og } \bar{\sigma}_*^2 = \sum_{i=}^m \hat{\sigma}_*^2(i)/m .$$

Et «direkte» standard konfidensintervall:

$$KI^* : \bar{\bar{y}}_s^* \pm 1{,}96 \cdot \bar{\sigma}_* \sqrt{\tfrac{1}{n} - \tfrac{1}{N}} .$$

Problemet nå er at $\bar{\sigma}_* \sqrt{\tfrac{1}{n} - \tfrac{1}{N}}$ måler variasjonen kun *innen* utvalgene. Det er nødvendig å inkludere et mål på variasjon *mellom* de *m* utvalgene; dvs., måle usikkerheten på grunn av imputering,

$$B_* = \frac{1}{m-1} \sum_{i=1}^m (\bar{y}_s^*(i) - \bar{\bar{y}}_s^*)^2 .$$

La f_{mis} være frafallsandelen, og erstatt $\bar{\sigma}_*^2 (\tfrac{1}{n} - \tfrac{1}{N})$ med

$$V_* = \bar{\sigma}_*^2 (\frac{1}{n} - \frac{1}{N}) + (\frac{1}{1 - f_{mis}} + \frac{1}{m}) B_*$$

Tilsvarende 95 % konfidensintervall: $\bar{\bar{y}}_s^* \pm 1{,}96 \sqrt{V_*}$.

Hvis imputeringene er basert på en såkalt Bayesiansk modell, dvs. at de imputerte verdiene er trukket fra en aposteriori fordeling gitt frafall, så skal 1/(1-f_{mis}) erstattes med 1. Slike Bayesianske imputeringer brukes ikke foreløpig i SSB.

6.4. Mer avanserte modellbaserte imputeringsmetoder*

Disse metodene krever en modellbasert tilnærming; y_i antas å være verdier av tilfeldige variable Y_i. Vi skal beskrive kort to regresjonsbaserte imputeringsmetoder.

1. Regresjonsimputering

Anta en regresjonsmodell for Y_i mot x_i, hvor x_i er tilgjengelig også for frafallsgruppen, for eksempel, $E(Y_i) = \beta x_i, Var(Y_i) = \sigma^2 x_i$. Estimer β fra svarutvalget s_r med

$$\hat{\beta} = \sum_{i \in s_r} Y_i / \sum_{i \in s_r} x_i$$

og for alle enheter i frafallsgruppen, prediker y_i med $y_i^* = \hat{\beta} x_i$.

Det er et problem med denne imputeringsmetoden, den gir ikke nok variasjon i de imputerte verdiene til å gjenskape variasjonen i frafallsgruppen. Da er det mulig å bruke en form for hot-deck imputering på observerte residualer som beskrives i neste avsnitt.

2. Residual regresjon imputering

Siden $Var\{(Y_i - \beta x_i)/\sqrt{x_i}\} = \sigma^2$, så er standardiserte observerte residualer $e_i = (y_i - \hat{\beta} x_i)/\sqrt{x_i}$.

For $i \in s - s_r$, trekkes en verdi e_i^* tilfeldig fra mengden av standardiserte residualer i svarutvalget, $\{e_j : j \in s_r\}$.

Imputert verdi er da gitt ved: $y_i^* = \hat{\beta} x_i + e_i^* \sqrt{x_i}$.

Underliggende antakelse for frafallsmekanismen er at vi har stratifisert tilfeldig frafall. Dvs., sannsynligheten for svar for enhet *i* kan avhenge av x_i, men er uavhengig av y_i.

Hvis basis estimator for fullt utvalg er rate-estimator, $\hat{t}_R = X_\circ (\sum_{i \in s} y_i / \sum_{i \in s} x_i)$, så blir den imputeringsbaserte estimatoren lik

$$\hat{t}_{R,I} = X_\circ \frac{\sum_{s_r} y_i + \sum_{s - s_r} y_i^*}{\sum_s x_i} .$$

7. Utvalgsplaner og estimering for økonomisk statistikk. Bedrifts- og foretaksundersøkelser

Temaer i dette kapitlet er følgende:

- Vanlige utvalgsplaner i bedrifts- og foretaksundersøkelser
  - Kriteriet dekningsgrad istedenfor representativitet
  - Store, mellomstore og små enheter
  - Stratifisering
- Illustrerende eksempel med industri-data
- Eksempel med rateestimator i en SSB-statistikk, modellbasert

7.1. SSBs økonomiske utvalgsplaner

I SSBs økonomiske utvalgsundersøkelser brukes ofte utvalgsplaner med følgende trekk:

- Ikke lenger et mål at utvalget skal være en miniatyr av populasjon
- Viktig å få med de største bedriftene, målt med antall ansatte
  - Viktige variabler har mye høyere variasjon i populasjonen enn for person-undersøkelser
  - variasjoner i analysevariable for de store enhetene er større enn for mindre enheter
- Interessert i høy dekningsgrad målt ved omsetningsandel i utvalget av total populasjons-omsetning, hvor omsetning = all salg av varer og tjenester.
- Vanlig utvalgsopplegg:
  - Stratum av «store» enheter: Fulltelling
  - Ellers: stratifisert enkelt tilfeldig utvalg
  - Cut-off: De minste enhetene som bidrar lite til totalen holdes utenfor utvalget.

Vi skal nå bruke data fra industribedrifter i tre næringer for å illustrere hvorfor dette utvalgsopplegget gir mest nøyaktig estimering, og det er den total omsetningen vi ønsker å estimere. I dette eksemplet kjenner vi den totale omsetningen, noe som gjør at det er mulig å se hva egenskapene er til forskjellige utvalgsplaner og estimeringsmetoder.
Vi starter med å se på egenskapene til estimeringsmetoden basert på vanlig enkelt tilfeldig utvalg.

Eksempel: Industri bedrifter med enkelt tilfeldig utvalg

Populasjonen er 2830 bedrifter (kalles virksomheter i SSB) fra næringene 10, 20 og 24 hvor

- 10 = næring- og nytelsesmidler
- 20 = kjemikalier og kjemiske produkter
- 24 = metaller

Den totale omsetningen i denne populasjonen er ***t* = 267 950.** Vi skal se på problemet med å *estimere* denne totale omsetningen *t* i 2011 basert på et enkelt tilfeldig utvalg *s* på *n* = 300 bedrifter. Dvs., studievariabelen er: y_i = omsetning for bedrift *i,* i millioner kroner.

Estimatoren er den gjennomsnittlige omsetningen i utvalget multiplisert med antallet i denne populasjonen, dvs., ekspansjonsestimatoren $\hat{t}_e = 2830 \cdot \bar{y}_s$.
Resultater ved trekking av ett enkelt tilfeldig utvalg:

- Estimat: 275 625 (= $2830 \cdot \bar{y}_s$)
- *SE* = standardfeil for $2830 \cdot \bar{y}_s$ i ETU = 45 030
- Feilmarginen = $2 \cdot SE$ = 90 060
- 95 % konfidensintervall: estimat ± $2 \cdot SE$ = 185 565 – 365 685
  (For enkelthets skyld bruker vi tallet 2 istedenfor 1,96. Uansett er dekningsgraden kun tilnærmet lik 0,95 for dette konfidensintervallet. Hvis utvalgsgjennomsnittet var eksakt normalfordelt ville dekningsgraden ha vært 0,9545 med 2*SE*.)

Hvis vi hadde trukket ut andre bedrifter ville vi fått et annet estimat og konfidensintervall. For å illustrere dette har vi trukket ni nye utvalg. Det gir oss muligheten til å beregne *empirisk* varians og standardfeil basert på de ti estimatene på følgende måte:

Estimater : $\hat{t}_1, \hat{t}_2, ..., \hat{t}_{10}$ med gjennomsnitt $\bar{\hat{t}}$. Empirisk varians er da

$$s_t^2 = \tfrac{1}{9} \sum_{i=1}^{10} (\hat{t}_i - \bar{\hat{t}})^2$$

og empirisk standardfeil til estimatoren er: $s_t = \sqrt{s_t^2}$.

Tabell 7.1 Ti estimater for total omsetning med tilhørende 95 % konfidensgrenser

Utvalg	Nedre konfidensgrense	Estimat	Øvre konfidensgrense	Feilmargin
1	185 565	275 625	365 685	90 060
2	132 178	203 334	274 490	71 156
3	112 308	274 913	437 518	162 605
4	147 309	226 218	305 127	78 909
5	161 928	276 099	390 270	114 171
6	81 731	157 519	233 308	75 789
7	134 092	218 248	302 404	84 156
8	56 583	326 655	596 726	270 071
9	118 244	183 039	247 834	64 795
10	152 522	225 269	298 017	72 748
Gjennomsnittlig estimat		236 692		
Empirisk standardfeil av 10 estimater		51 059		

Tolkning:

- Estimatoren er forventningsrett, så gjennomsnittet av estimatene ved mange gjentatte utvalg skal være nær sann verdi. Her er gjennomsnittet 236 692 langt under populasjonsverdien. Som vi ser er feilmarginnen store så med ti utvalg er det stor sjanse for å få slik gjennomsnittlig underestimering.
- 20 % (2 av 10) konfidensintervaller dekker ikke *t.* Ved mange gjentatte utvalg vil denne andelen være ca. 5 %.
- Variasjonen i feilmarginen (= $2 \cdot SE$) indikerer at det er stor usikkerhet i *variansestimatoren* for estimatoren.
- Empirisk standardfeil gir en bedre indikasjon på usikkerheten i estimatoren.
- Gjennomsnittlig *SE* basert på feilmarginene er 54 223 som er ganske lik empirisk standardfeil.

Eksempel: Industri bedrifter med stratifisert utvalg

En alternativ utvalgsplan er å fordele utvalget utover populasjonen ved stratifisering. Når populasjonen består av bedrifter er det vanlig å stratifisere etter

- Næringsgruppering
- Sysselsettingsgrupper (størrelse)
- Omsetningsgrupper (størrelse)
- Geografi

De fleste begunnelsene for stratifisering i bedriftsundersøkelser er de samme som for person- og husholdningsundersøkelser:

- Sikre data til å lage detaljert statistikk for publisering
- La strata bestå av homogene grupper og dermed redusere usikkerheten i estimeringen (uten å øke utvalget)
- Alle grupper i populasjonen er representert
- Ta hensyn til at i noen grupper er det større variasjon enn i andre grupper
- Forskjellig utforming av skjema eller spørsmål til de enkelte grupper
- Forskjellige metoder for innsamling av data

I eksempel med industridata fra delkapittel 7.1.1 så er det en mulighet å dele inn i tre strata, etter to-siffer næring:

- 10 = næring- og nytelsesmidler
- 20 = kjemikalier og kjemiske produkter
- 24 = metaller

Vi trekker et stratifisert enkelt tilfeldig utvalg på totalt $n = 300$ virksomheter med proporsjonal allokering, dvs. utvalgsandel er like for alle strata. Utvalgsplanen blir da:

Bedrifter	*N*	*N*	Utvalgsandel
Næring 10	2358	250	10,6 prosent
Næring 20	295	31	10,6 prosent
Næring 24	177	19	10,6 prosent

Det ble trukket ti stratifiserte utvalg og resultatene er gitt i tabell 7.2.

Tabell 7.2 Ti estimater med tilhørende 95 % konfidensgrenser for stratifisert utvalg

Utvalg	Nedre konf. grense	Estimat	Øvre konf. grense	Feilmargin
1	138 869	191 214	243 560	52 346
2	158 600	254 586	350 572	95 986
3	154 749	261 362	367 975	106 613
4	181 827	283 629	385 430	101 801
5	167 836	265 536	363 236	97 700
6	117 690	176 032	234 372	58 340
7	67 208	318 709	570 209	251 500
8	20 405	268 397	516 388	247 991
9	135 808	294 620	453 432	158 812
10	207 662	330 843	454 024	123 181
Gjennomsnittlig estimat		264 493		
Empirisk standardfeil av 10 estimater		49 368		

Tolkning:

- Estimatoren er forventningsrett, så gjennomsnittet av estimatene ved mange gjentatte utvalg skal være nær sann verdi. Her er gjennomsnittet 264 493,så stratifisert estimator treffer mye bedre enn estimator fra enkelt tilfeldig utvalg i delkapittel 71.1.
- Som i enkelt tilfeldig utvalg:
  - 20 % (2 av 10) av konfidensintervallene dekker ikke *t.* Ved mange gjentatte utvalg vil denne andelen være ca. 5 %.
  - Variasjonen i feilmarginen (= $2 \cdot SE$) indikerer at det er stor usikkerhet i *variansestimatoren* for estimatoren.
- Empirisk standardfeil gir en bedre indikasjon på usikkerheten i estimatoren, ikke mye forskjell fra enkelt tilfeldig utvalg.

Denne stratifiseringen hjalp noe, men kan forbedres. Noe vi skal se på i neste delkapittel.

Eksempel: Industri bedrifter med stratifisering og fulltelling av de største bedriftene

Vi oppnår en bedre stratifisering med å stratifisere på tre-siffer nivå. Samtidig tar vi fulltelling av de største bedriftene. Det gir følgende utvalgsplan:

- Stratifiserer på 3 siffer nivå
- Samtidig fulltelling av alle bedrifter med over 100 sysselsatte, i alt 154 bedrifter
  - Står for 61 % av total omsetning: 164 097
- Tar et 11 % stratifisert utvalg av den resterende populasjonen på 2830-154 = 2676 bedrifter. Blir et stratifisert utvalg på 299 bedrifter, små og mellomstore, slik at total utvalgsstørrelse blir 453.
- Velger proporsjonal allokering

Utvalgsplanen for mindre og mellomstor bedrifter er gitt i tabell 7.3.

Tabell 7.3 Utvalgsplan for restpopulasjon av mindre og mellomstore enheter

Bedrifter	*N*	*n*
Stratum 101 produksjon av kjøtt og kjøttvarer	111	13
Stratum 102 bearbeid. og konserv. av fisk, skalldyr og bløtdyr	645	71
Stratum 103 bearbeiding og konserv. av frukt og grønnsaker	395	44
Stratum 104 prod. vegetabilske, animalske oljer og fettstoffer	1118	123
Stratum 201 prod. av kjemiske råvarer, gjødsel, nitrogen-forbindelser, basisplast og syntetisk gummi	20	3
Stratum 202 produksjon av plantevern- og skadedyrmidler	51	6
Stratum 203 produksjon av maling, lakk, tetningsmidler	24	3
Stratum 204 produk. såpe og vaskemidler, parfyme, toalettart.	170	19
Stratum 241 produksjon av jern og stål, samt ferrolegeringer	15	2
Stratum 242 produksjon av andre rør og rørdeler av stål	29	4
Stratum 243 annen bearbeiding av jern og stål	27	3
Stratum 244 produksjon av ikke-jernholdige metaller	71	8

Det ble trukket 10 utvalg fra «restpopulasjonen» på 2676 bedrifter og resultatene er gitt i tabell 7.4.

Tabell 7.4. Ti estimater; sann verdi i utvalgsdelen er 103 853.

Utvalg	Estimat utvalg	Total estimat	Feilmargin
1	131 049	295 146	41 016
2	78 697	242 794	15 728
3	99 721	263 818	22 171
4	103 040	267 137	35 022
5	100 823	264 920	25 743
6	112 908	277 005	31 317
7	97 497	261 594	21 801
8	97 365	261 462	19 752
9	115 563	279 660	27 818
10	140 171	304 268	46 184
Gjennomsnittlig estimat		271 780	
Empirisk standardfeil av 10 estimater		17 868	

Tolkning:

- Estimatoren er forventningsrett, så gjennomsnittet av estimatene ved mange gjentatte utvalg skal være nær sann verdi. Her er gjennomsnittet 271 780, så den treffer bra.
- Empirisk standardfeil sier også at usikkerheten i estimatoren er mye mindre, denne estimatoren treffer mye bedre.
  - Gjennomsnittlig *SE* basert på feilmarginene er 14 328 som er ganske lik empirisk standardfeil.
- 10 % (1 av 10) av konfidensintervallene dekker ikke *t*. Ved mange gjentatte utvalg vil denne andelen være ca. 5 %.
- Variasjonen i feilmarginen (= $2 \cdot SE$) indikerer at usikkerheten i *variansestimatoren* for estimatoren er mye mindre nå enn i ren stratifisering eller enkelt tilfeldig utvalg.
- Konklusjon: Det er veldig viktig med fulltelling av store enheter.

Kommentarer:

- Utvalg og estimering er ikke helt sammenlignbart med de to andre estimeringsmetodene i delkapitlene 7.1.1 og 7.1.2 siden utvalget er 50 % større.
- Hvis utvalgene i 7.1.1 og 7.1.2 hadde vært 50 % større så ville usikkerheten målt ved *SE* ha blitt redusert med en faktor på ca. 20 prosent, dvs. til *SE* = 40 000 omtrent og fremdeles er det stor forskjell.

En illustrasjon på hvor estimatene fordeler seg for de tre utvalgsplanene vi har sett på:

ETU estimator:

Stratifisert estimator:

Fulltelling av store enheter + stratifisering:

Igjen, la oss understreke noen selvsagte momenter. I en realistisk situasjon trekker vi bare ett utvalg og populasjonstotalen *t* er ukjent. Vi kan derfor ikke sammenligne den estimerte verdien med det faktiske populasjonstotalen slik vi har gjort i eksemplene. Vi må bruke standardfeilen/feilmarginen for å si noe om usikkerhet.

7.2. Utvalgsplan og allokering for bedriftsundersøkelser

I person- og husholdningsundersøkelser er det viktig med et representativt utvalg. Derfor brukes ofte proporsjonal allokering hvor utvalgsandelen er lik for alle strataene.

I bedriftsundersøkelser er det viktigst med høy dekningsgrad. Derfor er følgende viktig som vi har sett:

- Fulltelling av store enheter, definert vanligvis som bedrifter med minst 100 ansatte
- For mellomstore enheter: Stratifisert utvalg etter næringsgruppe og sysselsettingsgruppe

I tillegg er det vanlig med et cut-off: Små enheter som bidrar lite til totalen holdes utenfor, vanligvis definert som bedrifter med mindre enn 10 ansatte.
Når det gjelder allokering i det stratifiserte utvalget av mellomstore bedrifter er det flere muligheter:

- En mulighet er *proporsjonal allokering*, som i eksempel, men det er ikke mye brukt.
- Kan vurdere *optimal allokering*; fordele *n* slik at variansen til totalestimatoren blir minst mulig:
  - Som i person-husholdningsundersøkelserløsning så avhenger det av valg av variable og ukjente stratumvarianser for den valgte variabelen. Man må i såfall velge et par hovedvariabler og anslå stratumvarianser fra tidligere undersøkelse.
- *Den vanligste allokeringen, for å øke dekningsgraden*: Allokeringsrutiner som gir størst trekkesannsynlighet til store enheter
  - Deler ofte utvalget i næringsgruppe og antall sysselsatte, dvs. innen hver næringsgruppe- stratifiserer etter sysselsettingsgruppe, for eksempel 10-19, 20-49 og 50-99 og lar utvalgsandelen innen hvert stratum øke med antall ansatte.

To viktige temaer i bedriftsundersøkelser er editering og oppgavebyrde.
I person- og husholdningsundersøkelser så er den største feilkilden frafall. I økonomiske undersøkelser er det oppgaveplikt og frafallet er lite, vanligvis 5 -10 %. Den største feilkilden er målefeil. For eksempel, 1000-feil er ganske vanlig, dvs. at måleenheten på verdier er i feil krone-enhet. Det er derfor med gode editeringsmetoder for å rette opp feil. Editering er et omfattende tema hvor viktige aspekter er automatiske kontroller, makro- og selektiv revisjon. SSB har lenge hatt et eget metodekurs i editering basert på en håndbok fra 2005.

I bedriftsundersøkelser er det viktig å redusere og fordele oppgavebyrden. Noen momenter er:

- Det er naturlig å rullere ut en større andel jo færre ansatte det er i strataene
- Der det er mulig skal en enhet delta i to år på rad i en strukturundersøkelse, og fire perioder på rad i en korttidsstatistikk
- Der det er mulig skal deltagelse i en undersøkelse etterfølges av en hvileperiode
- Det forventede forhold mellom antallet år av hvil og deltagelse er likt over tid for alle sammenlignbare enheter.

7.3. Bruk av stratifisert rate-estimator i SSBs ordrestatistikk i industrien

Vi skal se på bruk av rate-estimator i SSBs ordrestatistikk i industrien i 1. kvartal 2011. Populasjonen består av 5051 bedrifter innen næringene i tabell 7.5.

