Statistics Norway
Statistisk sentralbyrå

Jan F. Bjørnstad

En innføring i utvalgsundersøkelser

Korrigert versjon

Avdeling for samordning og utvikling/Seksjon for statistiske metoder og standarder
Emnegruppe: 00.90

Forord

Formålet med dette kompendiet er å gi en praktisk, ikkematematisk, innføring i

- planlegging av utvalgsundersøkelser
- begrepsapparatet og de viktigste statistiske estimeringsmetodene i utvalgsundersøkelser
- anvendelser av statistiske metoder for å handtere frafall (dvs. at det mangler opplysninger fra enkelte enheter i utvalget) som kan resultere i skjeve utvalg.

Matematiske utledninger og formler er forsøkt holdt til et minimum, men noe formelbruk er uunngåelig hvis man skal oppnå en viss basis innsikt, kunnskap og forståelse av utvalgsundersøkelser. For å helpe til med tilegnelse og fordypning inkluderes oppgaver, med løsninger. De mer teoretiske oppgavene er stjernemerket. For å kunne tilegne seg det statistiske språket og begrepsapparatet er det nødvendig å kjenne til litt elementær sannsynlighetsteori. Et appendiks er tatt med for de som trenger en repetering eller innføring i sannsynlighet.

Innhold

1. Noen grunnbegreper i utvalgsundersøkelser

1.1 Utvalgsundersøkelser - Innledning

Populasjon

Utgangspunktet er en *populasjon* U bestående av N *enheter*, hvor N er kjent. Vi er interessert i informasjon om denne populasjonen med hensyn til en eller flere variable. Formålet kan ,f.eks., være å anslå totalene til variablene av interesse.

Eksempel 1.1. Vi er interessert i å undersøke partitilhørighet for personer i Norge som er 18 år og eldre.Populasjonen er da alle personer i Norge som er minst 18 år gamle. Enheten er «person». Variabelen av interesse, for hver enhet i populasjonen, er partitilhørighet. Formålet er å si noe om andelene for de forskjellige partiene.

Eksempel 1.2. Problemet er å anslå antall sysselsatte i Norge. Populasjonen er personer i Norge over 15 år. For hver person er variabelen av interesse om personen er sysselsatt eller ikke. Denne kan formuleres slik: For person i, la $y_i = 1$ hvis personen er sysselsatt og 0 ellers. Antall sysselsatte er da lik summen av alle y_i i populasjonen, dvs. totalen til denne variabelen.

Eksempel 1.3. Vi ønsker å undersøke inflasjon og arbeidsledighet i EU og EØS. Populasjonen består da av alle land i EU og EØS. Enheten er «land». Formålet kan være å anslå gjennomsnittlig inflasjon og prosent arbeidsledighet.

Eksempel 1.4. I forbindelse med levekårsundersøkelsen i Statistisk sentralbyrå er vi interessert i å anslå gjennomsnittlig timelønn for ansatte fordelt på kjønn, landsdel og type bosted (etter bostedstetthet).

Alle enhetene i populasjonen kan identifiseres og er nummerert fra 1 til N. Dvs., vi kan betegne populasjonen symbolsk med $U = \{1,2, \ldots ,N\}$.

Utvalg

Vanligvis er det for kostbart og tidkrevende å observere alle enhetene i populasjonen. Vi tar da et *utvalg* s av enheter fra populasjonen som vi observerer, dvs. verdiene av de valgte variablene i undersøkelsen «måles» for enhetene i s. Tilgang til populasjonen er ved et *trekkeregister* U_F . Vi skal anta at U_F er en tilfredsstillende representasjon for U, dvs. $U = U_F$.

I eksemplene 1.1 og 1.2 består målingene av å spørre personene i utvalget. I eksempel 1.3 består målingene i å beregne inflasjonen og arbeidsledigheten i de utvalgte landene. Her kunne det tenkes at hele populasjonen ble valgt. Isåfall har vi en *fulltelling*. I eksempel 1.4 er målingen, for en ansatt, en beregning av vedkommendes timelønn.

Vi kaller s et *sannsynlighetsutvalg* hvis vi trekker enheter til utvalget tilfeldig med visse sannsynligheter.

Eksempel- enkelt tilfeldig utvalg. La n være størrelsen på utvalget, som er bestemt på forhånd. (Vi skal senere se på noen kriterier for valg av n.) Vi trekker nå en og en enhet *uten tilbakelegging* , dvs. en person kan bare trekkes en gang. Ved første trekking lar vi alle enhetene i populasjonen ha samme sannsynlighet for å bli valgt. Etter at første enhet er valgt, velges neste slik at alle N-1 gjenværende

enheter har samme sannsynlighet for å bli trukket ut osv., til n enheter er trukket ut til utvalget. En slik sannsynlighetsplan for å velge utvalget kalles *utvalgsplanen*.