Tabell 7.5 Næringene i ordrestatistikken

Næring	Antall bedrifter
Tekstil- og bekledningsindustri	406
Papir- og papirvareindustri	72
Kjemisk og farmasøytisk industri	219
Metallindustri	132
Metallvareindustri	1 417
Data- og elektrisk utstyrsindustri	542
Maskinindustri	829
Bygging av skip og oljeplattformer	333
Annen verkstedsindustri	134
Maskinreparasjon og -installasjon	967
I alt	5 051

Utvalgsplanen for denne undersøkelsen:

- I alt 900 bedrifter = 18 % av populasjonen
- Totaltelling: bedrifter med minst 100 sysselsatte
- Ingen bedrifter med mindre enn 10 sysselsatte
- Bedrifter med antall sysselsatte 10 – 99 trekkes stratifisert etter næringsgruppe og sysselsetting
- Innenfor hver næring på 3 siffer nivå (bearbeidingsnivå, som i tabell 7.5) stratifiser etter sysselsettingsgruppene 10-19, 20-49 og 50-99.
- Innenfor hvert stratum: Enkelt tilfeldig utvalg, andel i utvalgsstratumet øker med antall ansatte.

Utvalgsplanen sikrer at vi får med mange bedrifter med stor omsetning i utvalget. I 1. kvartal 2011 var dekningsgraden 78 % når det gjaldt omsetning.
Det er to hovedvariable, ordretilgang og ordrereserve, fordelt på hjemmemarkedet og eksportmarkedet. Det blir i alt seks analysevariable når vi tar med total ordretilgang og total ordrereserve. Den presise definisjonen av disse variablene er:

- Ordretilgang = verdien av alle ordrer og bestillinger en bedrift mottar i et gitt kvartal, unntatt ordrer og bestillinger på handelsvarer.
- Ordrereserven = verdien av bedriftens påbegynte og ikke påbegynte ordre, målt ved gjeldende kvartals utløp.

Det publiseres tall på næringsgruppene i tabell 7.5 samt kjemiske råvarer og ikke-jernholdige metaller.

Den stratifiserte rate-estimatoren
Innenfor hvert stratum brukes en rate-estimator med kvartalsvis omsetning som forklaringsvariabel. Omsetningen er en fjerdedel av total omsetning året før hentet fra VoF, Virksomhets- og Foretaksregisteret. Dvs., innenfor et gitt stratum h så estimeres den totale ordretilgangen/ordrereserven med $b_h X_h$ hvor

$$b_h = \frac{\text{sum rapport ordretilgang/ordrereserve i stratumutvalget}}{\text{sum omsetning i utvalgsstratumet}}$$

og X_h = sum omsetning i hele stratumet i populasjonen.

Med y_i = ordretilgang/ordrereserve for bedrift i og x_i = omsetning for bedrift i (=1/4 av fjorårets omsetning), så er stratumtotalen t_h for y – variabelen estimert med

$$\hat{t}_{h,R} = X_h \cdot \frac{\sum_{i \in s_h} y_i}{\sum_{i \in s_h} x_i}, \text{ hvor } s_h \text{ er utvalget fra stratum } h.$$

Hvis det er fulltellingsstratum, så er $X_h = \sum_{i \in s_h} x_i$, og estimatet blir $\hat{t}_{h,R} = \sum_{i \in s_h} y_i = t_h$, stratumtotalen.

Estimatet for totalen t er da: $\hat{t} = \sum_h \hat{t}_h$. For et gitt publiseringsnivå er estimatet $\hat{t}_{publ}$ = sum av $\hat{t}_h$ over de strataene som inngår i publiseringsnivået. Denne estimeringsmetoden er basert på en såkalt *ratemodell* innen hvert stratum.

Stratifisert ratemodell
y_i´ene antas å være verdier av stokastiske variable Y_i, og innenfor hvert stratum en egen ratemodell:
$E(Y_i) = \beta_h x_i$ og $Var(Y_i) = \sigma_h^2 x_i$.

For denne undersøkelsen passer denne ratemodellen rimelig bra, litt bedre for ordretilgang enn ordrereserver. Noen observasjoner er «avvikere» fra den antatte modellen samtidig som de har stor innflytelse på estimeringen; de fjernes fra estimeringen og representerer kun seg selv. I kapittel 8 skal vi betrakte modellbasert statistisk inferens i utvalgsundersøkelser generelt.

Strataene defineres ved bearbeidingsnivå, en del bearbeidingsnivå blir slått sammen for å få nok utvalgsenheter i strataene for estimeringen. Vi betrakter modellbasert varians som vi skal se på i detalj i kapittel 8, $Var(\hat{T} - T)$. Usikkerheten basert på estimert modell-varians og variasjonskoeffisienten, CV = standardfeil/estimat, er for 2011 1. kvartal mellom 0,8 og 6,8 prosent for næringene som det publiseres for. For total ordretilgang så er den 1,1 prosent. I tabell 7.6 presenteres resultatene for ordretilgang på næringsnivå. Vi har tatt med tilsvarende estimater for 2015 2. kvartal.

Tabell 7.6. Estimater-ordretilgang, hjemme- og eksportmarked i alt (2005 =100)

Næring	Ordretilgang 2011K1	Ordretilgang 2015K2
Tekstil- og bekledningsindustri	145,3	124,2
Papir- og papirvareindustri	83,6	53,1
Kjemisk og farmasøytisk industri (i alt)	152,0	177,8
Kjemiske råvarer	153,5	149,0
Metallindustri (i alt)	131,7	99,9
Ikke-jernholdige metaller	128,9	94,6
Metallvareindustri	151,5	142,0
Data- og elektrisk utstyrsindustri	137,2	139,3
Maskinindustri	159,9	135,5
Bygging av skip og oljeplattformer	115,2	108,7
Annen verkstedsindustri	158,5	180,7
Maskinreparasjon og -installasjon	130,9	125,4
Totalt	*134,4*	*125,2*

Fra 2011 til 2015 ser vi at det noen næringer som har store endringer. For eksempel, sterk nedgang i papir- og papirvareindustrien, metallindustrien og maskinindustrien, og stor oppgang i kjemisk og farmasøytisk industri og annen verkstedsindustri.

8. Alternative tilnærminger for statistisk inferens basert på utvalgsundersøkelser

Dette kapitlet inneholder følgende temaer:

- Alternative tilnærminger for statistisk inferens basert på utvalgsundersøkelser
- Vitenskapelige problemer med design-basert inferens
- Likelihoodprinsippet
- Likelihoodfunksjonen i design-basert inferens

8.1. Alternative tilnærminger

Til nå har vi betraktet en design-basert tilnærming:

- Ingen modellering, eneste stokastiske element er utvalget *s* med *kjent fordeling.*
- Utvalgsvariansen sier hvor mye estimatoren varierer når utvalget trekkes gjentatte ganger *på samme tid.*

En modell-basert tilnærming betyr at verdiene y_i i populasjonen antas å være verdier av tilfeldige variable Y_i, og at det antas en (populasjons-) modell. Et eksempel er stratifisert ratemodell for ordrestatistikken i delkapittel 7.3. I modell-basert analyse så vil

- modellvariansen fortelle oss hvor mye estimatoren varierer ved gjentakelser av ***Y*** for en *gitt* utvalg**.** Modellvariansen er et mål for usikkerhet for det utvalget *s som er valgt.*

Statistiske inferensprinsipper impliserer at den modell-baserte tilnærmingen er den mest «trygge» og vitenskapelig gyldige måten å foreta statistisk analyse i utvalgsundersøkelser. I SSB og de fleste nasjonale statistikkbyråer brukes det *design-baserte opplegget* i de fleste undersøkelser. Den er den mest *anvendte* måten å behandle analyser i utvalgsundersøkelser. Design-basert utvalgsteori skiller den fra vanlig statistisk analyse. Denne tradisjonelle tilnærmingen har sitt utgangspunkt i en artikkel av Jerzy Neyman. På en måte er den en slags "minste motstands vei" i den forstand at vi trenger ikke å jobbe med den vanlige starten for statistiske analyser, modellering av data.

Vitenskapelige problemer med design-basert inferens

Generelt så er design-basert inferens relatert til *hypotetiske* gjentagelser av trekking av utvalg for en *fast* populasjonsvektor **y.** Det betyr at variansestimater kan gi feilaktig informasjon for et *gitt utvalg.* Et aspekt ved måling av usikkerhet for korttidsstatistikken er:

- Hvis vi ønsker å måle hvor godt en estimeringsmetode virker i kvartalsvis eller månedlige undersøkelser, så vil **y** variere fra kvartal til kvartal eller måned til måned
  - trenger å anta at **y** er en realisering av tilfeldig vektor.

Vi skal bruke *likelihood* og *likelihoodprinsippet* som rettledning for hvordan disse temaene skal behandles. Men først nå skal vi illustrere problemer som kan oppstå ved bruk av utvalgsbasert varians.

Problem med design-basert variansmål. Illustrasjon 1

Utvalgsplanen er slik at med sannsynlighet ½ så velges fulltelling av hele populasjonen, og med sannsynlighet ½ så velges tilfeldig *en* enhet. Det betyr at det er $N + 1$ mulige utvalg: {1}, {2},…,{N}, {1,2,…, N}, og utvalgsplanen er gitt ved:

- $p(\{i\}) = 1/2N$, for $i = 1,..,N$; $p(\{1,2,\ldots N\}) = ½$.

Vi bruker utvalgsgjennomsnittet $\bar{y}_s$ som estimator for populasjonsmiddel μ. Den er forventningsrett:

$$E(\bar{y}_s = \sum_s p(s) \cdot \bar{y}_s = \sum_{i=1}^{N} \tfrac{1}{2N} \cdot y_i + \tfrac{1}{2} \cdot \mu = \mu$$

La $\tilde{\sigma}^2 = \sum_{i=1}^{N} (y_i - \mu)^2 / N$. Design-variansen blir:

$$Var(\bar{y}_s) = E(\bar{y}_s - \mu)^2 = \sum_{i=1}^{N} (y_i - \mu)^2 \cdot \tfrac{1}{2N} = \tfrac{1}{2} \cdot \tilde{\sigma}^2 .$$

Anta nå at vi trekker «utvalget» {1, 2, …, N}, dvs., fulltelling. Da påstår vi at «presisjonen» for det estimatet vi fikk (som vi vet er feilfritt) er lik $\tilde{\sigma}^2 / 2$.

Problem med design-basert variansmål. Illustrasjon 2

Anta vi har to eksperter til å utføre en undersøkelse. Ekspert trekker et vanlig enkelt tilfeldig tuvalg, mens ekspert 2 trekker et enkelt tilfeldig utvalg med tilbakelegging. Anta at begge eksperter velger samme utvalg og beregner samme estimat. Da har vi følgende situasjon:

- Ekspert 1: Enkelt tilfeldig utvalg og estimat $\bar{y}_s$. Presisjon er målt ved

$$(1-f)\frac{\sigma^2}{n}$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2, f = n/N$$

- Ekspert 2: Enkelt tilfeldig utvalg med tilbakelegging og estimat $\bar{y}_s$, måler presisjon med $\tilde{\sigma}^2/n$.
- Begge eksperter velger *samme* utvalg, beregner det *samme* estimatet, men gir *forskjellige* mål på presisjon…

8.2. Likelihood og likelihoodprinsippet (LP), generell modell

Modell: Data y er verdi av stokastisk variabel Y med statistisk modell $f_\theta(y), \theta \in \Omega$; θ er de ukjente parametrene i modellen.

Eksempel:
Anta vi trekker tilfeldig 100 personer og spør om de er arbeidsledig. La θ være sannsynligheten for å være arbeidsledig. La Y være antall arbeidsledige i utvalget. Da er

$$f_\theta(y) = P(Y=y) = C_y \cdot \theta^y (1-\theta)^{100-y}$$

hvor C_y er alle mulige måter å få akkurat y ledige.

Likelihoodfunksjonen er definert, med *data* y: $l_y(\theta) = f_\theta(y)$.

Denne funksjonen av den ukjente parameteren θ måler "likelihood" (har ikke noe godt norsk ord) for forskjellige θ-verdier i lys av data.

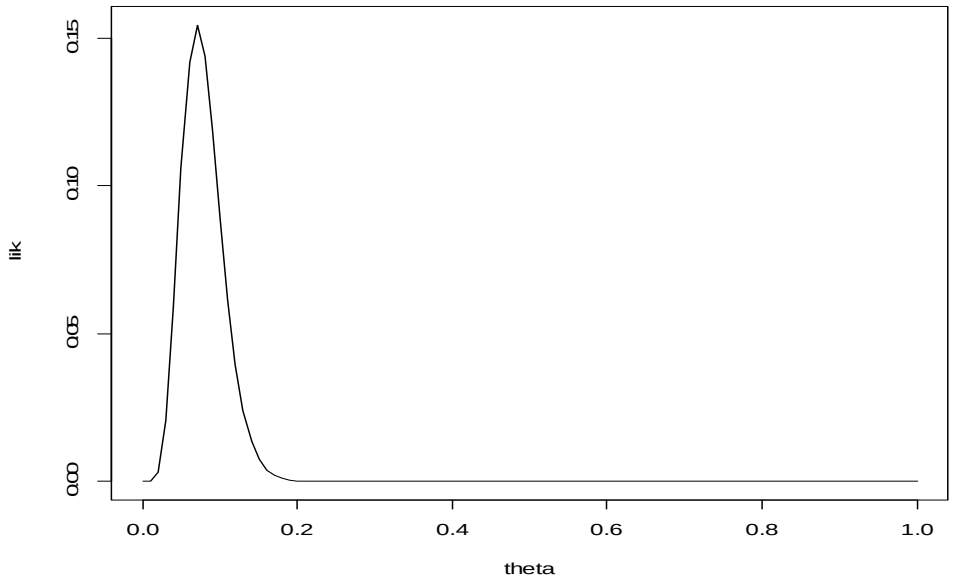
Likelihoodprinsippet (*LP*) sier: Likelihoodfunksjonen inneholder all informasjon om de ukjente parametrene θ.

I eksemplet, anta vi observerer y =7 arbeidsledige blant de 100. Da er

$$l(\theta) = C_7 \cdot \theta^7 (1-\theta)^{93} \quad \text{hvor } C_7 = 1{,}6 \cdot 10^{10}.$$

Figur 8.1 viser en graf av likelihoodfunksjonen.

Figur 8.1 Likelihoodfunksjonen for theta (andel arbeidsledige) for 7 av 100 arbeidsledige i utvalget



Maksimumspunktet er for theta = 7/100 = 0,07. Denne verdien kalles maximum-likelihood estimatet.
*Utledning: Maksimer log likelihood:

$$\log l = \log(C) + 7\log(\theta) + 93\log(1-\theta).$$

Deriverer med hensyn på θ og løser ligningen
$\partial \log l / \partial \theta = 0$:

$$\frac{7}{\theta} - \frac{93}{1-\theta} = 0 \Leftrightarrow \frac{7}{\theta} = \frac{93}{1-\theta} \Leftrightarrow 7(1-\theta) = 93\theta \Leftrightarrow \theta = 7/100.$$

Noen viktige aspekter ved likelihoodprinsippet:

- Vanlig "frekventistisk" tilnærming: Statistiske metoder evalueres pre-eksperimentelt. Dvs., hvordan metoder gjør det i det lange løp, ikke nødvendigvis hvordan metoden fungerer for akkurat de *data* vi jobber med her og nå.
- *LP* evaluerer statistisk metode post-eksperimentelt, gitt data. Dvs., ser på informasjonen vi har akkurat for de dataene vi har her og nå.
- *LP* er kontroversielt, men vanskelig å argumentere mot på grunn av et fundamentalt av Birnbaum i 1962:
  - *LP* følger fra suffisiens- (SP) og betingings-prinsipper (CP) som "ingen" er uenig i.
  - SP: Statistisk inferens skal baseres på suffisiente observatorer
  - BP: Hvis du har to mulige undersøkelser og velger en av de tilfeldig, så skal inferensen kun avhenge av den valgte undersøkelsen.

Illustrasjon av BP
Et valg skal tas mellom en fulltelling eller ta et utvalg på størrelse 1. Hver med sannsynlighet ½.
La oss si at fulltelling er valgt. En ubetinget vurdering gir følgende trekkesannsynligheter:

$\pi_i = P$(fulltelling) + P(utvalg av størrelse 1og enhet i velges)
= ½ + P(utvalg av størrelse 1)P(enhet i velges | utvalg av størrelse 1) =1/2+(1/2)·(1/N) ≈1/2.

Horvitz-Thompson estimatoren blir: $\hat{t}_{HT} \approx 2\sum_{i=1}^{N} y_i = 2t$!

Med betinget opplegg så er $\pi_i = 1$ og H-T estimatet er t.

8.3. Likelihoodfunksjon og likelihoodprinsippet i design-basert inferens

Vi trenger å «oversette» den generelle situasjonen i delkapittel 8.3 til utvalgsundersøkelser:

- Ukjent parameter θ: $\mathbf{y} = (y_1, \ldots, y_N)$
- Likelihoodfunksjon = sannsynligheten for data, betraktet som funksjon av parametrene
- La nå x betegne data. Da er x utvalget s med tilhørende y-verdier
- Likelihoodfunksjonen er sannsynligheten for data x med utvalgsplan $p(\cdot)$, som funksjon av $\mathbf{y}$:
  - $l_x(\mathbf{y}) = p(s)$, for alle $\mathbf{y}$ som er mulige.

Dette betyr at likelihoodfunksjonen er konstant for alle mulige verdier av $\mathbf{y}$. Dvs., alle mulige verdier av $\mathbf{y}$ er like "sannsynlige".
Likelihoodprinsippet (*LP*) sier at likelihoodfunksjonen inneholder all informasjon om de ukjente parametrene.
Det gir følgende implikasjoner ifølge *LP*:

- Design-modellen er slik at data inneholder ingen informasjon om den ikke-observerte delen av $\mathbf{y}$, $\mathbf{y}_{\text{unobs}}$.
- Må anta på forhånd at det er en relasjon mellom data og $\mathbf{y}_{\text{unobs}}$:
  - Som en konsekvens av *LP*: Det er nødvendig å anta en statistisk modell.
- Utvalgsplanen er irrelevant for den statistiske analysen, fordi to utvalgsplaner som leder til samme s vil ha proporsjonal likelihood.
- Samme inferens med de to forskjellige utvalgsplaner som resulterer i samme utvalg s. Dette er direkte i opposisjon til vanlig design-basert inferens, hvor den eneste stokastiske evalueringen er via utvalgsplanen, for eksempel Horvitz-Thompson estimatoren.
- Begreper som design forventingsrett og designvarians er irrelevante ifølge *LP* når det dreier seg om den faktiske statistiske analysen.

Det er viktig å merke seg to sentrale momenter i anvendelse av likelihoodbetraktninger i design-basert utvalgsteori:

- *LP* ikke dreier seg ikke om metodevurdering, men den statistiske analysen *etter* at data har blitt observert.
- Likelihoodvurderingen *betyr ikke* at utvalgsplanen ikke er viktig. Det er viktig å forsikre oss om at vi får et godt representativt utvalg. Men når utvalgsdata er blitt samlet inn så skal ikke utvalgsplanen spille noen rolle i inferensfasen, ifølge *LP*.

9. Modell-basert statistisk inferens i utvalgsundersøkelser

Dette kapitlet inneholder følgende temaer:

- Modell-baserte estimatorer, ratemodellen
- Varianstolkninger; utvalgsbasert, modell-basert og metodebasert
- Robust variansestimering

9.1. Modell-basert tilnærming

- Antar en modell for y vektoren,
  $y_1, y_2, \ldots, y_N$ er realiserte verdier av tilfeldige variable $Y_1, Y_2, \ldots, Y_N$

Vi bruker modellen til å konstruere estimator, for eksempel rate-estimatoren

- Modellbasert inferens/analyse:
  - Inferens er basert på antatt modell
  - Behandler utvalget *s* som gitt, betinger på det faktiske utvalget
- Optimale estimatorer kan utledes: Beste lineære forventningsrette estimatorer
- Forskjellige variansmål.
- Som i all statistisk modellering så er analysen avhengig av valgt modell:
  - Introduserer et subjektivt element
  - nesten umulig å modellere alle variablene i en undersøkelse
  - Modellerer hovedvariablene

Vi kan dekomponere totalen *t* som følger:

$$t = \sum_{i=1}^{N} y_i = \sum_{i \in s} y_i + \sum_{i \notin s} y_i \ .$$

Siden $\sum_{i \in s} y_i$ er kjent, problemet er å estimere $z = \sum_{i \notin s} y_i$, realisert verdi av $Z = \sum_{i \notin s} Y_i$.

Det kan gjøres ved å estimere hver uobservert y_i: $\hat{y}_i, i \notin s.$ Estimator blir da:

$$\hat{t}_{pred} = \sum_{i \in s} y_i + \sum_{i \notin s} \hat{y}_i = \sum_{i \in s} y_i + \hat{z}$$

hvor $\hat{z}$ er en estimator for *z*.

Kommentarer:

1. Alle estimatorer kan beskrives på «prediksjonsform»:

$$\hat{t} = \sum_{i \in s} y_i + \hat{z}_t$$

$$\text{ved å la } \hat{z}_t = \hat{t} - \sum_{i \in s} y_i$$

2. Kan bruke denne formen for å se om estimatoren gir mening.

Eksempel

$$\hat{t}_e = N\bar{y}_s = \sum_{i \in s} y_i + (N-n)\bar{y}_s = \sum_{i \in s} y_i + \sum_{i \notin s} \bar{y}_s$$

$$\text{Derav}: \hat{z} = \sum_{i \notin s} \bar{y}_s \text{ og } \hat{y}_i = \bar{y}_s, \text{ for alle } i \notin s.$$

Vi kan nevne to vanlige modeller.

I. Ratemodellen, spesielt for bedriftsundersøkelser:

Anta tilleggsvariabel *x* er tilgjengelig alle enheter i populasjonen. Modellen er:

$Y_i = \beta x_i + \varepsilon_i$, for $i = 1, \ldots, N$, hvor feilleddet varierer rundt 0 med varians $\sigma^2 x_i$.

Variabelen x_i er vanligvis et mål på "størrelsen" til enhet *i*, og y_i øker med økende x_i. I bedriftsundersøkelser så kan *x* være antall ansatte, og regresjonen går gjennom origo i mange tilfeller.

II. Enkel middel modell:

$Y_i = \beta + + \varepsilon_i$, hvor feilleddet har varians σ^2.

9.2. Modellbaserte optimale estimatorer

En estimator kan vi nå generelt uttrykke på følgende måte:

$$\hat{T} = \sum_{i \in s} Y_i + \hat{Z} .$$

La θ betegne modell-parametrene. Definisjon av forventningsrette estimatorer og varians i en modell-basert tilnærming er følgende:

- $\hat{T}$ er modell-forventningsrett hvis $E_\theta(\hat{T} - T) = 0, \forall \theta, T = \sum_{i=1}^{N} Y_i$
- Modellvarians er variansen til *estimeringsfeilen,* også kalt *prediksjonsvariansen*

$$Var_\theta(\hat{T} - T) .$$

Prediksjonsvariansen er et variansmål for det faktisk observerte utvalget. For de to illustrasjonene i delkapittel 8.2 så gir prediksjonsvariansen det «riktige» svar.

Illustrasjon 1

$N + 1$ mulige utvalg: {1}, {2}, …, {N}, {1, 2, …, N}. Vi bruker $\hat{T} = N\overline{Y}_s$ som estimator for totalen T. Anta at «utvalget» blir {1, 2, …, N}. Da er $\hat{T} = N\overline{Y} = T$, og prediksjonsvariansen blir $Var(\hat{T} - T) = Var(0) = 0,$ som den bør være for dette datasettet.

Illustrasjon 2. Eksakt samme prediksjonsvarians for de to utvalgsplanene.

Optimale lineære estimatorer

Vi skal betrakte *lineære* estimator, dvs. estimatoren er en lineær kombinasjon av *y-v*erdiene i utvalget:

$$\hat{T} = \sum_{i \in s} a_i(s) Y_i .$$

BLU (best linear unbiased) estimator:

$\hat{T}_0$ er beste lineære forventningsrette (engelsk: BLU) estimator for T hvis

1) $\hat{T}_0$ er modell-forventningsrett
2) $\hat{T}_0$ har uniformt minimum prediksjonsvarians blant alle modell- forventningsrette estimatorer:

   For alle modell-forventningsrette lineære estimatorer $\hat{T}$, $Var_\theta(\hat{T}_0 - T) \le Var_\theta(\hat{T} - T)$ for alle θ.

Vi skal nå se på BLU estimatorene i ratemodellen og middel modell.

Optimalitet i ratemodellen, $E(Y_i) = \beta x_i$ og $Var(Y_i) = \sigma^2 x_i$:

BLU estimator:

$$\hat{T}_R = \sum_{i \in s} Y_i + \sum_{i \notin s} \hat{\beta}_R x_i$$

hvor $\hat{\beta}_R$ er den beste lineære forventningsrette estimator (BLUE) av β.

$$\hat{\beta}_R = \sum_{i \in s} Y_i / \sum_{i \in s} x_i = \hat{R},$$ den vanlige utvalgsraten.

Vi ser at

$$\hat{T}_R = \sum_{i \in s} Y_i + \sum_{i \notin s} \hat{R} x_i = \hat{R} \sum_{i \in s} x_i + \hat{R} \sum_{i \notin s} x_i = \hat{R} \cdot t_x,$$

den vanlige rate-estimatoren.

Estimert prediksjonsvarians til BLU estimator:

$$\hat{V}(\hat{T}_R - T) = N^2 \frac{1-f}{n} \cdot \frac{\overline{x}_r \overline{x}}{\overline{x}_s} \hat{\sigma}^2,$$

hvor $f = n/N, \overline{x}_r = \sum_{i \notin s} x_i / (N-n)$ og $\overline{x} = \sum_{i=1}^{N} x_i / N$

og

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i \in s} \frac{1}{x_i} (Y_i - \hat{R} x_i)^2 .$$

Tilnærmet 95 % konfidensintervall for store n, N-n for verdien t av T:

$$\hat{t}_R \pm 1{,}96 \sqrt{\hat{V}(\hat{T}_R - T)}$$

Dette kalles også et 95 % *prediksjonsintervall* for stokastisk variabel *T*.

Optimal estimator i enkel middel modell, $Y_i = \beta + \varepsilon_i, E(\varepsilon_i) = 0$ og $Var(\varepsilon_i) = \sigma^2$:

$$\hat{\beta}_{opt} = \overline{Y}_s, \text{ og } \hat{T}_{pred} = \sum_{i \in s} Y_i + \sum_{i \notin s} \overline{Y}_s = N \cdot \overline{Y}_s$$

$$Var(N \cdot \overline{Y}_s - T) = N^2 (1-f) \frac{\sigma^2}{n}.$$

Dette er også den vanlige, design-baserte variansformel under enkelt tilfeldig utvalg.
Vi ser at variansestimatet er gitt ved

$$N^2 (1-f) \frac{\hat{\sigma}^2}{n} \text{ med } \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \overline{y}_s)^2, \text{ utvalgsvariansen,}$$

eksakt det samme som det design-baserte variansestimat, men tolkningen er forskjellig.

Tilslutt i kapittel 9.2, noen kommentarer om robust variansestimering for modellbasert varians:

- Modellen er egentlig en "arbeidsmodell"
- Spesielt variansantakelsen kan være betydelig feil
  - som konstant varians
  - varians proporsjonal med et størrelsesmål x_i
    og det er ikke alltid så lett å oppdage en slik modellfeil.
- Standard minste kvadraters variansestimat er følsom for feil i variansantakelsen
- Det er mulig å konstruere robuste variansestimatorer.

I tillegg til design-basert og modell-basert varians så finnes det et tredje variansmål som kan benyttes ved gjentatte månedlige eller kvartalsvise undersøkelser. Det beskriver egenskapen ved selve metoden og vi kaller det metodevarians.

9.3. Metodevarians

Formålet er nå å definere et variansmål som sier noe om den forventede usikkerhet i *gjentatte* undersøkelser. Her skal vi både bruke både utvalgsplanen og at populasjonsvektoren **Y** antas å være stokastisk for å beskrive at **y**-verdiene endrer seg ved gjentatte undersøkelser.

1. Betinget på utvalget *s,* med modell-forventningsrett $\hat{T}$:
   $Var(\hat{T} - T)$ måler usikkerheten for *dette* spesielle utvalget *s*
2. Forventet usikkerhet for gjentatte undersøkelser måles ved
   $E_p\{Var(\hat{T} - T)\}$, over utvalgsfordelingen $p(\cdot)$.

Denne kalles *metodevariansen* (engelsk: *anticipated* (forventet) *variance*), og kan betraktes som et variansmål som beskriver hvordan estimerings*metoden* gjør det i gjentatte undersøkelser.

En anvendelse:
Enkel lineær regresjon og enkelt tilfeldig utvalg
Lineær modell: $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $Var(\varepsilon_i) = \sigma^2$.
Hvis $N \cdot \overline{Y}_s$ benyttes så kan det vises at metodevariansen er lik

$$N^2 \frac{1}{n}(1 - \frac{n}{N})(\sigma^2 + \beta_2^2 S_x^2), \quad S_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

BLU estimatoren er gitt ved

$$\hat{T} = N[\overline{Y}_s + \hat{\beta}_2 (\overline{x} - \overline{x}_s)]$$

hvor $\hat{\beta}_2$ er minste kvadraters estimator (BLUE) for β_2.
Metodevariansen for BLU estimatoren er tilnærmet lik

$$N^2 \frac{1}{n}(1 - \frac{n}{N})\sigma^2 .$$

Vi ser at $\hat{T}$ fjerner $\beta_2^2 S_x^2$ og er mye mer effisient enn $N \cdot \overline{Y}_s$.

Kommentarer:

- Fra et design-basert synspunkt så er utvalgsmiddel estimatoren forventningsrett, mens den lineære regresjonsestimatoren ikke er det
- Hvis kun *design-skjevhet* betraktes, så kunne vi valgt utvalgsmiddel estimatoren
- Den lineære regresjonsestimatoren vil kun velges framfor utvalgsmiddel estimatoren fordi den har mindre metodevarians
- Det er derfor vanskelig å se design - forventningsretthet som et kriterium for å velge estimatorer.

Øvelser for KLAR 311 Introduksjonskurs i statistiske metoder

Øvelse 1. Begrepsforståelse. En innføring i sannsynlighetsbegreper

Formålet med denne øvelsen som består av tre oppgaver, er å illustrere de viktigste begrepene i estimering fra enkelt tilfeldig utvalg:

- Sannsynlighet
- Forventning til en estimator
- Varians og standardfeil til en estimator
- Estimert varians og standardfeil til en estimator

Alle oppgavene er basert på følgende situasjon:
Vi har en populasjon på 5 personer, nummerert fra 1 til 5. Personene 1,3 og 4 er sysselsatte. De andre er arbeidsledige. Vi trekker et enkelt tilfeldig utvalg på 3 personer og lar X betegne antall sysselsatte i utvalget.

1.1 Sannsynlighetsberegninger

(a) List opp alle mulige utvalg. Hvor mange mulige utvalg er det?
(b) Hva er sannsynligheten for et vilkårlig utvalg på 3 personer?
(c) Hva er sannsynligheten for at person 1 er i utvalget?
(d) Hva er sannsynligheten for at personene 1 og 2 begge er i utvalget?
(e) Hva er sannsynligheten for at alle i utvalget er sysselsatt?

1.2 Forventning og varians

(a) Beregn sannsynlighetene for at X tar verdiene 0,1,2,3; $P(X = x)$ for $x = 0,1,2,3$.
(b) Beregn forventningen til X ved å bruke formelen:
  - Forventning er lik summen av verdier·sannsynlighet.
(c) Betrakt $\hat{p} = X/3$ som et estimat for andel sysselsatte i populasjonen og vis at $\hat{p}$ er en forventningsrett estimator for andel sysselsatte i populasjonen.
(d) Vis at variansen og standardfeilen til estimatoren er lik 0,04 og 0,20.

1.3 Tolkning av forventning, standardfeil og estimert standardfeil

En forventningsrett estimator for variansen til $\hat{p}$ er $\hat{V} = \hat{p}(1-\hat{p})/5$ slik at den estimerte standardfeilen er gitt ved $SE = \sqrt{\hat{p}(1-\hat{p})/5}$.

Vi skal nå illustrere og tolke begrepene forventning og standardfeil til estimatoren $\hat{p}$ ved å foreta mange trekkinger (simuleringer) av enkelt tilfeldig utvalg på 3 personer og beregne gjennomsnitt av estimatene, den empiriske standardfeilen og gjennomsnittet av de estimerte standardfeilene.

La b betegne antall simuleringer, og la $\hat{p}_1,...,\hat{p}_b$ være verdiene av $\hat{p}$. Da har vi:

- Gjennomsnitt av estimatene: $\overline{p} = \sum_{i=1}^{b} \hat{p}_i / b$
- Empirisk varians: $v = \frac{1}{b-1}\sum_{i=1}^{b}(\hat{p}_i - \overline{p})^2$
- Empiriske standardfeil: $\sqrt{v}$
- Gjennomsnitt av SE, $S\overline{E}$: $S\overline{E} = \sum_{i=1}^{b} SE_i / b$ hvor $SE_i = \sqrt{\hat{p}_i(1-\hat{p}_i)/5}, i = 1,...,b$,

Tolkning, som vi skal vise ved å se på verdier av b fra 5 til 5 millioner:
Når $b \to \infty$ så vil
$\overline{p} \to E(\hat{p}) = 0{,}6$ og $\sqrt{v} \to$ standardfeil = 0,2.

Vi vil også se at selv om $\hat{V}$ er forventningsrett for V= 0,04, så er ikke SE helt forventningsrett for standardfeilen 0,2.

```
R-program:
y=c(1,0,0,1,1,0)# Her indikerer 1 sysselsatt og 0 arbeidsledig
```

Følgende R-funksjon kan benyttes:

```
forv2se=function(b)
{estimat=numeric(b)
estse=numeric(b)
for(k in 1:b){
s=sample(5,3)
estimat[k]=sum(y[s])/3
estse[k]=sqrt(estimat[k]*(1-estimat[k])/5)
}
mean(estimat)
meanse=mean(estse)
se=sqrt(var(estimat))
list(mean=mean(estimat),se=se,meanse=meanse)
}
```

Her er:
mean(estimat) = $\bar{p}$

mean(estse) = gjennomsnitt av estimert standardfeil= $S\bar{E}$.

se = empirisk standardfeil= $\sqrt{v}$

Basert på antall simuleringene i tabellen nedenfor, beregn $\bar{p}, \sqrt{v}$ og $S\bar{E}$. Sett verdiene inn i tabell 1:

Tabell 1

Antall simuleringer av utvalg	Gjennomsnitt av estimat	Empirisk standardfeil	Gjennomsnitt av SE, $S\bar{E}$
5			
10			
50			
100			
500			
1000			
10000			
100000			
1 mill			
5 mill			
Sanne verdier – forventning og standardfeil	**0,6**	**0,2**	

Øvelse 2. Innføring i enkel bruk av R

Vi skal bruke datasettet «api» i R.

- Academic Performance Index (API) for California skoler
- Basert på standardisert testing av elevene
- Data fra alle skoler med minst100 elever
- Enhet i populasjon = skole (Grunnskole/Ungdomsskole/Videregående)
- Populasjonen består av $N = 6194$ observasjoner
- Ser på variabelen: y = api00 = API i 2000
- Middel(y) = 664,7 med min(y) =346 og max(y) =969

Henter datasettet i R:

```
library(survey)
data(api)
```

Alle variablene i api fås ved R-koden `?api.`

NB! Legg all R-kode du vil spare på i «script», uten > tegnet. Det enkleste er faktisk å først skrive alle R-koder i et script, og så kopiere og lime inn i R.

2.1 Først studier av populasjonen

Studere variabelen api00 = API i 2000

```
y=apipop$api00
```

(a) Lag et histogram av y-populasjonen ved R-koden `hist(y).`

(b) Lager et histogram med bredde 5, med relativ frekvens på y-aksen.

```
hist(y,seq(min(y)-5,max(y)+5,5),prob=TRUE)
```

(c) Beregn gjennomsnittet av y i populasjonen ved å bruke R-koden `mean(y)`.

(d) Beregn populasjonsvariansen til y, σ^2, ved R-koden: `var(y)`

(e) Beregn populasjonsstandardavvik, σ, ved R-funksjonen `sqrt`

(f) Lag et boxplott med hensyn til skoletype, stype: E,H,M (grunnskole,videregående,ungdomsskole). Her er R-koden:

```
y=apipop$api00
x=apipop$stype
plot(x,y)
```

Forklaring: linjen i midten = medianen, boksen =1.kvartil -3.kvartil, med 50 % av de sentrale verdiene.

(g) Lag et vanlig spredningsplott for y med hensyn på variabelen « col.grad=prosent av foreldre med college utdannelse» og beregn korrelasjonskoeffisienten. R-koden er:

```
x=apipop$col.grad
plot(x,y)
cor(x,y)
```

2.2 Enkelt tilfeldig utvalg (ETU), beregning av estimat for mean(y), SE og KI

Populasjonsstørrelse: N = 6194
Utvalgsstørrelse: n = 100
R-kode for ETU: `s=sample(1:N,n)`

(a) Trekk et ETU på n=100, estimer med utvalgsmiddel og beregn 95 % konfidensintervall.
R-kode for ETU, estimering og KI:

```
N=6194
n=100
s=sample(1:N,n)
ybar=mean(y[s])
#y[s]er y-verdiene i utvalget s
se=sqrt(var(y[s])*(N-n)/(N*n))
#Merk at (1-f)/n=(N-n)/(N•n)
```

```
ybar
var(y[s])
se
CI=ybar +qnorm(c(0.025,0.975))*se
CI
```

(b) Lag et histogram for utvalgsdata.

2.3 Testing av faktisk konfidensnivå (dekningsgrad) ved å trekke flere utvalg. La b betegne antall utvalg som trekkes, antall simuleringer

(a) Estimer konfidensnivået når n=5, 10, 50 100, 500, basert på 1000 og 10 000 trekkinger av utvalget. Oppsummer resultatene i tabell 2 nedenfor.

Her er det laget en R-funksjon for simulering og beregning av konfidensnivå:

```
sim=function(b,n,N)
{
ybar=numeric(b)
se=numeric(b)
 for (k in 1:b){
 s=sample(1:N,n)
 ybar[k]=mean(y[s])
se[k]=sqrt(var(y[s])*(N-n)/(N*n))
 }
dek=sum(mean(y)<ybar+1.96*se)-sum(mean(y)<ybar-1.96*se)
konf.nivå=dek/b
list(konf.nivå=konf.nivå)
}
```

Et eksempel på bruk av funksjonen for 1000 simuleringer og n =100:

```
sim(1000,100,6194)
$konf.nivå
[1] 0.951
```

Tabell 2 Konfidensnivå

n	1000 simuleringer	10000 simuleringer
5		
10		
50		
100		
500		

(b) Beregn feilmarginen for estimering av et konfidensnivå med sann verdi 0,95, basert på 1000 og 10 000 simuleringer.

2.4 Histogram av 1000 og 10000 simulerte ybar med normalfordelingstilpasning. For å illustrere at ybar har en sannsynlighetsfordeling som ligner på normalfordelingen*

(a) Lag histogram med tilpasset normalfordeling for utvalgsgjennomsnittet `ybar` for n = 5, 10 og 100 basert 1000 ETU.
(b) Lag histogram med tilpasset normalfordeling for utvalgsgjennomsnittet `ybar` for n = 5, 10 og 100 basert 10 000 ETU.