Estimering

En undersøkelse vil vanligvis dreie seg om mange variable. Formålet med å trekke et utvalg fra populasjonen er å få informasjon om funksjoner av disse variable. La y betegne en av disse variable. Til hver enhet i i populasjonen er det tilordnet et tall y_i som er verdien av y for denne enheten. Vanligvis er vi interessert i å anslå eller *estimere* totalen t i populasjonen, dvs. summen av alle y-verdiene eller gjennomsnittet, som er t delt på antallet N , basert på det utvalget vi observerer.

Eksempel 1.4, forts. Anta at utvalget i et spredtbygd område bestod av 8 personer, hvorav 2 ikke ga opplysninger om timelønn og de resterende 6 ga følgende verdier : 125, 87, 95, 160, 141 og 112 med gjennomsnittsverdi 120. Kr. 120 er da et *estimat* for gjennomsnittlig timelønn for dette området.

Feilkilder

1. Trekkeregisteret U_F . Utvalget trekkes fra U_F . Det kan tenkes at U_F ikke er en korrekt representasjon av populasjonen. F.eks., et bedriftsregister er vanligvis ikke en korrekt beskrivelse av en populasjon av bedrifter. F.eks., nye bedrifter er ikke kommet med i U_F (underdekning) og bedrifter som er nedlagt kan fremdeles være med i U_F (overdekning).

2. Frafall. I eksemplet ovenfor ser vi et problem som vi ofte har ; manglende opplysninger for enkelte enheter i utvalget. Et viktig problem å undersøke er om et slikt frafall av observasjoner medfører at utvalget blir skjevt slik at estimatet blir mer usikkert.

3. Målefeil. Nå utvalget observeres kan det oppstå feil, ved at vi ikke måler den riktige verdien av y_i. For eksempel, i intervju-undersøkelser kan det være flere årsaker til svarfeil: feil avmerking, intervjuerpåvirkning, sosial ønskelige svar (eks.: underrapportering av alkohol og tobakk forbruk), manglende kunnskaper, dårlig hukommelse, misforstår spørsmålet.

4. Utvalgsfeil . Siden vi observerer bare en del av populasjonen vil det alltid være knyttet en viss usikkerhet til estimatene våre. Dette kurset vil konsentrere seg om denne feilkilden og se på hvordan denne usikkerheten kan tallfestes. Blant annet skal vi se på en metode for å beregne et intervallestimat som er slik at vi er «rimelig sikker» på at den sanne verdien (f.eks. populasjonsgjennomsnittet) ligger innenfor intervallet. Et slikt intervall kalles et *konfidensintervall.* I eksempel 1.4 vil et 95% konfidensintervall for gjennomsnittlig timelønn være $120 \pm 22 = (98, 142)$. Dette betyr at vi er 95% sikker på at den faktiske gjennomsnittlige timelønnen (i området) ligger mellom 98 og 142.

De tre ikke-utvalgsfeilene, frafall, målefeil,register, kan være betraktelige, faktisk atskillig større enn utvalgsfeilen. Vi skal i kurset se på hvordan skjevheter på grunn av frafall kan rettes opp.

Oppsummering av begreper
- populasjon
- enhet
- utvalg
- utvalgsplan
- estimering
- feilkilder i utvalgsundersøkelser : register, målefeil, frafall,utvalgsfeil

For å kunne tilegne seg det *statistiske språket* og *begrepsapparatet* for utvalgsundersøkelser er det nødvendig å kunne litt om elementær sannsynlighetsteori. Appendiks A er tatt med for de som trenger å friske opp og/eller tilegne seg denne teorien.

Begreper som er fundamentale i utvalgsteorien og som trenger en presis oppbygging gjelder:

- Utvalgsplan; sannsynlighetsutvalg, trekkesannsynlighet for de enkelte enhetene i populasjonen

- Estimering; estimator , mål for skjevhet, mål for usikkerhet, konfidensintervall

For å kunne gjennomgå disse begrepene trenger vi først, i kapittel 1.2, en oppsummering av to av de viktigste sannsynlighetsfordelingene i utvalgsteorien. For en grundigere gjennomgang henviser vi til Appendiks A, kap. 2.

I kapittel 1.3 gis en kort gjennomgang av estimeringsbegrepene anvendt på utvalgsundersøkelser, og i kapittel 1.4 gis en generell beskrivelse av utvalgsplaner.