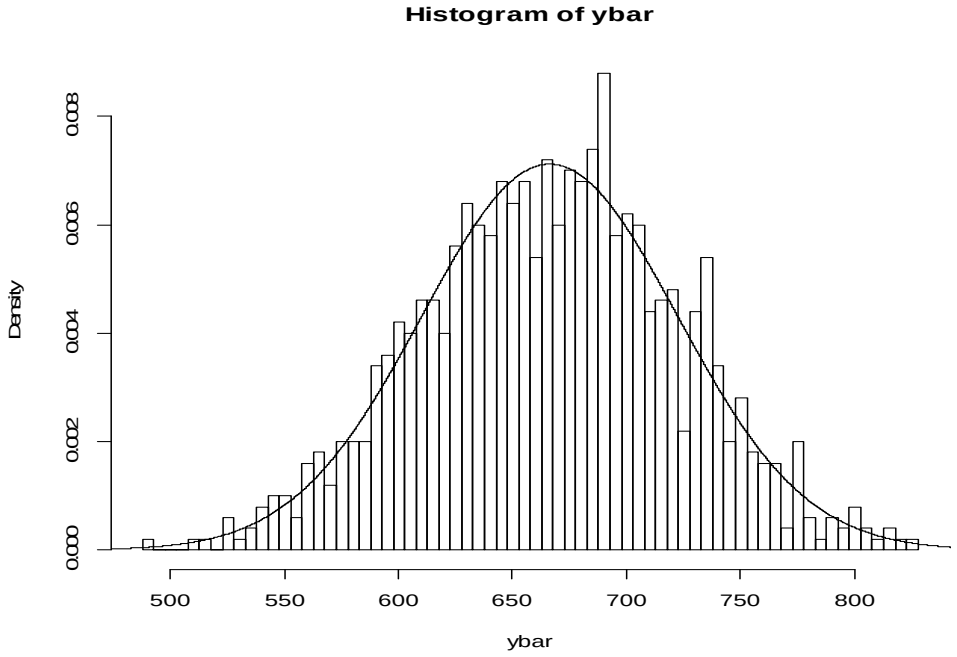
Bruk følgende R-funksjon for relativ frekvens histogram and tilpasset normalfordeling:

```
Tilpasning = function(b,n,N)
{
ybar=numeric(b)
for (k in 1:b){
s=sample(1:N,n)
ybar[k]=mean(y[s])
}
hist(ybar,seq(min(ybar)-5,max(ybar)+5,5),prob=TRUE)
x=seq(mean(ybar)-4*sqrt(var(ybar)),mean(ybar)+4*sqrt(var(ybar)),0.05)
z=dnorm(x,mean(ybar),sqrt(var(ybar)))
```

```
lines(x,z)
}
```

Her er et eksempel på bruk av R-funksjonen. Med n = 5, basert på b = 1000 simulerte utvalg:

```
Tilpasning(1000,5,6194)
```



Øvelse 3. Illustrasjon av Horvitz-Thompson. Tolkning av begrepene forventning og standardfeil til en estimator*

Denne oppgaven skal illustrere og tolke begrepene forventning og varians/standardfeil. Samtidig er det også en illustrasjon på det faktum at Horvitz-Thompson estimatoren kan ha meget stort standardfeil hvis y- verdiene og trekkesannsynlighetene ikke har noen sammenheng. I slike tilfeller bør ikke denne estimatoren brukes selv om den er forventningsrett.

Vi skal betrakte en forenklet versjon av Basus elefanteksempel. En populasjon på tre elefanter skal sendes med båt og vi trenger et estimat for den totale vekten. Å veie en elefant er ingen enkel affære. Eieren ønsker å anslå total vekt ved å veie kun en elefant. Fra tidligere har det vist seg at elefant 2 har a vekt y_2 nær gjennomsnittsvekten for de tre elefantene. Så eieren ønsker å veie denne elefanten og bruke $3y_2$ som estimat, men for å få en forventningsrett estimator så må alle trekkesannsynlighetene være positive.

Følgende utvalgsplan velges for å gjøre det høyst sannsynlig at elefant 2 blir veid:

$|s| = 1$, dvs., $n = 1$, med trekkesannsynlighetene: $\pi_2 = 0{,}90$ og $\pi_1 = \pi_3 = 0{,}05$.

La oss anta at de sanne vektene for elefantene 1,2,3 er 1, 2, 4 tonn, med total vekt = 7 tonn.
Vi skal sammenligne H-T estimatoren med estimatoren $3y$, hvor y er vekten til den valgte elefanten.

3.1 Estimat-verdiene

List opp de mulige verdiene disse to estimatorene kan ha.

3.2 Trekking av utvalg

(a) Trekk et utvalg i R etter utvalgsplanen ovenfor og beregn verdiene av de to estimatorene.

En måte å gjøre det på er å lage en ny populasjon med 100 y-verdier hvor 90 er lik 2, 5 er lik 1 og 5 er lik 4 og så trekke ETU utvalg på størrelse 1. Da vil trekkesannsynlighetene for verdiene 1, 2, 4 være 0,05, 0,90 og 0,05 henholdsvis.

Følgende R-kode kan brukes til å konstruere en vektor med disse verdiene:

```
x=c(2)[rep(c(1),times=90)]
#x er en vector som gjentar verdien 2 90 ganger
y=c(1,1,1,1,1,x,4,4,4,4,4)
```

Utvalget og $3y$ estimatet:

```
s=sample(1:100,1)
ytot=3*y[s]
```

For å beregne HT-estimatoren så definerer vi en trekkepopulasjon med tilsvarende trekkesannsynligheter som y:

```
p1=c(0.05,0.05,0.05,0.05,0.05)
p2=c(0.9)[rep(c(1),times=90)]
p=c(p1,p2,p1)
```

H-T estimatoren:

```
ht=y[s]/p[s]
```

(b) Skriv ut `p`-vektoren.

3.3 Trekking av ti utvalg og empirisk standardfeil

Trekk 10 utvalg og skriv inn estimatene i tabell 3, sammen med estimatenes gjennomsnitt og empiriske standardfeil (kvadratroten av den empiriske variansen til estimatene).

Gjennomsnittet til et datasett y beregnes med `mean(y)`, den empiriske variansen beregnes ved `var(y)` og `sqrt(var(y))` beregner kvadratroten. Skriv ned R-programmet som trekker utvalget, beregner estimatene, og gjennomsnitt og standardfeil for de 10 estimatene.

Tabell 3

Utvalgsnr.	*s*	*y*-verdi	3*y*-estimat	HT-estimat
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
Gjennomsnitt av estimatene				
Empirisk standardfeil				

For å finne ut hva de korrekte verdiene for forventning og standardfeil er for disse to estimatorene skal vi gjenta utvalgstrekkingen mange ganger og beregne gjennomsnittene og de empiriske standardfeilene for H-T estimatene og 3*y* estimatene.

3.4 Simulering av opptil 1 million utvalg, gjennomsnitt og empirisk standardfeil

(a) Beregn gjennomsnitt og empirisk standardfeil for HT-estimatene og 3*y*-estimatene basert på 100, 1000, 10000, 100000 og 1 million trekkinger av ETU. Sett verdiene inn i tabellen nedenfor.

Her er hvordan vi kan gjøre dette ved å trekke 100 ganger.

```
b=100
ytot=numeric(b)
ht=numeric(b)
for(k in 1:b){
s=sample(1:100,1)
ytot[k]=3*y[s]
ht[k]=y[s]/p[s]
 }
mean(ytot)
[1] 6.36
mean(ht)
[1] 8.022222
se1=sqrt(var(ytot))
se2=sqrt(var(ht))
se1
[1] 1.611324
se2
[1] 20.00348
```

Dette betyr at vi generelt kan lage følgende R-funksjon:

```
forv.se=function(b)
{ytot=numeric(b)
ht=numeric(b)
for(k in 1:b){
s=sample(1:100,1)
ytot[k]=3*y[s]
ht[k]=y[s]/p[s]
}
mean(ytot)
mean(ht)
se1=sqrt(var(ytot))
se2=sqrt(var(ht))
list(mean1=mean(ytot),mean2=mean(ht),se1=se1,se2=se2)
```

```
}
```

Tabell 4

Antall simuleringer av utvalg	Gjennomsnitt		Empirisk standardfeil	
	ytotal	HT	ytotal	HT
100				
1000				
10000				
100000				
1 million				
Sann verdi				

(b) Basert på simuleringene, anslå forventning og standardfeil til estimatorene. Hvilken av estimatorene vil dere velge?

Øvelse 4. Frafall

4.1 Valgundersøkelsen 1993 og bruk av etterstratifisering

Vi skal se på Valgundersøkelsen utført av Statistisk sentralbyrå i 1993 (se s.117,118 i forelesningene). Det ble, bl.a., stilt spørsmål om de intervjuede personene stemte ved Stortingsvalgene i 1993 og 1989. Det ble tatt et utvalg på 3000 personer etter SSBs utvalgsplan, og foretatt ialt 11 gjenbesøk. Vi skal bruke data etter to gjenbesøk. Da hadde i alt 1403 personer svart.

(a) Blant de 1403 personene hadde 1190 stemt ved Stortingsvalget i 1993. Hvis det antas at utvalget er et rent tilfeldig utvalg og frafallet også er tilfeldig, utled et estimat og et 95% konfidensintervall for stemmeandelen i 1993.

(b) Den sanne stemmeandelen i 1993 var 0,755. Sammenlign med estimatet og konfidens-intervallet i punkt (a). Hva kan du si om antagelsene som ble gjort i punkt (a) ?

- - - - - - - - - - -

For å prøve å rette opp noe av skjevheten med estimeringen i punkt (a) viste vi på forelesning (s.118 i forelesningene) at etterstratifisering etter valgdeltakelse i 1989 ga estimatet 0,818.

(c) Sammenlign etterstratifiseringsestimatet med estimatet i punkt (a) og den sanne verdien 0,755. Kommentér.

(d) Anta at svarutvalgene er representative for frafallsgruppene i de 3 strataene, dvs. at det er samme stemmeandel i frafallsgruppene som i svarutvalgene i de 3 etterstrataene. Anta at frafallsgruppene fordeler seg slik på det 3 etterstrataene:

Etterstratum 1: 850

Etterstratum 2: 550

Etterstratum 3: 197

Beregn etterstratifiseringsestimatoren i det tilfelle at det ikke er noe frafall.

(e) Under hvilken forutsetning vil etterstratifiseringsestimatoren være forventningsrett. Vurder om det gjelder i denne situasjonen, og eventuell angrepsmåte hvis det ikke holder.

4.2 Hot-deck imputering og multippel imputering

Vi skal estimere middel inntekt i en stor populasjon og tar et enkelt tilfeldig utvalg $n = 20$ personer. 10 personer svarte med følgene inntekter (i 1000 kr.) 600, 520, 620, 500, 380, 460, 450, 250, 400 and 780. Vi antar tilfeldig frafall.

(a) Bruk R til å gjennomføre en hot-deck imputering for frafallet. Beregn standard 95 % konfidensintervall for middel inntekt i populasjonen, basert på det komplette datasettet med observerte og imputerte verdier.

R-kode for hot-deck imputering:

```
y=c(600,520,620,500,380,460,450,250,400,780)
s=c(1,2,3,4,5,6,7,8,9,10)
simp=sample(s,10,replace=TRUE)
#imputerte verdier:
yimp=y[simp]
#Komplett utvalg med observerte og imputerte verdier
ycomp=c(y,yimp)
```

(b) Bruk R til å beregne standard 95 % konfidensintervall for middel inntekt, kun basert på svarutvalget.

(c) Bruk R til å gjennomføre multippel imputering ved å kombinere 5 hot-deck imputeringer. Bruk både 1 og $1/(1-f_{mis})$ i kombinasjonsformelen på s. 137 i forelesningene. Sammenlign med de to intervallene i (a) og (b).

R-kode for multippel imputering:

```
y=c(600,520,620,500,380,460,450,250,400,780)
s=c(1,2,3,4,5,6,7,8,9,10)
b=5
n=20
```

```
nmis=10
m=5
ybar=numeric(b)
var=numeric(b)
for(k in 1:b){
simp=sample(s,nmis,replace=TRUE)
yimp=y[simp]
ycomp=c(y,yimp)
ybar[k]=mean(ycomp)
var[k]=var(ycomp)/n
ymean=sum(ybar)/b
varimp1=var(ybar)*(1+1/m)
varimp2=var(ybar)*(n/(n-nmis)+1/m)
varbar=sum(var)/b
se1=sqrt(varbar+varimp1)
se2=sqrt(varbar+varimp2)
}
CI_1=ymean+qnorm(c(0.025,0.975))*se1
CI_2=ymean+qnorm(c(0.025,0.975))*se2
```

Øvelse 5. En innføring i betydningen av utvalgsplan og valget av estimator for bedriftsundersøkelser

Vi har en populasjon på 4 bedrifter. Variabelen av interesse er omsetningen y i løpet av et år. Anta at omsetningen for et gitt år for de fire bedriftene 1, 2, 3, 4 er 100, 200, 300 og 1000 millioner kroner. Antall sysselsatte (x) i hver bedrift er kjent på forhånd. Anta at x = 20, 30, 50 og 200 for bedriftene 1-4. Vi skal på forskjellige måter trekke et utvalg på 2 bedrifter for å estimere den totale omsetningen (som vi vet er 1600, selvsagt). I de første tre oppgavene skal vi se på estimatorer som ikke bruker tilleggsinformasjonen x. De tre neste oppgavene ser på bruk av rate-estimatoren.

Tre merknader til oppgavene:

1. Med *utvalgsplan* menes samlingen av alle sannsynlighetene $p(s)$ for alle mulige utvalg s. Dvs., utvalgsplanen angir alle sannsynlighetene $p(s)$.

2. Med *standardfeil* (*SE*) til estimatoren menes kvadratroten av variansen, og ikke som vanlig, den estimerte standardfeilen.

3. Med middel kvadrat feil (*MSE*, for engelsk «mean squared error») for en estimator $\hat{t}$, som ikke er forventningsrett for totalen t, menes $E(\hat{t}-t)^2$. *MSE* kan beregnes som summen av variansen og kvadratet av estimatorens skjevhet; $MSE = Var(\hat{t}) + [E(\hat{t})-t]^2$.

5.1 Utvalgsplan 1

Bedrift 4 skal være med, og den andre bedriften trekkes fra bedriftene 1, 2, 3 med sannsynligheter proporsjonal med antall sysselsatte:

bedrift 1: 0,2
bedrift 2: 0,3
bedrift 3: 0,5
Skriv ned utvalgsplanen.

(a) Skriv ned Horvitz-Thompson (HT) estimatoren og vis at den er forventningsrett. Beregn standardfeilen (SE) til estimatoren.

5.2 Utvalgsplan 2

Bedrift 4 skal være med, og den andre bedriften trekkes fra bedriftene 1, 2, 3 med sannsynligheter

bedrift 1: 0,5
bedrift 2: 0,3
bedrift 3: 0,2
Skriv ned utvalgsplanen.

(a) Vis at HT-estimatoren er forventningsrett, og beregn standardfeilen til estimatoren.
(b) Det viser seg at standardfeilen til HT-estimatoren er mye større enn i utvalgsplan 1. Finn en estimator, uten bruk av x, som vil være mer nøyaktig. Beregn forventning, *SE* og $\sqrt{MSE}$.

5.3 Utvalgsplan 3 (ETU)

Vi trekker et enkelt tilfeldig utvalg på 2 bedrifter. Skriv ned utvalgsplanen.

(a) Vis at ekspansjonsestimatoren er lik Horvitz-Thompson estimatoren, og beregn standardfeilen til estimatoren.

5.4 Rate-estimatoren i utvalgsplan 1

Finn forventningen til rate-estimatoren.
(a) Beregn standardfeilen til rate-estimatoren.
(b) Beregn middel kvadratfeil, *MSE*, og $\sqrt{MSE}$.

5.5 Rate-estimatoren i utvalgsplan 2

(a) Finn forventningen til rate-estimatoren.
(b) Beregn standardfeilen til rate-estimatoren.
(c) Beregn middel kvadratfeil, *MSE*, og $\sqrt{MSE}$.

5.6 Rate-estimatoren i utvalgsplan 3 (ETU)

(a) Finn forventningen til rate-estimatoren.
(b) Beregn standardfeilen til rate-estimatoren.
(c) Beregn middel kvadratfeil, *MSE*, og $\sqrt{MSE}$.

5.7 Sammenlikning av utvalgsplaner og estimatorer

(a) Hvis vi ikke har kjennskap til antall sysselsatte, hvilken utvalgsplan av de tre i oppgavene 1-3 ville du valgt?
(b) Med kjennskap til antall sysselsatte i alle bedriftene, hvilken utvalgsplan og estimator ville du valgt? Hvilke andre trekk ved resultatene i oppgavene 1-6 synes du er viktige?

Øvelse 6. Bedriftsundersøkelser for økonomisk statistikk

Vi skal bruke datasettet pop_industri som inneholder en populasjon på 415 observasjoner av bedrifter innenfor en gitt næring. Datasettet ligger på mappen Q:\Introkurs. For hver bedrift er det registrert 4 variable:

idnr: konstruert identitetsnummer for bedriften, fra 001 til 415
oms: bedriftens omsetning, angitt i 1000 kroner
syss: antall sysselsatte i bedriften, antas kjent for hele populasjonen
stratum: inndeling etter sysselsetting i 6 strata;
stratum 1: syss >100
stratum 2: 50< syss ≤100
stratum 3: 20< syss ≤50
stratum 4: 10< syss ≤20
stratum 5: 5< syss ≤10
stratum 6: syss ≤ 5

Omsetning og sysselsetting er hentet fra Virksomhets- og foretaksregisteret. Variablene i R defineres ved:

```
y=pop_industri$oms
x=pop_industri$syss
z=pop_industri$stratum
```

Vi skal estimere total omsetning i populasjonen på 415 bedrifter med forskjellige utvalgsplaner og estimeringsmetoder.

6.1 Fulltelling og enkelt tilfeldig utvalg

(a) Det skal være fulltelling for alle bedrifter med mer enn 50 sysselsatte, og enkelt tilfeldig utvalg blant resten. Størrelsen på utvalget skal være slik at fulltellingsstratumet og utvalget i alt består av 25 bedrifter. Beregn estimater og 95 % konfidensintervaller basert på rate- og ekspansjonsestimatorene. Foreta trekkingen ti ganger og fyll ut tabell 5 nedenfor. Sammenlign de to estimatorene. Hvilken vil du foretrekke?

R-koden for enkelt tilfeldig utvalg og totalestimat basert på utvalgsmiddel er gitt i øvelse 2. Sammen med rate-estimering kan følgende R-kode brukes for enkelt tilfeldig utvalg med utvalgsstørrelse n:

```
s=sample(415,n)
totalest=415*mean(y[s])
totalest
se=415*sqrt(var(y[s])*(415-n)/(415*n))
se
CI=totalest+qnorm(c(0.025,0.975))*se
CI
rateest=sum(x)*mean(y[s])/mean(x[s])
rateest
r=mean(y[s])/mean(x[s])
ssqr=(1/(n-1))*sum((y[s]-r*x[s])^2)
ser=415*sqrt((mean(x)/mean(x[s]))^2*((415-n)/415)*ssqr/n)
ser
CIr=rateest+qnorm(c(0.025,0.975))*ser
Cir
```

Merk sann verdi:

```
t=sum(y)
t
[1] 3309622
```

En måte å definere y- og x-verdiene i de 6 strata er å bruke stratumvariabelen på følgende måte:

```
y1=y[z==1]
y2=y[z==2]
y3=y[z==3]
y4=y[z==4]
y5=y[z==5]
y6=y[z==6]
x1=x[z==1]
x2=x[z==2]
x3=x[z==3]
```

```
x4=x[z==4]
x5=x[z==5]
x6=x[z==6]
```

Definerer «restpopulasjonsverdiene» av y og x:

```
yrest=c(y3,y4,y5,y6)
xrest=c(x3,x4,x5,x6)
```

(b) Skriv ned R-koden for å trekke utvalg, beregn de to estimatene og tilhørende standardfeil og konfidensintervaller, ved å bruke R-programmet for ETU med yrest og xrest.

Tabell 5 Ti 95 % konfidensintervaller i millioner kroner

Utvalgsnr	Ekspansjons-estimat (se i parentes)	KI basert på ekspansjons-estimat	Ratebasert estimat (se)	KI basert på rate-estimat
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

(c) I en kjøring av ti enkelt tilfeldige utvalg av 25 bedrifter ble den empiriske standardfeilen 1589 for ekspansjonsestimatoren og 730 for rate-estimatoren. Sammenlign utvalgsplanen i punkt (a) med enkelt tilfeldig utvalg.

6.2 Fulltelling, cut-off og enkelt tilfeldig utvalg*

La det igjen være fulltelling for alle bedrifter med mer enn 50 sysselsatte, men nå skal ingen bedrifter med 5 eller færre sysselsatte være med. Trekk et enkelt tilfeldig utvalg blant resten med samme størrelse som i oppgave 1. Beregn estimater basert på rate- og ekspansjonsestimatorene. Foreta trekkingen og estimeringen ti ganger og fyll inn resultatene i tabell 6 nedenfor. Sammenlign denne utvalgsplanen med utvalgsplanen i oppgave 6.1.

Tabell 6 Ti estimater for total omsetning, i millioner kroner

Utvalgsnr	Ekspansjonsestimat	Ratebasert estimat
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

Løsninger til øvelser i KLAR 311

Øvelse 1. Begrepsforståelse. En innføring i sannsynlighetsbegreper

1.1 Sannsynlighetsberegninger

(a) List opp alle 10 utvalg på 3 personer, nummerert 1-10:

Utvalgsnr.	1	2	3	4	5	6	7	8	9	10
Utvalg	(1, 2,3)	(1,2 ,4)	(1,2, 5)	(1,3, 4)	(1,3, 5)	(1,4, 5)	(2,3, 4)	(2,3, 5)	(2,4, 5)	(3,4, 5)

(b) Alle utvalg er like sannsynlige, dvs., hvert mulig utvalg på 3 personer har sannsynlighet 1/10.
(c) P(person 1 i utvalget) = P(utvalgene 1-6)=6/10 = 0,6.
Kan også bruke at trekkesannsynligheten for hver enhet er n/N =3/5.
(d) P(personene 1 og 2 i utvalget) = P(utvalgene 1-3)=3/10 = 0,3.
(e) P(alle i utvalget er sysselsatt) =P(utvalg 4) = 1/10 =0,1.

1.2 Forventning og varians

(a)
Først ser vi at $P(X = 0)=0$.
$P(X=1)$=P(utvalgene 3,8,9)=3/10
$P(X=3)$ = P(utvalg 4)=1/10
Dermed, siden summen av sannsynlighetene må bli 1:
$P(X=2)$ = 6/10 ($X = 2$ inntreffer for utvalgene 1,2,5,6,7,10).

Merk: X har en hypergeometrisk fordeling, i tilfelle noen kjenner til denne sannsynlighetsfordelingen:
$P(X = x) = \binom{3}{x}\binom{2}{3-x} / \binom{5}{3}$ = 3/10, 6/10 og 1/10 for x=1,2,3.

(b) $E(X)= 1\cdot 0,3+2\cdot 0,6+3\cdot 0,1=1,8$.

(c) $E(\hat{p}) = E(X/3)= \frac{1}{3}\cdot 0,3+\frac{2}{3}\cdot 0,6+\frac{3}{3}\cdot 0,1 = 0,1+0,4+0,1 = 0,6 = 3/5 = p.$

Vi ser at $E(\frac{X}{3}) = E\left(\frac{1}{3}X\right) = \frac{1}{3}E(X)$. Det er ingen tilfeldighet. Hvis a er en konstant så vil generelt for en *stokastisk variabel X,* $E(aX) = aE(X)$.

(d) $Var(\hat{p}) = E(\hat{p}-p)^2 = E(\frac{X}{3}-\frac{3}{5})^2 =$

$$(\frac{1}{3}-\frac{3}{5})^2\cdot 0,3+(\frac{2}{3}-\frac{3}{5})^2\cdot 0,6+(\frac{3}{3}-\frac{3}{5})^2\cdot 0,1 = (\frac{5-9}{15})^2\cdot 0,3+(\frac{10-9}{15})^2 0,6+(\frac{2}{5})^2\cdot 0,1$$

$$= \frac{4^2}{15^2}\cdot\frac{3}{10}+\frac{1}{15^2}\cdot\frac{3}{5}+\frac{2^2}{5^2}\cdot\frac{1}{10} = \frac{16}{750}+\frac{1}{375}+\frac{4}{250} = \frac{16+2+12}{750} = \frac{30}{750} = \frac{1}{25} = 0,04$$

Standardfeil = $\sqrt{0,04} = 0,20$.

Vi kan også bruke "regneregelen" fra punkt (c).
Vi får da:

$$Var(\hat{p}) = E(\hat{p}-p)^2 = E\left(\frac{X-1,8}{3}\right)^2 = E\left(\frac{1}{9}(X-1,8)^2\right) = \frac{1}{9}E(X-1,8)^2$$

$$= \frac{1}{9}[(1-1,8)^2\cdot 0,3+(2-1,8)^2\cdot 0,6+(3-1,8)^2\cdot 0,1]$$

$$= \frac{1}{9}[0,192+0,024+0,144] = 0,36/9 = 0,04.$$

1.3 Tolkning av forventning, standardfeil og estimert standardfeil

En R-kjøring med 200 simuleringer:

```
> forv2se(200)
$mean
[1] 0.5733333

$se
[1] 0.189507

$meanse
[1] 0.1981694
```

Dvs., gjennomsnitt av 200 estimater er 0,573, gjennomsnitt av 200 estimerte standardfeil er 0,198 og empirisk standardfeil er 0,190. I Tabell 1 har vi følgende beregninger:

Gjennomsnitt av estimatene: $\bar{p} = \sum_{i=1}^{b} \hat{p}_i / b$

Empirisk varians: $v = \frac{1}{b-1} \sum_{i=1}^{b} (\hat{p}_i - \bar{p})^2$

Empiriske standardfeil: $\sqrt{v}$

Gjennomsnitt av SE, $\overline{SE}$,: $SE_i = \sqrt{\hat{p}_i(1-\hat{p}_i)/5}, i = 1,...,b$, $\overline{SE} = \sum_{i=1}^{b} SE_i / b$.

Tabell 1

Antall simuleringer av utvalg	Gjennomsnitt av estimat	Empirisk standardfeil	Gjennomsnitt av *SE*
5	0,4667	0,1826	0,2108
10	0,7000	0,1892	0,1687
50	0,5533	0,1858	0,2024
100	0,5967	0,1970	0,1918
500	0,5980	0,1944	0,1923
1000	0,6073	0,2035	0,1870
10000	0,6064	0,1998	0,1886
100000	0,6001	0,2006	0,1895
1 mill	0,6001	0,2003	0,1896
5 mill	0,6000	0,2000	0,1897
Sanne verdier – forventning og standardfeil	**0,6**	**0,2**	

Fra simuleringene, ser vi at $E(SE)$=0,19, mens $\sqrt{Var(\hat{p})} = 0{,}20$.

Øvelse 2. Innføring i enkel bruk av R

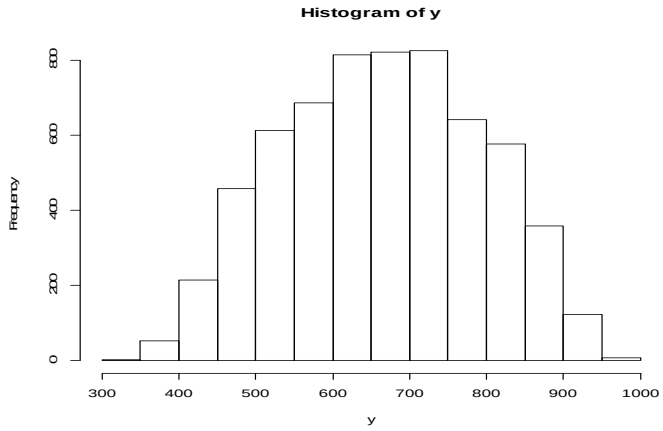
2.1 Først studier av populasjonen

Studere variabelen api00 = API i 2000

```
y=apipop$api00
```

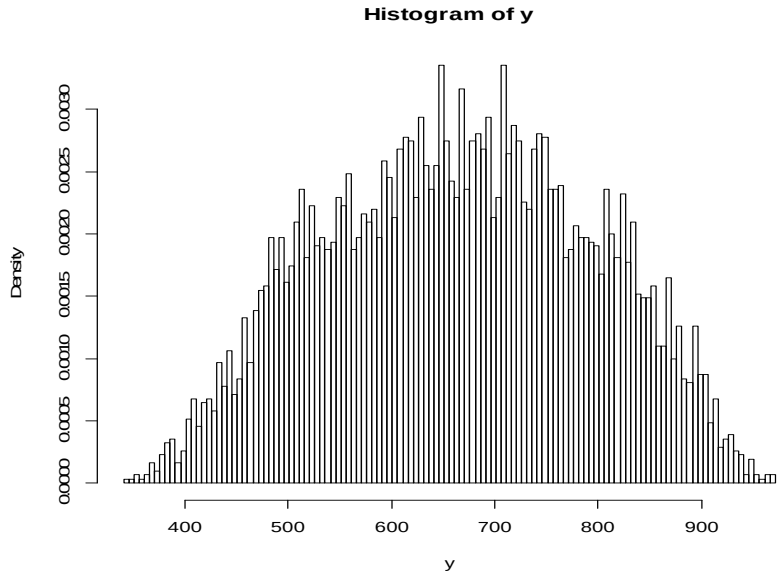
(a) Lag et histogram av y-populasjonen.

```
hist(y)
```



(b) Lag et histogram med bredde 5, med relativ frekvens på y-aksen.

```
hist(y,seq(min(y)-5,max(y)+5,5),prob=TRUE)
```



(c) Beregn gjennomsnittet av y i populasjonen.

```
mean(y)
[1] 664.7126
```

(d) Beregn populasjonsvariansen til y, σ^2.

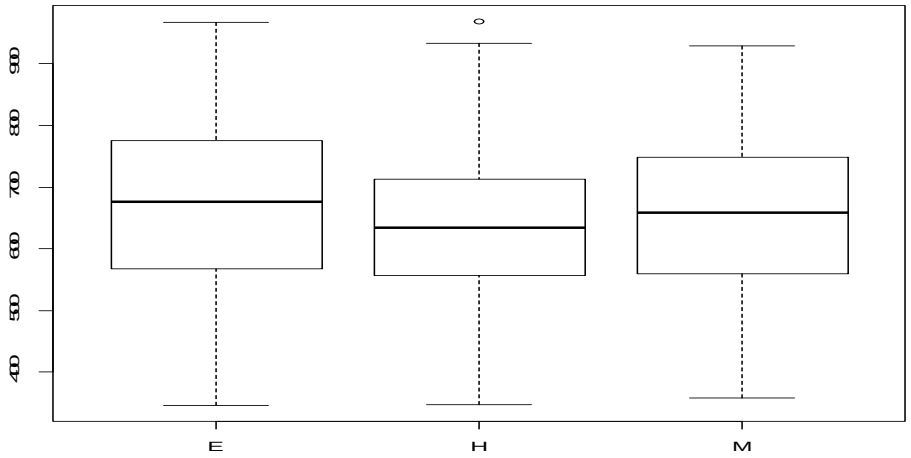
```
var(y)
[1] 16446.56
```

(e) Beregn populasjonsstandardavvik, σ.

```
sqrt(var(y))
[1] 128.2441
```

(f) Lag et boxplott med hensyn til skoletype, stype: E,H,M (grunnskole,videregående, ungdomsskole)

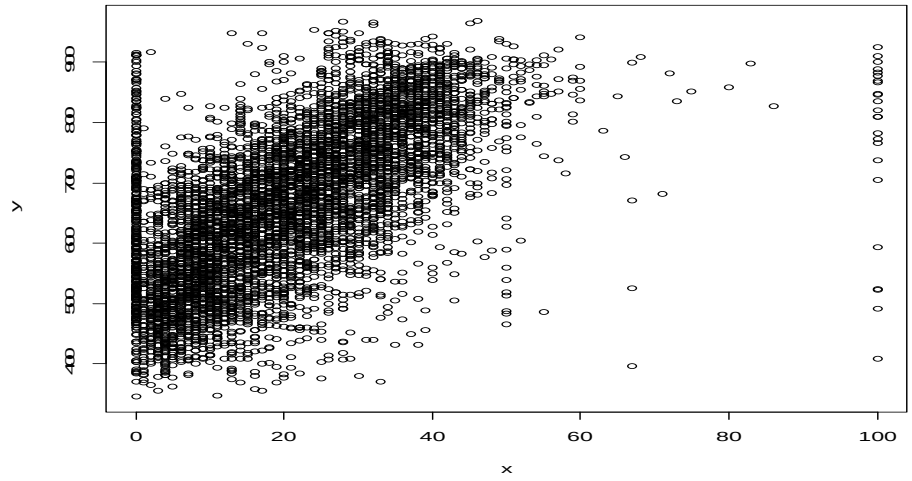
```
y=apipop$api00
x=apipop$stype
plot(x,y)
```



Forklaring: linjen i midten = medianen, boksen =1.kvartil -3.kvartil, med 50% av de sentrale verdiene.

(g) Lag et vanlig spredningsplott for y med hensyn på variabelen « col.grad=prosent av foreldre med college utdannelse» og beregn korrelasjonskoeffisienten.

```
x=apipop$col.grad
plot(x,y)
```



Korrelasjonskoeffisienten

```
cor(x,y)
[1] 0.6263411
```

2.2 Enkelt tilfeldig utvalg (ETU), beregning av estimat for mean(y), SE og KI

(a) Trekk et ETU på n=100, estimer med utvalgsmiddel og beregn 95 % konfidensintervall.

R-kode for ETU, estimering og KI:

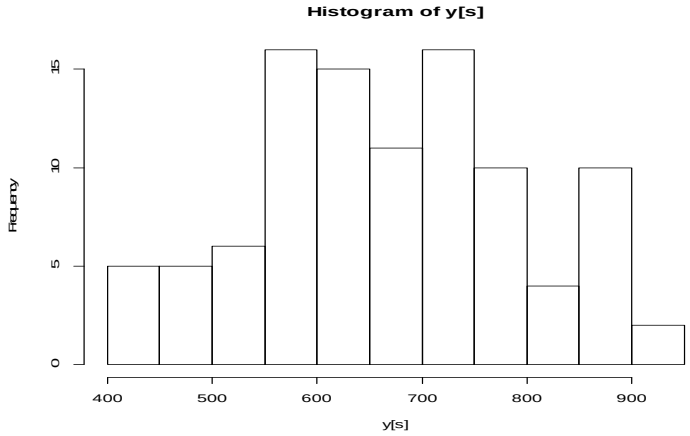
```
N=6194
n=100
s=sample(1:N,n)
ybar=mean(y[s])
se=sqrt(var(y[s])*(N-n)/(N*n))
#Merk at (1-f)/n=(N-n)/(N•n)
ybar
[1] 671.61
var(y[s])
[1] 16045.92
se
[1] 12.56458
```

```
CI=ybar +qnorm(c(0.025,0.975))*se
CI
[1] 646.9839 696.2361
```

Konfidensintervallet: 647,0 – 696,2

(b) Lag et histogram for utvalgsdata:

```
hist(y[s])
```



2.3 Testing av faktisk konfidensnivå (dekningsgrad) ved å trekke flere utvalg. La b betegne antall utvalg som trekkes, antall simuleringer

(a) Har laget en R-funksjon for simulering og beregning av konfidensnivå:

```
sim=function(b,n,N)
{
ybar=numeric(b)
se=numeric(b)
 for (k in 1:b){
 s=sample(1:N,n)
 ybar[k]=mean(y[s])
se[k]=sqrt(var(y[s])*(N-n)/(N*n))
 }
dek=sum(mean(y)<ybar+1.96*se)-sum(mean(y)<ybar-1.96*se)
konf.nivå=dek/b
list(konf.nivå=konf.nivå)
}
```

Et eksempel på bruk av funksjonen for 1000 simuleringer og n =100:

```
sim(1000,100,6194)
$konf.nivå
[1] 0.951
```

Tabell 2. Konfidensnivå

n	1000 simuleringer	10000 simuleringer
5	0,885	0,874
10	0,908	0,917
50	0,940	0,945
100	0,952	0,947
500	0,937	0,953

(b) Beregn feilmarginen for estimering av et konfidensnivå med sann verdi 0,95, basert på 1000 og 10 000 simuleringer.

$b = 1000 : \text{se} = \sqrt{0{,}95 \cdot 0{,}05/1000} = 0{,}0069$ og feilmarginen $= 2 \cdot \text{se} = 0{,}0139$

$b = 10\,000 : \text{se} = \sqrt{0{,}95 \cdot 0{,}05/10000} = 0{,}0022$ og feilmarginen $= 2 \cdot \text{se} = 0{,}0044$

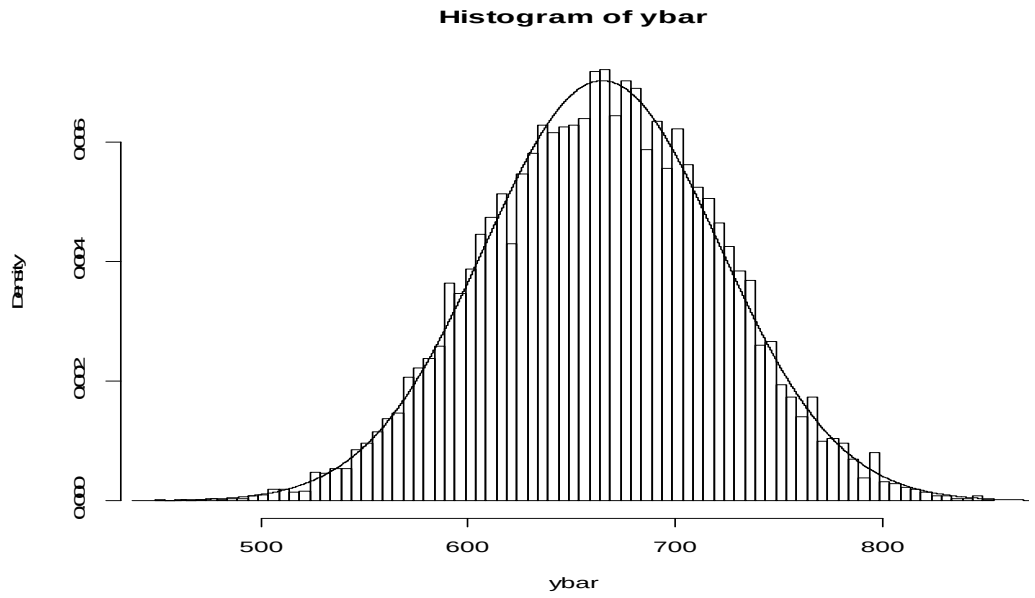
2.4 Histogram av 1000 og 10000 simulerte ybar med normalfordelingstilpasning. For å illustrere at ybar har en sannsynlighetsfordeling som ligner på normalfordelingen*

(a) Lag histogram med tilpasset normalfordeling for utvalgsgjennomsnittet `ybar` for n = 5, 10 og 100 basert 1000 ETU. Bruk R-funksjonen nedenfor.
(b) Lag histogram med tilpasset normalfordeling for utvalgsgjennomsnittet `ybar` for n = 5, 10 og 100 basert 10 000 ETU.

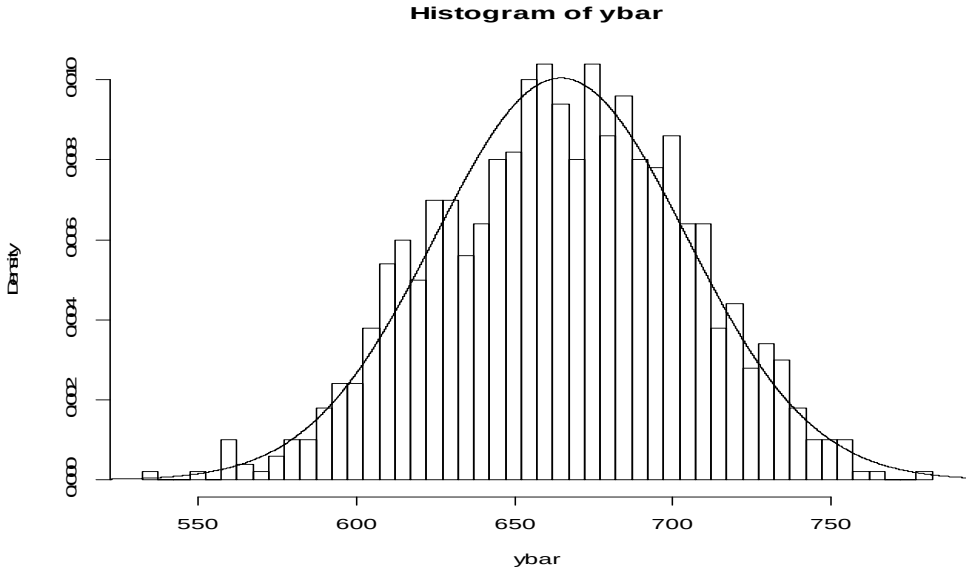
Bruk følgende R-funksjon for relativ frekvens histogram and tilpasset normalfordeling:

```
Tilpasning = function(b,n,N)
{
ybar=numeric(b)
for (k in 1:b){
s=sample(1:N,n)
ybar[k]=mean(y[s])
}
hist(ybar,seq(min(ybar)-5,max(ybar)+5,5),prob=TRUE)
x=seq(mean(ybar)-4*sqrt(var(ybar)),mean(ybar)+4*sqrt(var(ybar)),0.05)
z=dnorm(x,mean(ybar),sqrt(var(ybar)))
lines(x,z)
}
```

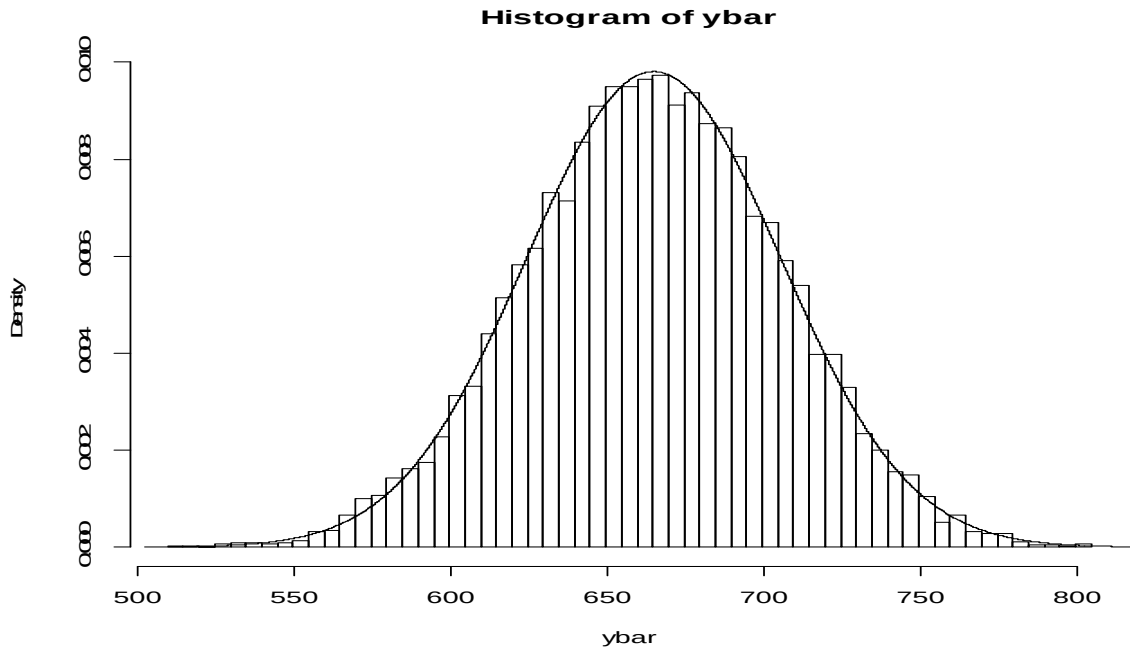
Med n = 5, basert på b= 10000 simuleringer:
`Tilpasning(10000,5,6194)`



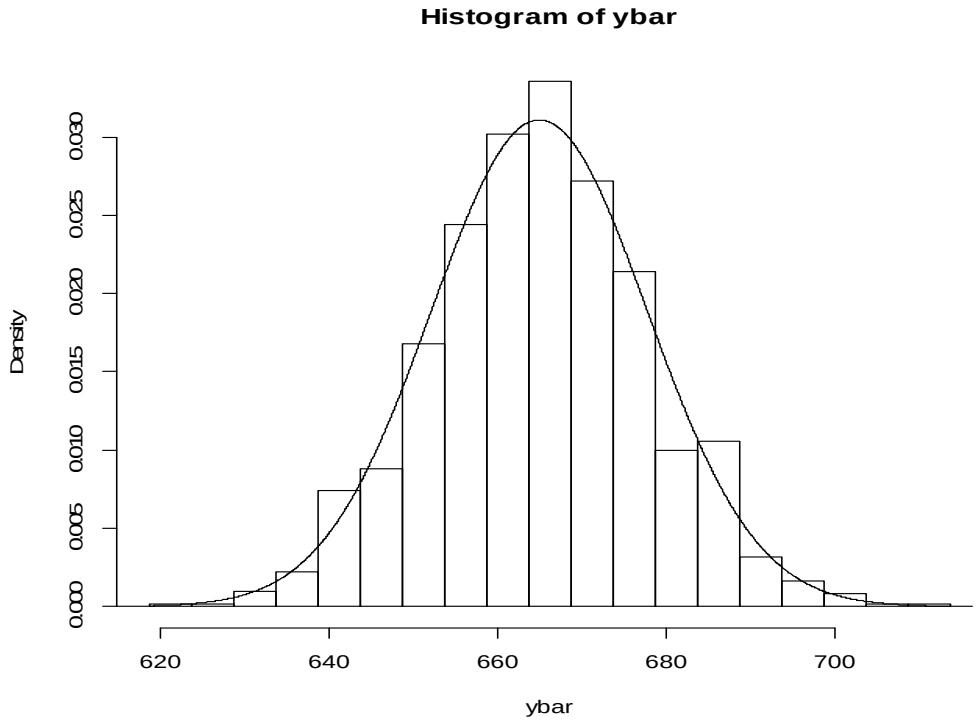
Med n=10, basert på 1000 simuleringer:
`Tilpasning(1000,10,6194)`



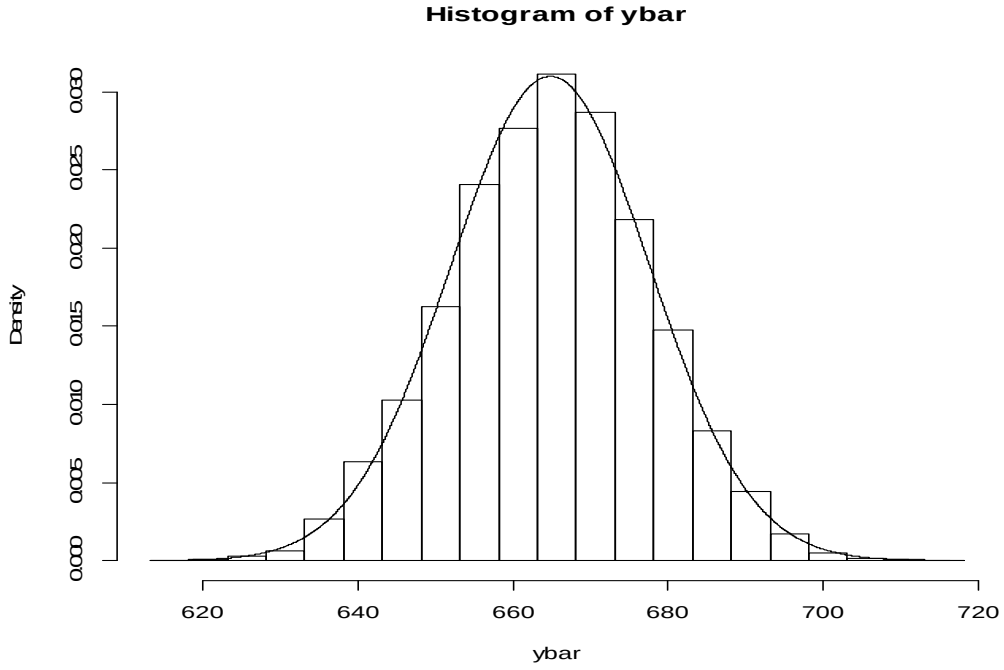
Med n=10, basert på 10000 simuleringer:
`Tilpasning(10000,10,6194)`



Med n= 100, basert på 1000 simuleringer.
`Tilpasning(1000,100,6194)`



Med n= 100, basert på 10000 simuleringer.
`Tilpasning(10000,100,6194)`



Øvelse 3. Illustrasjon av Horvitz-Thompson. Tolkning av begrepene forventning og standardfeil til en estimator*

Følgende utvalgsplan velges for å gjøre det høyst sannsynlig at elefant 2 blir veid:
$|s| = 1$, dvs., $n = 1$, med trekkesannsynlighetene: $\pi_2 = 0{,}90$ og $\pi_1 = \pi_3 = 0{,}05$.
De sanne vektene for elefantene 1,2,3 er 1, 2, 4 tonn, med total vekt = 7 tonn.
Vi skal sammenligne H-T estimatoren med estimatoren $3y$, hvor y er vekten til den valgte elefanten.

3.1 Estimat-verdiene

H-T estimatoren:

$$\hat{t}_{HT} = y_i / \pi_i \text{ hvis } s = \{i\}$$

$$= \begin{cases} 20 & \text{hvis } s = \{1\} \\ 2{,}22 & \text{hvis } s = \{2\} \\ 80 & \text{hvis } s = \{3\} \end{cases}$$

Håpløs! Alltid langt fra den sanne totalen 7. Kan ikke brukes, selv om den er forventningsrett.

Den planlagte estimatoren, selv om vi ikke har ETU:
$ytot = 3y_i$ hvis $s = \{i\}$
Mulige verdier: 3, 6, 12

3.2 Trekking av utvalg

(a) Trekk et utvalg i R etter utvalgsplanen ovenfor og beregn verdiene av de to estimatorene.
Følgende R-kode kan brukes til å konstruere en vektor med disse verdiene:

```
x=c(2)[rep(c(1),times=90)]
#x er en vector som gjentar verdien 2 90 ganger
y=c(1,1,1,1,1,x,4,4,4,4,4)
```

Utvalget og 3*y* estimatet:

```
s=sample(1:100,1)
ytot=3*y[s]
```

For å beregne HT-estimatoren så definerer vi trekkepopulasjonen:

```
p1=c(0.05,0.05,0.05,0.05,0.05)
p2=c(0.9)[rep(c(1),times=90)]
p=c(p1,p2,p1)
```

HT - estimatoren:

```
ht=y[s]/p[s]
```

(b) Skriv ut p-vektoren.

```
p
 [1] 0.05 0.05 0.05 0.05 0.05 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90
 [16] 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90
 [31] 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90
 [46] 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90
 [61] 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90
 [76] 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90 0.90
 [91] 0.90 0.90 0.90 0.90 0.90 0.05 0.05 0.05 0.05 0.05
```

3.3 Trekking av ti utvalg og empirisk standardfeil

Tabell 3

Utvalgsnr.	*s*	*y*-verdi	3*y*-estimat	HT-estimat
1	82	2	6	2,22
2	1	1	3	20
3	86	2	6	2,22
4	3	1	3	20
5	62	2	6	2,22
6	11	2	6	2,22
7	98	4	12	80
8	2	1	3	20
9	68	2	6	2,22
10	66	2	6	2,22
Gjennomsnitt av estimatene			5,7	15,3
Empirisk standardfeil			2,6	63,9

R-programmet:

```
x=c(2)[rep(c(1),times=90)]
y=c(1,1,1,1,1,x,4,4,4,4,4)
p1=c(0.05,0.05,0.05,0.05,0.05)
p2=c(0.9)[rep(c(1),times=90)]
p=c(p1,p2,p1)
s=sample(1:100,1)
ytot=3*y[s]
ht=y[s]/p[s]
ytot
ht
```

Fra verdiene i tabellen:

```
t=c(2.22,2.22,2.22,2.22,2.22,2.22,20,20,20,80)
var(t)
[1] 4087.736
sqrt(var(t))
[1] 63.9354
yt=c(6,6,6,6,6,6,3,3,3,12)
var(yt)
[1] 6.9
sqrt(var(yt))
[1] 2.62678
```

3.4 Simulering av opptil 1 million utvalg, gjennomsnitt og empirisk standardfeil

R-funksjonen:

```
forv.se=function(b)
{ytot=numeric(b)
ht=numeric(b)
for(k in 1:b){
s=sample(1:100,1)
ytot[k]=3*y[s]
ht[k]=y[s]/p[s]
}
mean(ytot)
mean(ht)
se1=sqrt(var(ytot))
se2=sqrt(var(ht))
list(mean1=mean(ytot),mean2=mean(ht),se1=se1,se2=se2)
}
```

En kjøring:

```
forv.se(100)
$mean1
[1] 6.18
$mean2
[1] 7.955556
$se1
[1] 1.641507
$se2
[1] 18.77713
```

Tabell 4

Antall simuleringer av utvalg	Gjennomsnitt		Empirisk standardfeil	
	ytot	HT	ytot	HT
100	6,18	7,96	1,64	18,78
1000	6,14	7,31	1,55	17,54
10000	6,15	7,01	1,49	17,23
100000	6,15	7,03	1,50	17,24
1 million	6,15	6,99	1,49	17,17
Sann verdi	**6,15**	**7,00**	**1,49**	**17,19**

(b) Anslag for forventning og standardfeil til estimatorene.

HT- estimatoren kan ikke brukes, selv om $E(\hat{t}_{HT}) = 7 = t$.

Sann $SE(\hat{t}_{HT}) = \sqrt{Var(\hat{t}_{HT})} = 17{,}2$!!!

$E(ytot) = 6{,}15$; ikke forventningsrett, men se på $SE(ytot) = 1{,}49$.

3*y* er klart å foretrekke framfor HT-estimatoren.

Øvelse 4. Frafall

4.1 Valgundersøkelsen 1993 og bruk av etterstratifisering

(a) De 1403 personene utgjør et enkelt tilfeldig utvalg. $\hat{p} = 1190/1403 = 0{,}848$. $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p}}{n-1} \cdot \frac{N-n}{N}} = 0{,}00959$

.

95 % konfidensintervall: $\hat{p} \pm 1{,}96 \cdot SE(\hat{p}) = 0{,}848 \pm 0{,}019 = (0{,}829 \text{ - } 0{,}867)$.

(b) Den sanne stemmeandelen i 1993 var 0,755. Estimat og konfidensintervall har stor skjevhet. Frafallet er ikke tilfeldig. Det er større frafall blant de som ikke stemte.

- - - - - - - - - - -

Etterstratifisering etter valgdeltakelse i 1989 ga estimatet 0,818.

(c) Estimatet basert på etterstratifisering er også skjevt. Etterstratifisering er ikke nok.

(d) Antar at svarutvalgene er representative for frafallsgruppene i de 3 strataene, dvs. at det er samme stemmeandel i frafallsgruppene som i svarutvalgene i de 3 etterstrataene. Antar at frafallsgruppene fordeler seg slik på det 3 etterstrataene:

Etterstratum 1: 850

Etterstratum 2: 550

Etterstratum 3: 197

Etterstratifiseringsestimatoren i det tilfelle at det ikke er noe frafall blir:

Etterstrata blir nå totalt, uten frafall:

Etterstratum 1: 2042

Etterstratum 2: 665

Etterstratum 3: 293

Etterstratifiseringsestimatet: 2042x0,889+665x0,496+293x0,760 = 2368

Estimert andel = 2368/3000= 0,789.

(e) Etterstratifiseringsestimatoren er forventningsrett hvis svarutvalget er representativt for frafallsgruppen innen hvert etterstratum. Det holder ikke her. Her trenger man å anta at sannsynligheten for at en person svarer er avhengig av om personen stemte eller ikke.

4.2 Hot-deck imputering og multippel imputering

(a) Bruker R til å gjennomføre en hot-deck imputering for frafallet.

```
y=c(600,520,620,500,380,460,450,250,400,780)
s=c(1,2,3,4,5,6,7,8,9,10)
simp=sample(s,10,replace=TRUE)
simp
 [1]  4 10  5 10  5  4  2  6  3  4
yimp=y[simp]
yimp
 [1] 500 780 380 780 380 500 520 460 620 500
ycomp=c(y,yimp)
ycomp
 [1] 600 520 620 500 380 460 450 250 400 780 500 780 380 780 380 500 520 460 620
[20] 500
mean(ycomp)
[1] 519
mean(y)
    [1] 496

var(ycomp)
[1] 20472.63
v=var(ycomp)/20
se=sqrt(v)
se
[1] 31.99424
CIimp=mean(ycomp)+qnorm(c(0.025,0.975))*se
CIimp
[1] 456.2924 581.7076
```

Standard 95 % konfidensintervall, basert på det komplette datasettet: (456.3, 581.7)

(b) Standard 95 % konfidensintervall for middel inntekt, kun basert på svarutvalget.

```
vresp=var(y)/10
seresp=sqrt(vresp)
seresp
[1] 46.43275
CIresp=mean(y)+qnorm(c(0.025,0.975))*seresp
CIresp
[1] 404.9935 587.0065
```

Standard 95% konfidensintervall basert på svarutvalget: (405.0, 587.0)
Konfidensintervallet i (a) er altfor kort.

(c) Bruker R til å gjennomføre multippel imputering ved å kombinere 5 hot-deck imputeringer. Bruker både 1 og $1/(1-f_{mis})$ i kombinasjonsformelen på s. 137. R-kode for multippel imputering:

```
y=c(600,520,620,500,380,460,450,250,400,780)
s=c(1,2,3,4,5,6,7,8,9,10)
b=5
n=20
nmis=10
m=5
ybar=numeric(b)
var=numeric(b)
for(k in 1:b){
simp=sample(s,nmis,replace=TRUE)
yimp=y[simp]
ycomp=c(y,yimp)
ybar[k]=mean(ycomp)
var[k]=var(ycomp)/n
ymean=sum(ybar)/b
varimp1=var(ybar)*(1+1/m)
varimp2=var(ybar)*(n/(n-nmis)+1/m)
```

```
varbar=sum(var)/b
se1=sqrt(varbar+varimp1)
se2=sqrt(varbar+varimp2)
}
CI_1=ymean+qnorm(c(0.025,0.975))*se1
CI_2=ymean+qnorm(c(0.025,0.975))*se2
CI_1
[1] 409.5318 566.4682
# Dette er konfidensintervallet som bruker 1 i kombinasjonsformelen: (409.5,
566.5)
CI_2
[1] 396.3449 579.6551
# Dette er konfidensintevallet som bruker 1/(1-f_mis) i kombinasjonsformelen:
(396,3, 579.7))
se1
[1] 40.03551
se2
[1] 46.76368
ymean
[1] 488
```

Vi ser at CI_2 og se2 ligner veldig på resultatene basert på svarutvalget i (b), og er derfor den korrekte måten å kombinere hot-deck imputeringer. CI_1 er for smalt.

Øvelse 5. En innføring i betydningen av utvalgsplan og valget av estimator for bedriftsundersøkelser

5.1 Utvalgsplan 1

(a)
$p(\{1,4\}) = 0,2$
$p(\{2,4\}) = 0,3$
$p(\{3,4\}) = 0,5$
$p(\text{s}) = 0$ ellers.

(b) Horvitz-Thompson estimatoren, med y_i = omsetning for bedrift i:

$$\hat{t}_{HT} = \sum_{i \in s} y_i / \pi_i$$
$$= y_1 / \pi_1 + y_4 = 100 / 0,2 + 1000 = 1500, \text{ hvis } s = \{1,4\}$$
$$= y_2 / \pi_2 + y_4 = 200 / 0,3 + 1000 = 1666,67, \text{ hvis } s = \{2,4\}$$
$$= y_3 / \pi_3 + y_4 = 300 / 0,5 + 1000 = 1600, \text{ hvis } s = \{3,4\}$$

$$E(\hat{t}_{HT}) = 1500 \cdot 0,2 + 1666,67 \cdot 0,3 + 1600 \cdot 0,5 = 300 + 500 + 800 = 1600 = t$$

$$Var(\hat{t}_{HT}) = E(\hat{t}_{HT} - t)^2 = (1500 - 1600)^2 \cdot 0,2 + (1666,67 - 1600)^2 \cdot 0,3 + (1600 - 1600)^2 \cdot 0,5$$

$$= 2000 + 1333,33 = 3333,33$$

$$\Rightarrow SE = \sqrt{3333,33} = \mathbf{57,7}.$$

5.2 Utvalgsplan 2

(a)
$p(\{1,4\}) = 0,5$
$p(\{2,4\}) = 0,3$
$p(\{3,4\}) = 0,2$
$p(\text{s}) = 0$ ellers.

(b)

$$\hat{t}_{HT} = \sum_{i \in s} y_i / \pi_i$$

$= y_1/\pi_1 + y_4 = 100/0{,}5 + 1000 = 1200$, hvis $s = \{1,4\}$

$= y_2/\pi_2 + y_4 = 200/0{,}3 + 1000 = 1666{,}67$, hvis $s = \{2,4\}$

$= y_3/\pi_3 + y_4 = 300/0{,}2 + 1000 = 2500$, hvis $s = \{3,4\}$

$E(\hat{t}_{HT}) = 1200 \cdot 0{,}5 + 1666{,}67 \cdot 0{,}3 + 2500 \cdot 0{,}2 = 600 + 500 + 500 = 1600 = t$

$Var(\hat{t}_{HT}) = E(\hat{t}_{HT} - t)^2 = (1200-1600)^2 \cdot 0{,}5 + (1666{,}67-1600)^2 \cdot 0{,}3 + (2500-1600)^2 \cdot 0{,}2$

$= 80000 + 1333{,}3 + 162000 = 243333{,}33$

$\Rightarrow SE = \sqrt{243333{,}33} = \mathbf{493{,}3}$.

(c) Ny estimator:
Generelt opplegg: Vi har to strata:

- Fulltellingsstratum og utvalgsstratum. Beregn et estimat for utvalgsstratum basert på et utvalg og legg til totalen fra fulltellingsstratumet. I dette tilfellet blir estimatet, hvis det totale utvalget er $s = \{i,4\}$: $\hat{t} = 3 \cdot y_i + y_4$. De mulige verdiene med sannsynligheter:

s	{1,4}	{2,4}	{3,4}
$\hat{t}$	1300	1600	1900
$p(s)$	0,5	0,3	0,2

Dette gir: $E(\hat{t}) = 1300 \cdot 0{,}5 + 1600 \cdot 0{,}3 + 1900 \cdot 0{,}2 = \mathbf{1510}$

$Var(\hat{t}) = E(\hat{t} - 1510)^2 = (1300-1510)^2 \cdot 0{,}5 + (1600-1510)^2 \cdot 0{,}3 + (1900-1510)^2 \cdot 0{,}2$

$= 22050 + 2430 + 30420 = 54900$

$\Rightarrow SE = \sqrt{54900} = \mathbf{234{,}3}$

$\Rightarrow MSE = 54900 + 90^2 = 63000$ og $\sqrt{MSE} = \mathbf{251{,}0}$.

5.3 Utvalgsplan 3 (ETU)

(a) Det er 6 mulige utvalg av 2 bedrifter. Hvert utvalg har sannsynlighet 1/6. Dvs.:
$p(\{i,j\}) = 1/6$ for $\{i,j\} = (1,2), (1,3), (1,4), (2,3), (2,4)$ og $(3,4)$.

(b) Alle trekkesannsynlighetene er $n/N = ½$. Da blir $\hat{t}_{HT} = \sum_{i \in s} y_i/\pi_i = 2 \cdot \sum_{i \in s} y_i = 4 \cdot \bar{y}_s$.

$$Var(4\bar{y}_s) = E(4\bar{y}_s - 1600)^2 = \sum_{s:|s|=2} (4\bar{y}_s - 1600)^2 \cdot \frac{1}{6}$$

$$= \frac{1}{6}\left[(600-1600)^2 + (800-1600)^2 + (2200-1600)^2 + (1000-1600)^2 + (2400-1600)^2 + (2600-1600)^2\right]$$

$$= \frac{1}{6} \cdot 4000000 = 666666{,}667$$

$\Rightarrow SE = \sqrt{666666{,}667} = \mathbf{816{,}5}$

Vi kan også beregne SE fra formelen for variansen til denne estimatoren,

$$Var(N\bar{y}_s) = N^2 \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$$

hvor $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2, \mu = t/N,$ er populasjonsvariansen.

Her får vi at $\sigma^2 = (300^2 + 200^2 + 100^2 + 600^2)/3 = 500000/3$. Det gir at

$Var(4\bar{y}_s) = 4^2 \frac{500000/3}{2}\left(1 - \frac{1}{2}\right) = 4 \cdot 500000/3 = 666666{,}667,$ som før.

5.4 Rate-estimatoren i utvalgsplan 1

(a)

Totale antall sysselsatte = 20+30+50+200=300. Rate-estimatoren er da:

$$\hat{t}_R = 300 \cdot \frac{\sum_s y_i}{\sum_s x_i}.$$

De mulige verdiene av rate-estimatoren med sannsynligheter:

s	{1,4}	{2,4}	{3,4}
$\hat{t}_R$	1500	1565,22	1560
$p(s)$	0,2	0,3	0,5

Forventningen:

$E(\hat{t}_R) = 1500 \cdot 0,2 + 1565,22 \cdot 0,3 + 1560 \cdot 0,5$

$= 300 + 469,6 + 780 = \mathbf{1549,6}$

(b)

$Var(\hat{t}_R) = E(\hat{t}_R - 1549,6)^2 = 49,6^2 \cdot 0,2 + 15,62^2 \cdot 0,3 + 10,4^2 \cdot 0,5 = 492,03 + 73,20 + 54,08 = 619,31.$

Det gir at $SE = \sqrt{619,31} = \mathbf{24,9}$.

(c)

Vi kan samle skjevhet (engelsk:bias) og varians med middel kvadratfeil :

$MSE =$

$E(\hat{t}_R - t)^2 = Var(\hat{t}_R) + [E(\hat{t}_R) - t]^2$

$= 619,31 + 50,4^2 = 3159,5$

og

$\sqrt{E(\hat{t}_R - t)^2} = \sqrt{3159,5} = 56,2.$

$\sqrt{MSE} =$ **56,2**. (engelsk notasjon: root mean squared error, RMSE)

$\sqrt{MSE}$ er sammenlignbart med SE for forventningsrette estimatorer.

5.5 Rate-estimatoren i utvalgsplan 2

(c) De mulige verdiene av rate-estimatoren med sannsynligheter:

s	{1,4}	{2,4}	{3,4}
$\hat{t}_R$	1500	1565,22	1560
$p(s)$	0,5	0,3	0,2

$E(\hat{t}_R) = 1500 \cdot 0,5 + 1565,22 \cdot 0,3 + 1560 \cdot 0,2$

$= 750 + 469,6 + 312 = \mathbf{1531,6}$

(b) $Var(\hat{t}_R) = E(\hat{t}_R - 1531,6)^2 = 31,6^2 \cdot 0,5 + 33,62^2 \cdot 0,3 + 28,4^2 \cdot 0,2 = 499,28 + 339,09 + 161,31 = 999,68.$

Det gir at SE = $\sqrt{999,68} =$ **31,6**.

(c) MSE =

$E(\hat{t}_R - t)^2 = Var(\hat{t}_R) + [E(\hat{t}_R) - t]^2$

$= 999,68 + 68,4^2 = 5678,24$

og $\sqrt{\text{MSE}} = \sqrt{5678,24} =$ **75,4**.

5.6 Rate-estimatoren i utvalgsplan 3 (ETU)

(a) De mulige verdiene av rate-estimatoren, hver med sannsynlighet 1/6, blir:

s	{1,2}	{1,3}	{1,4}	{2,3}	{2,4}	{3,4}
$\hat{t}_R$	1800	1714,29	1500	1875	1565,22	1560

Forventningen:

$E(\hat{t}_R) = \frac{1}{6}(1800 + 1714{,}29 + 1500 + 1875 + 1565{,}22 + 1560) = \mathbf{1669{,}1}$.

(b)

$Var(\hat{t}_R) = E(\hat{t}_R - 1669{,}1)^2$

$= \frac{1}{6}\left[(1800 - 1669{,}1)^2 + (1714{,}29 - 1669{,}1)^2 + (1500 - 1669{,}1)^2 + (1875 - 1669{,}1)^2\right]$

$+ \frac{1}{6}\left[(1565{,}22 - 1669{,}1)^2 + (1560 - 1669{,}1)^2\right]$

$= \frac{1}{6}[17134{,}81 + 2042{,}14 + 28594{,}81 + 42394{,}81 + 10791{,}05 + 11902{,}81]$

$= 112860{,}43/6 = 18810{,}07$

Det gir: $SE = \sqrt{18810{,}07} = \mathbf{137{,}2}$.

(c) $MSE =$

$E(\hat{t}_R - t)^2 = Var(\hat{t}_R) + [E(\hat{t}_R) - t]^2$

$= 18810{,}07 + 69{,}1^2 = 23584{,}88$

og $\sqrt{MSE} = \sqrt{23584{,}88} = \mathbf{153{,}6}$.

5.7 Sammenlikning av utvalgsplaner og estimatorer

(a) HT- estimatorene er alle forventningsrette og standardfeilene er følgende:

Utvalgsplan 1: 57,7

Utvalgsplan 2: 493,3 (Alternativ estimator: SE= 234,3 og skjevhet = -90,0; $\sqrt{MSE} = 251{,}0$.)

Utvalgsplan 3: 816,5

Overlegent best: Utvalgsplan 1, jo større bedrift jo høyere trekkesannsynlighet og tar alltid med den største.

(b) Oppsummert:

Utvalgsplan 1 og HT-estimator: SE = 57,7

Utvalgsplan 2 og HT-estimator: SE = 493,3 (Alternativ estimator SE = 234,3: $\sqrt{MSE} = 251{,}0$.)

Utvalgsplan 3 og HT-estimator: SE = 816,5

Utvalgsplan 1 og rate-estimator: SE = 24,9 og estimeringsskjevhet = - 50,4 ; $\sqrt{MSE}$ = 56,2

Utvalgsplan 2 og rate-estimator: SE = 31,6 og estimeringsskjevhet = - 68,4; $\sqrt{MSE}$ = 75,4

Utvalgsplan 3 og rate-estimator: SE = 137,2 og estimeringsskjevhet = + 69,1; $\sqrt{MSE}$ = 153,6

Noen konklusjoner og vurderinger:

- For det første ser vi, i utvalgsplan 2, at HT-estimatoren fungerer ekstremt dårlig når trekkesannsynlighetene er veldig negativt korrelert med selve y-verdiene. Her er SE 8,5 ganger større enn SE i utvalgsplan 1. Rate-estimatoren er mye mer robust. $\sqrt{MSE}$ i utvalgsplan 2 er kun 1,3 ganger større i utvalgsplan 2.
- Det er ikke smart å trekke enkelt tilfeldig utvalg i bedriftsundersøkelser med stor variasjon i y-verdiene.
- Hvis vi er så «uheldig» at vi velger en utvalgsplan som utvalgsplan 2 så er HT-estimatoren ubrukelig selv om den er forventningsrett. Rate-estimatoren har en $\sqrt{MSE}$ som er kun 15 % av HT-estimatorens SE. Hvis vi ikke har tilleggsinformasjon så vil det her vært bedre å bruke $\hat{t} = 3 \cdot y_i + y_4$, selv om den ikke er forventningsrett.
- Ikke så enkelt å konkludere her på estimator. Utvalgsplan 1 er fremdeles opplagt den beste av de tre utvalgsplanene. Rate-estimatoren har mye mindre SE enn HT-estimatoren. Samtidig er skjevheten såpass stor

at $\sqrt{MSE}$ er omtrent lik *SE* for HT-estimatoren. For populasjoner og utvalg av vanlig størrelse vil skjevheten til rate-estimatoren være betraktelig mindre slik at den vanligvis vil foretrekkes.

Øvelse 6. Bedriftsundersøkelser for økonomisk statistikk

Vi skal estimere total omsetning i populasjonen på 415 bedrifter med forskjellige utvalgsplaner og estimeringsmetoder.

6.1 Fulltelling og enkelt tilfeldig utvalg

(a) Fulltelling for alle bedrifter med mer enn 50 sysselsatte, og enkelt tilfeldig utvalg blant resten, i alt 25 bedrifter. Skal beregne estimater og 95 % konfidensintervaller basert på rate- og ekspansjonsestimatorene, ti ganger.

y-verdiene og x-verdiene i de 6 strata:

```
y1=y[z==1]
y2=y[z==2]
y3=y[z==3]
y4=y[z==4]
y5=y[z==5]
y6=y[z==6]
x1=x[z==1]
x2=x[z==2]
x3=x[z==3]
x4=x[z==4]
x5=x[z==5]
x6=x[z==6]
```

```
y1
[1] 107298 247791
y2
 [1] 52070 95016 75494 24848 103865 32029 90780 84508 32205 132924
[11] 38640
```

Det betyr 13 bedrifter i fulltellingsstratumet. Da skal utvalgsstørrelsen være 12. Kan også finne dette direkte ved R-koden:

```
> length(y1)
[1] 2
> length(y2)
[1] 11
```

(b) R koden:

```
yrest=c(y3,y4,y5,y6)
xrest=c(x3,x4,x5,x6)
s=sample(402,12)
fulltelling=sum(y1)+sum(y2)
fulltelling
[1] 1117468
yrest[s]
 [1] 5836 5923 25439 161 402 2129 312 73 3965  3744  1427  1331
totrest=402*mean(yrest[s])
totrest
[1] 1699857
totest=totrest+fulltelling
totest
[1] 2817325
rateest=sum(xrest)*mean(yres[s])/mean(xrest[s])
rateest
[1] 1740313
ratetot=rateest+fulltelling
ratetot
[1] 2857781
mean(xrest)
[1] 6.313433
```

```
mean(xrest[s])
[1] 6.166667

se=402*sqrt(var(yres[s])*(402-12)/(402*12))
 se
[1] 800924.6
CI=totrest+qnorm(c(0.025,0.975))*se
CI
[1]  130073.6 3269640.4
r=mean(yres[s])/mean(xrest[s])
ssqr=(1/11)*sum((yres[s]-r*xrest[s])^2)
ser=415*sqrt((mean(xrest)/mean(xrest[s]))^2*((402-12)/402)*ssqr/12)
ser
[1] 344698.1
CIr=rateest+qnorm(c(0.025,0.975))*ser
CIr
[1] 1064718 2415909
CItot=fulltelling+CI
CItot
[1] 1247542 4387108
   # KI = 1 247 542 – 4 387 108
CIrtot=fulltelling+CIr
CIrtot
[1] 2182186 3533377
# KI = 2 182 186 – 3 533 377, mye mindre enn CItot
```

For å gjøre dette ti ganger, legg følgende R-program inn i et script:

```
s=sample(402,12)
fulltelling=sum(y1)+sum(y2)
yres[s]
totrest=402*mean(yres[s])
totest=totrest+fulltelling
totest
rateest=sum(xrest)*mean(yres[s])/mean(xrest[s])
rateest
ratetot=rateest+fulltelling
ratetot
se=402*sqrt(var(yres[s])*(402-12)/(402*12))
se
CI=totest+qnorm(c(0.025,0.975))*se
CI
r=mean(yres[s])/mean(xrest[s])
ssqr=(1/11)*sum((yres[s]-r*xrest[s])^2)
ser=415*sqrt((mean(xrest)/mean(xrest[s]))^2*((402-12)/402)*ssqr/12)
ser
CIr=ratetot+qnorm(c(0.025,0.975))*ser
CIr
```

Deretter kopier R-koden 1 gang og lim inn i R 10 ganger. Får da umiddelbart resultatene.

Tabell 5 Ti 95 % konfidensintervaller i millioner kroner

Utvalgsnr	Ekspansjons-estimat (se i parentes)	KI basert på ekspansjonsestimat	Ratebasert estimat	KI basert på rate-estimat
1	2817 (801)	1248 – 4387	2858 (345)	2182 – 3533
2	3510 (857)	1831 – 5190	2895 (424)	2064 – 3725
3	2010 (712)	614 – 3405	2127 (170)	1794 – 2460*
4	2083 (397)	1305 – 2861*	2819 (318)	2196 – 3442
5	3200 (992)	1255 – 5146	2801 (311)	2191 – 3411
6	5375 (1684)	2073 – 8676	3657(383)	2906 – 4408
7	5211 (1121)	3014 – 7408	3815 (422)	2987 – 4642

8	3701 (1626)	514 – 6887	3874 (622)	2656 – 5092
9	1697 (266)	1175 – 2219*	2372 (164)	2051 – 2693*
10	2278 (647)	1009 – 3547	2988(620)	1774 – 4204

Rateestimatoren er klart mer presis enn ekspansjonsestimatoren og foretrekkes. Vi ser at den empiriske standardfeilen er mye mindre for rate-estimatoren, 588 mot 1293. Gjennomsnittlige verdier av estimatene er 3021 for rateestimaoren og 3188 for ekspansjonsestimatoren. Disse verdiene kan beregnes ved:

```
rt =c(2858,2895,2127,2819,2801,3657,3815,3874,2372,2988)
mean(rt)
[1] 3020.6
sqrt(var(rt))
[1] 588.3795
mid=c(2817,3510,2010,2083,3200,5375,5211,3701,1697,2278)
mean(mid)
[1] 3188.2
sqrt(var(mid))
[1] 1292.71
```

(c) I en kjøring av ti enkelt tilfeldige utvalg av 25 bedrifter ble den empiriske standardfeilen 1589 for ekspansjonsestimatoren og 730 for rate-estimatoren.
Standardfeilene for begge estimeringsmetodene er blitt mindre. Empirisk standardfeil i disse ti estimeringene for ekspansjonsestimatoren er her 1293 mot 1589 i rent ETU og for rateestimatoren er det 588 her mot 730 i rent ETU. Det virker ganske sikkert at utvalgsplanen (a) gir større treffsikkerhet enn rent ETU.

6.2 Fulltelling, cut-off og enkelt tilfeldig utvalg*

Igjen fulltelling for alle bedrifter med mer enn 50 sysselsatte, men nå skal ingen bedrifter med 5 eller færre sysselsatte være med. Trekker et enkelt tilfeldig utvalg blant resten med samme størrelse som i oppgave 1. Skal beregne estimater basert på rate- og ekspansjonsestimatorene, og gjenta det ti ganger. Skal sammenligne denne utvalgsplanen med utvalgsplanen i oppgave 1.

```
yrev=c(y3,y4,y5)
xrev=c(x3,x4,x5)
```

Vi ser at

```
sum(z==3)+sum(z==4)+sum(z==5)
[1] 121
```

Det betyr at populasjonstørrelsen på strataene 3,4,5 er 121 hvor vi skal ta et ETU på 12 bedrifter.

```
s=sample(121,12)
totrev=121*mean(yrev[s])
totrev
[1] 1720630
total=totrev+fulltelling
total
[1] 2838098
rateest=sum(xrev)*mean(yrev[s])/mean(xrev[s])
rateest
[1] 2115745
ratetot=rateest+fulltelling
ratetot
[1] 3233213
```

Merk at

```
sum(y6)
[1] 295651. # Det er 8,9% av totalen.
```

Litt vel mye til å bli fullstendig utelukket, kanskje.

Tabell 6 Ti estimater for total omsetning, i millioner kroner

Utvalgsnr	Ekspansjonsestimat	Ratebasert estimat
1	2838	3233
2	2735	2772
3	2714	2920
4	2791	2489
5	2391	3508
6	3068	2992
7	2694	2425
8	2431	2926
9	2005	2466
10	2412	2709

Ser at vi får generell underestimering, men estimatene varierer mye mindre enn før. Allikevel, her utelukkes for stor andel av populasjonen.

Appendiks A*. Utledning av resultatene for eksemplene i 2.6.

Spill med tre dører

Deltaker 1: Velger dør 1.
Deltaker 2: Velger dør 1 og bytter etter at TV-verten har åpnet en dør.

La B_i være begivenheten at bilen er bak dør i, i = 1, 2, 3. $P(B_i)$ =1/3 for i = 1, 2, 3. La D_i være begivenheten at TV-verten åpner dør i. Dersom bilen er bak dør 1 så vil hun vilkårlig velge dør 2 eller 3. Dette betyr at

$P(D_3|B_1)$ = P($D_2|B_1$) = ½ og $P(D_3|B_2)$ = P($D_2|B_3$) = 1.

La V være begivenheten at deltaker 2 vinner. Anta nå at dør 3 åpnes. Vi skal først se på den betingede sannsynligheten for at deltaker vinner gitt D_3. Dvs., hvor ofte vil deltaker 2 vinne blant de spill der dør 3 åpnes? Denne sannsynligheten er $P(V|D_3) = P(B_2|D_3)$. Vi kjenner: $P(D_3|B_1)$ = ½, $P(D_3|B_2)$ = 1 og $P(D_3|B_3)$ = 0. Her skal vi bruke Bayes formel:

$$P(B_2 \mid D_3) = \frac{P(B_2)P(D_3 \mid B_2)}{P(D_3)} = \frac{P(B_2)P(D_3 \mid B_2)}{P(B_1)P(D_3 \mid B_1) + P(B_2)P(D_3 \mid B_2)} = \frac{\frac{1}{3}}{\frac{1}{3}\cdot\frac{1}{2} + \frac{1}{3}} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

Dette gir også at P(deltaker 1 vinner|D_3) = $P(B_1|D_3)$ = 1 - $P(B_2|D_3)$ = 1/3.

Opprinnelig var vi interessert i å finne ut hvor ofte deltaker 2 vinner totalt, uansett hvilken dør som åpnet. Dvs., vi ønsker å finne den ubetingede sannsynligheten $P(V)$. Vi har da:

$P(V) = P(V \cap D_3) = P(V \cap D_2) = P(V \mid D_3)P(D_3) + P(V \mid D_2)P(D_2)$.
Siden $P(D_2) + P(D_3)$ = 1 og $P(D_3)$ = ½, så er $P(D_2)$ = ½.

Og: $P(V \mid D_2) = P(B_3 \mid D_2) = \frac{P(B_3)P(D_2 \mid B_3)}{P(D_2)} = \frac{\frac{1}{3}\cdot 1}{\frac{1}{2}} = \frac{2}{3}$.

Dermed: $P(V) = P(V \mid D_3)P(D_3) + P(V \mid D_2)P(D_2) = \frac{2}{3}\cdot\frac{1}{2} + \frac{2}{3}\cdot\frac{1}{2} = \frac{2}{3}$.

Vi vil få samme svar om vi lar deltakerne velge forskjellige dører ved hvert spill. For eksempel, hvis deltaker 2 hver gang velger dør i med sannsynlighet p_i, hvor $p_1+p_2+p_3$ = 1, så vil $P(V|D_i)$ = 2/3 og $P(V)$ =2/3 uansett hva p_1, p_2, p_3 er.

At $P(V)$ = 2/3 kan anskueliggjøres ved å sette opp alle mulige utfall av spillet for deltaker 2 i en tabell:

Tabell A1

Bil er bak dør nr.	Deltakers valg	TV-verten åpner dør nr.	Deltaker bytter	Deltaker vinner
1	1	2 eller 3	Fra 1 til 2 eller 3	Nei
1	2	3	Fra 2 til 1	Ja
1	3	2	Fra 3 til 1	Ja
2	1	3	Fra 1 til 2	Ja
2	2	1 eller 3	Fra 2 til 1 eller 3	Nei
2	3	1	Fra 3 til 2	Ja
3	1	2	Fra 1 til 3	Ja
3	2	1	Fra 2 til 3	Ja
3	3	1 eller 2	Fra 3 til 1 eller 2	Nei

Hvis $p_1=p_2=p_3$ = 1/3, så er alle utfall like sannsynlige og $P(V)$ = 6/9 = 2/3. Også når p_i - ene er forskjellige kan tabellen brukes for å få samme resultat. La (i,j) stå for begivenheten «Bil bak dør i og deltakere velger dør j». Da har vi:

$P(V)$ = P(1,2)+P(1,3)+P(2,1)+P(2,3)+P(3,1)+P(3,2)

$= \frac{1}{3}(p_2 + p_3 + p_1 + p_3 + p_1 + p_2) = \frac{1}{3}\cdot 2(p_1 + p_2 + p_3) = 2/3.$

Finaler i fotball

Anta to lag A, B med egenskapene:

P(A vinner) = p

P(B vinner) = q, $p > q$

P(uavgjort) = $1 - p - q$

Her er altså A det beste laget.

La n være antall kamper som spilles. Vi skal altså bestemme n slik at sannsynligheten for at A vinner flere kamper enn B er 0,95.

La X være antall A-seire i løpet av n kamper, Y antall B-seire og Z antall uavgjorte kamper. Vi antar at resultatene av de n er uavhengige, slik at de n kampene utgjør et multinomisk eksperiment.

Vi ønsker å bestemme n slik at $P(X - Y \geq 1) = 0{,}95$.

(X, Y, Z) multinomisk fordelt medfører at

$E(X) = np$, $V(X) = np(1\text{-}p)$

$E(Y) = nq$, $V(Y) = nq(1\text{-}q)$

$Cov(X,Y) = -\, npq$

$X - Y$ er tilnærmet normalfordelt med forventning og varians gitt ved:

$$E(X - Y) = n(p - q)$$

$$V(X - Y) = V(X) + V(Y) - 2Cov(X,Y)$$
$$= np(1 - p) + nq(1 - q) + 2npq$$
$$= n[p + q - (p - q)^2] = \sigma_n^2$$

La nå U være $N(n(p - q), \sigma_n)$. Da er (med kontinuitetskorreksjon) og Φ lik fordelingsfunskjonen til $N(0,1)$:

$$P(X - Y) \geq 1) \approx P(U \geq 0{,}5) = 1 - \Phi\left(\frac{0{,}5 - n(p - q)}{\sigma_n}\right) = \Phi\left(\frac{n(p - q) - 0{,}5}{\sigma_n}\right).$$

Det betyr at n skal bestemmes slik at:

$$\Phi\left(\frac{n(p - q) - 0{,}5}{\sigma_n}\right) = 0{,}95 \text{, dvs. } \frac{n(p - q) - 0{,}5}{\sigma_n} = 1{,}645 = z$$

Dette er ekvivalent med:

$$\frac{n(p - q) - 0{,}5}{\sqrt{n}} = \sqrt{p + q - (p - q)^2} \cdot 1{,}645$$
$$\Leftrightarrow \frac{[n(p - q) - 0{,}5]^2}{n} = [p + q - (p - q)^2] \cdot 1{,}645^2$$
$$\Leftrightarrow n(p - q)^2 - (p - q) + \tfrac{1}{4n} = [p + q - (p - q)^2] \cdot 1{,}645^2$$

Siden $1/4n = 0$, så får vi:

$$n = \frac{1}{p - q} + [\frac{p + q}{(p - q)^2} - 1] \cdot 1{,}645^2 \qquad \text{(I)}$$

Bestemmer da n som minste hele tall større eller lik (I). Vi ser at n øker når $p - q$ avtar, også når $p + q$ øker. Når $p + q = 1$ så bestemmes i tillegg n som et odde tall.

Tabell A2. Verdier av *n* for utvalge verdier av *p* og *q*

p - q	*p*	*q*	*n*
0,60	0,7	0,1	5
0,60	0,8	0,2	7
0,50	0,7	0,2	9
0,50	0,75	0,25	11
0,40	0,6	0,2	14
0,40	0,7	0,3	17
0,35	0,55	0,2	17
0,35	0,65	0,3	22
0,30	0,5	0,2	22
0,30	0,65	0,35	31
0,25	0,50	0,25	34
0,25	0,60	0,35	43
0,20	0,50	0,30	57
0,20	0,60	0,40	71
0,10	0,40	0,30	197
0,10	0,55	0,45	279

Appendiks B. Functions most commonly used in R

Below is a summary of the functions most commonly used in R with short descriptions (see references). You can use help (function) or ? function to get more information about the usage of the function specified. Please see more detailed explanations about commonly used functions in R in the '**R Reference Card'**, which can be downloaded from here: https://cran.r-project.org/doc/contrib/Short-refcard.pdf

General functions

```
help()                          # Opens help page
library(packageName)            # Loads the package specified
builtins()                      # List all built-in functions
?functionName                   # Helps on the package specified
ls()                            # List objects in the workspace
rm(x)                           # Removes x from the workspace
rm(list=ls())                   # Remove all the variables from the workspace
Sys.time()                      # Return system time
Sys.Date()                      # Return system date
getwd()                         # Return working directory
setwd()                         # Set working directory
list.files()                    # List files in a given directory
file.info()                     # Get information about files
attach(data)                    # Makes the names of the variables in a matrix or data frame available in the
                                # workspace
detach(data)                    # Releases the names (remember to do this each time you attach something
append()                        # Add elements to a vector
c(x)                            # A generic function which combines its arguments
                                x = c(1,2,4,8,16 )# create a data vector with specified elements
                                y = c(1:10)# create a data vector with elements 1-10
                                x1 = c(rnorm(n))# create a n item vector of random normal deviates
                                y1 = c(runif(n))+n   # create another n item vector that has n added to
                                # each random uniform distribution
                                vect = c(x,y) # combine them into one vector
cat(x)                          # Prints the arguments
cbind()                         # Combine vectors by row/column
                                mat = cbind(x,y)
identical()                     # Test if 2 objects are *exactly* equal
length(x)                       # Return no. of elements in vector x
mat() or vec()                  # Create a matrix or vector
data[4,2]                       # display the 4th row and the 2nd column
data[3,]                        # display the 3rd row
data[,2]                        # display the 2nd column
data[,-j]                       # drop the jth column
data[-i,]                       # drop the ith row
data[,-c(i,j)]                  # drop the ith and jth column
data[n1:n2,n3:n4]               # select the n1 through n2 rows of variables n3 through n4
sort(x)                         # Sort the vector x
order(x)                        # list sorted element numbers of x
data[order(data$B),]            # sort a data frame by the order of the elements in B
data[rev(order(data$B)),]               # Sort the data frame x in reverse order
tolower(),toupper()             # Convert string to lower/upper case letters
unique(x)                       # Remove duplicate entries from vector
substr(x, start=n1, stop=n2)            # Extract or replace substrings in a character vector.
                                        x = "abcdef"
                                        substr(x, 2, 4) is "bcd"
                                        # substr(x, 2, 4) ="22222" is "a222ef"
strsplit(x, split)              # Split the elements of character vector x at split.
strsplit("abc", "")             # returns 3 element vector "a","b","c"
paste(..., sep="")              # Concatenate strings after using sep string to seperate them.
                                # paste("x",1:3,sep="") returns c("x1","x2" "x3")
```

```
                        # paste("x",1:3,sep="M") returns c("xM1","xM2" "xM3")
                        # paste("Today is", date())
toupper(x)                      # Uppercase
tolower(x)                      # Lowercase
seq(from, to, by)               # generate a sequence
                        # indices 0 seq(1,10,2)
                        # indices is c(1, 3, 5, 7, 9)
rep(x, ntimes)          # repeat x n times
                        y = rep(1:3, 2)# y is c(1, 2, 3, 1, 2, 3)
cut(x, n)               # divide continuous variable in factor with n levels
                      y = cut(x,5)
x%in%y                  # tests each element of x for membership in y
all(x%in%y)             # true if x is a proper subset of y
all(x)                  # for a vector of logical values, are they all true?
any(x)                  # for a vector of logical values, is at least one true?
colSums (x, na.rm = FALSE, dims = 1)        # Form column sums for numeric array
rowSums (x, na.rm = FALSE, dims = 1)        # Form row sums for numeric array
colMeans(x, na.rm = FALSE, dims = 1)        # Form column means for numeric array
rowMeans(x, na.rm = FALSE, dims = 1)        # Form row means for numeric array
rowsum(x, group, reorder = TRUE, ...)       # finds row sums for each level of a grouping variable
apply(x,1,min)          # finds the minimum for each row
apply(x,2,max)          # finds the maximum for each column
which.min(x)            # prints the observation number that has the minimum value
which.max(x)            # prints the observation number that has the maximum value
apply(x,1,which.min)    # tells the row with the minimum value for every column
```

Mathematical functions

You can use the option na.rm, wherever available, to exclude missing values before numerical calculations. Otherwise the presence of missing values will lead to a missing result.

```
abs(x)                  # absolute value
sqrt(x)                 # square root
ceiling(x)              # ceiling(3.475) is 4
floor(x)                # floor(3.475) is 3
trunc(x)                # trunc(5.99) is 5
round(x, digits=n)      # round(3.475, digits=2) is 3.48
signif(x, digits=n)     # signif(3.475, digits=2) is 3.5
cos(x), sin(x), tan(x)  # also acos(x), cosh(x), acosh(x), etc.  # trigonometric functions
log(x)                  # natural logarithm
log10(x)                # common logarithm
exp(x)                  # e^x
sign(x)                 # Returns the signs of the elements of x
sum(x)                  # mean and sum of object x
rev(x)                  # reverse the order of values in x
t(X)                    # transpose of matrix X
X %*% Y                 # matrix multiply X by Y
solve(A)                # inverse of A
solve(A,B)              # inverse of A * B
```

Statistical functions

```
help(package=stats)     # List all stats functions i.e. ks.test(), ?Chisquare, etc.
mean(x, trim=0, na.rm=FALSE) # mean of object x, removing any missing values
                        # trimmed mean, removing any missing values and 5 percent of highest and lowest
                        scores: mx = mean(x,trim=.05,na.rm=TRUE)
weighted.mean(x)        # weighted mean
sd(x), var(x), mad(x)   # standard deviation,  variance and median absolute deviation of x
min(x), max(x), range(x)# minimum, maximum and range
median(x)               # median
quantile(x, probs)      # quantiles where x is the numeric vector whose quantiles are
```

```
                        # desired and probs is a numeric vector with probabilities in [0,1].
                        y = quantile(x, c(.3,.84)) #30th and 84th percentiles of x
mad()                   # calculate median absolute deviation
summary(x)              # Returns a summary of x: mean, min, max etc.
summary.factor(x)       # Returns a frequency table of all categories of a factor variable x
table(x)                # frequency counts of entries, ideally the entries are factors
                        # (although it works with integers or even reals)
scale(data,scale=FALSE) # centers around the mean but does not scale by the sd
cor(x,y,use="pair")     # correlation matrix for pairwise complete data, use="complete" for
                        # complete cases
cor.test()              # Perform correlation test
cumsum(); cumprod()     # Cumulative sum and product for vectors
cummin(); cummax()      # Cumulative minimum and maximum for vectors
density(x)              # Compute kernel density estimates
t.test()                # Student's t-test
pairwise.t.test(x,g)    # Calculate pairwise comparisons between group levels
aov(x~y,data=datafile)  # where x and y can be matrices. Do the analysis-of-variance
aov(x~y*z,data=datafile)  # do a two way analysis of variance
summary(aov(x~y,data=datafile) )    # show the summary table
sample(x, size, replace = FALSE, prob = NULL)   # samples with or without replacement
ecdf()                  # Empirical Cumulative Distribution Function
qqplot()                # quantile-quantile plot
lm(x~y,data=dataset)    # basic linear model
```

Statistical probability functions

For random number generators below, you can use `set.seed(1234)` or some other integer to create reproducible pseudo-random numbers.

```
dnorm(x)        # normal density function (by default m=0 sd=1). Plot standard normal curve
                x = pretty(c(-3,3), 30)
                y = dnorm(x)
                plot(x,y,type='l',xlab="Normal Deviate",ylab="Density", yaxs="i")
pnorm(q)        # cumulative normal probability for q (area under the normal curve to the left of q)
                # pnorm(1.96) is 0.975
qnorm(p)        # normal quantile. Value at the p percentile of normal distribution
                # qnorm(.9) is 1.28, 90th percentile
rnorm(n, m=0,sd=1)    # n random normal deviates with mean m and standard deviation sd.
                      # 50 random normal variates with mean=50, sd=10:
                      x = rnorm(50, m=50, sd=10)
dbinom(x, size, prob) # binomial distribution where size is the sample size
                      # prob is the probability of a head (pi)
                      # prob of 0 to 5 heads of fair coin out of 10 flips: dbinom(0:5,10,.5)
pbinom(q, size, prob) # prob of 5 or less heads of fair coin out of 10 flips: pbinom(5, 10, .5)
qbinom(p, size, prob)
rbinom(n, size, prob)
dpois(x, lamda)       # poisson distribution with m=std=lamda
                      # probability of 0,1, or 2 events with lamda=4: dpois(0:2, 4)
ppois(q, lamda)       # probability of at least 3 events with lamda=4: 1-ppois(2,4)
qpois(p, lamda)
rpois(n, lamda)
dunif(x, min=0, max=1) # uniform distribution, follows the same pattern as the normal distribution
punif(q, min=0, max=1)
qunif(p, min=0, max=1)
runif(n, min=0, max=1)#10 uniform random variates: x = runif(10)
```

Graphical functions

```
help(package=graphics) # List all graphics functions
hist()                 # histogram
plot()                 # Generic function for plotting of R objects
```

```
boxplot()           # graphical summary appears in graphics window
par()               # Set or query graphical parameters
                    par(mfrow=c(1,2))# divides the graphics window into two parts vertically so that
                    two graphs can be viewed in one row next two each other
title( "some title") # add a title to the first graph
curve(5*x^3,add=T)  # Plot an equation as a curve
points(x,y)         # Add another set of points to an existing graph
arrows()            # Draw arrows [see errorbar script]
abline()            # Adds a straight line to an existing graph
lines()             # Join specified points with line segments
segments()          # Draw line segments between pairs of points
hist(x)             # Plot a histogram of x
pairs()             # Plot matrix of scatter plots
identify()          # reads the position of the graphics pointer when the (first) mouse button is pressed
legend()            # add legends to plots
matplot()           # Plot columns of matrices
?device             # Help page on available graphical devices
postscript()        # Plot to postscript file
pdf()               # Plot to pdf file
png()               # Plot to PNG file
jpeg()              # Plot to JPEG file
X11()               # Plot to X window
persp()             # Draws perspective plot
contour()           # Contour plot
image()             # Plot an image
```

Operators used in R

Example:

```
x = c(1:10) # yields 1 2 3 4 5 6 7 8 9 10
x > 8 # yields F F F F F F F F T T
x > 8 | x < 5       # yields T T T T F F F F T T
x[(x>8) | (x<5)] # yields 1 2 3 4 9 10
x[c(T,T,T,T,F,F,F,F,T,T)] # yields 1 2 3 4 9 10
```

Arithmetic operators

```
+               # addition
-               # subtraction
*               # multiplication
/               # division
^ or **         # exponentiation
x %% y#         modulus (x mod y) 5%%2 is 1
x %/% y         # integer division 5%/%2 is 2
```

Logical operators

```
<               # less than
<=              # less than or equal to
>               # greater than
>=              # greater than or equal to
==              # exactly equal to
!=              # not equal to
!x              # Not x
x | y           # x OR y
x & y           # x AND y
isTRUE(x)       # test if X is TRUE
```

Useful links to know more about R

https://cran.r-project.org/

http://www.statmethods.net/index.html

https://www.r-project.org/

http://www.r-bloggers.com/

References on R

http://www.statmethods.net/management/functions.html

http://www.sr.bham.ac.uk/~ajrs/R/r-function_list.html

http://personality-project.org/r/r.commands.html

Figurregister

Tabellregister

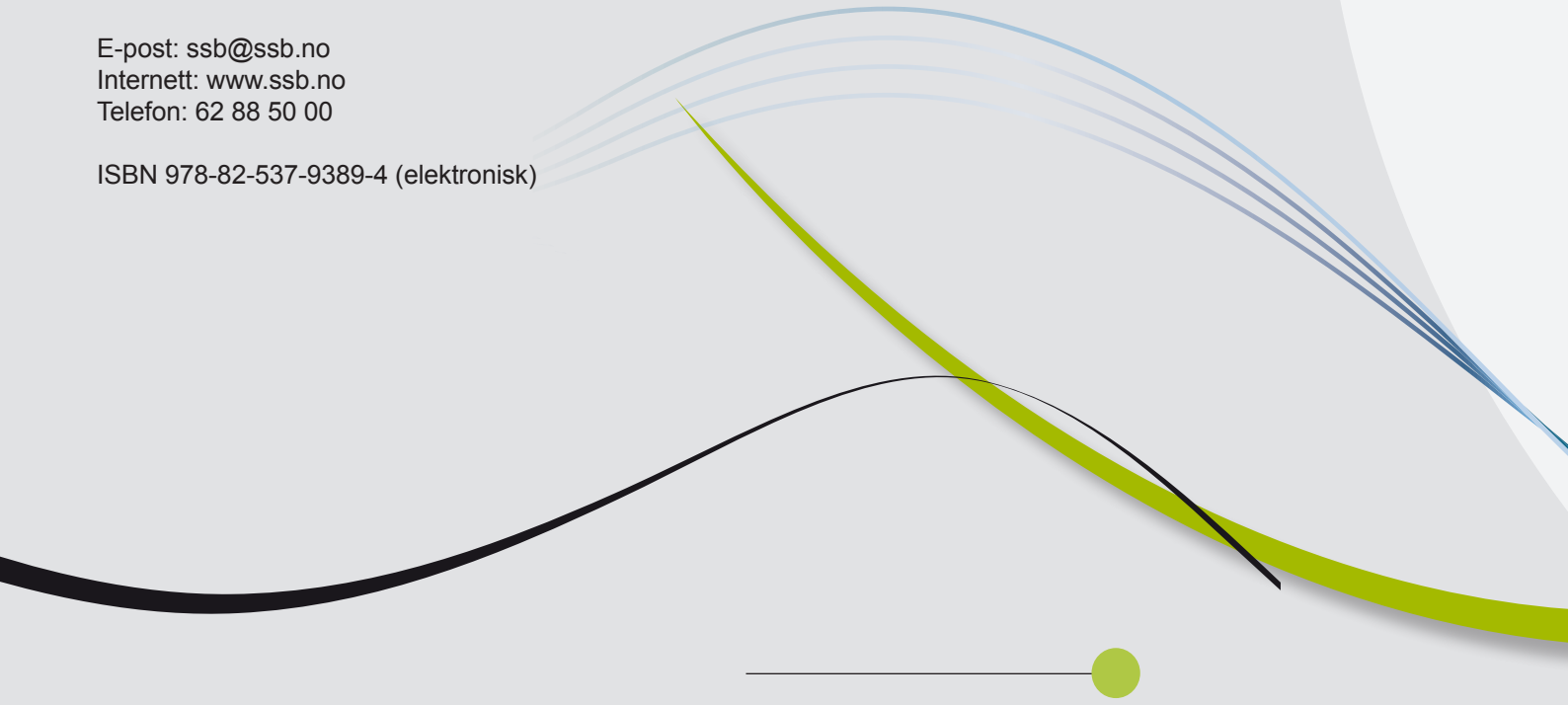
Statistisk sentralbyrå

Postadresse:
Postboks 8131 Dep
NO-0033 Oslo

Besøksadresse:
Akersveien 26, Oslo
Oterveien 23, Kongsvinger

E-post: ssb@ssb.no
Internett: www.ssb.no
Telefon: 62 88 50 00

ISBN 978-82-537-9389-4 (elektronisk)



Statistisk sentralbyrå
Statistics Norway

Design: Siri Boquist