1.2 Hypergeometriske og binomiske fordelinger

Enkelt tilfeldig utvalg- hypergeometrisk fordeling

La s betegne de enhetene som blir trukket til utvalget. Det kan vises at hvert utvalg s av størrelse n har samme sannsynlighet for å bli valgt, og dette er også den vanlige definisjonen på *enkelt tilfeldig utvalg :*

DEFINISJON. Vi har et *enkelt tilfeldig utvalg* av størrelse n, hvis alle utvalg s av størrelse n har samme sannsynlighet for å bli valgt.

Anta vi er interessert i å finne ut noe om antallet M i populasjonen med et visst kjennetegn (egenskap) A, f.eks. antall sysselsatte. La p være andelen med kjennetegn A i populasjonen slik at p er lik M delt på N, dvs. $p = M/N$. La X være antall med kjennetegn A i utvalget, slik at X/n er et anslag, dvs. estimator, for p. La M være antall med kjennetegn A i populasjonen, X kan ta verdiene 1, ... n med sannsynligheter bestemt ved den den *hypergeometriske* fordelingen. Vi sier at X er hypergeometrisk fordelt. Vi henviser til Appendiks A, kapittel 2.1 for formelen til denne fordelingen.

Et eksempel er lottospillet.

Eksempel 1.5. Lottospillet. Populasjonen består av $N = 34$ tall , og vi velger $M = 7$ tall. Så trekkes tilfeldig $n = 7$ tall som «korrekte». Vi er interessert i kjennetegn A : « tallet er på vår lottokupong». La X = antall korrekte tall på vår kupong. X er da hypergeometrisk fordelt med N,M,n gitt ovenfor. Fra Appendiks A, eksempel 2.2,

k	7	6	5	4
P($X = k$)	0,00000019	0,000035	0,0014	0,0190

For eksempel, P($X \geq 4$) = 0,020.

Trekking med tilbakelegging - binomisk fordeling

Vi trekker nå en og en enhet *med tilbakelegging* , dvs. en og samme person kan trekkes flere ganger. Ved hver trekking lar vi alle enhetene i populasjonen ha samme sannsynlighet $1/N$ for å bli valgt.

Vi er fremdeles interessert i å finne ut noe om antallet i populasjonen med et visst kjennetegn (egenskap) A. Som under enkelt tilfeldig utvalg skal vi se på sannsynlighetsfordelingen til variabelen X = antall med kjennetegn A i utvalget. Som før, la M være antall med kjennetegn A i populasjonen, slik at $p = M/N$ er populasjonsandelen med egenskapen A. Trekningene av enhetene er nå uavhengige. Dette medfører at X får en annen sannsynlighetsfordeling. Den kalles den *binomiske* fordelingen. Igjen henviser vi til Appendiks A, kapittel 2.1 for formelen til fordelingen.

Mer generelt så forekommer den binomiske fordelingen i de såkalte binomiske forsøk som kan beskrives på følgende måte:

Binomiske forsøk

(1) Består av n uavhengige enkeltforsøk med to mulige utfall, begivenhet A eller A^c .

(2) p = sannsynligheten for at A skal inntreffe, $p = P(A)$, er den samme for alle forsøkene.

A kalles gjerne «suksess», A^c «fiasko», for å ha en generell betegnelse .

La X være antall suksesser i de n forsøkene. Da er X binomisk fordelt.

Forventning og varians

Forventning

To sentrale egenskaper til en sannsynlighetsfordeling er forventningen og variansen. Forventningen er et mål for midtpunktet i fordelingen. Den forteller oss verdien av X i gjennomsnitt ved mange gjentatte observasjoner av X . Forventningen til X betegnes med $E(X)$ og defineres ved:

$$E(X) = \sum_x xP(X = x) .$$

Dvs., $E(X)$ = summen av verdi multiplisert med sannsynlighet, og betegnes vanligvis med den greske bokstaven μ (my).

Varians

Variansen til X, $Var(X)$, er et mål for spredningen i fordelingen rundt $\mu = E(X)$. $Var(X)$ gir gjennomsnittlig verdi av $(X - \mu)^2$ ved mange gjentatte observasjoner av X, og er definert ved:

$$Var(X) = E(X - \mu)^2 = \sum_x (x - \mu)^2 P(X = x) .$$

Dvs., $Var(X)$ = summen av $(x-\mu)^2$ multiplisert med sannsynligheten for verdien x. Notasjon, $\sigma^2 = Var(X)$ (σ er den greske bokstaven sigma).

Variansen måles i kvadratet av måle-enheten til X. For å uttrykke spredningen i samme enhet så brukes kvadratroten til $Var(X)$, kalt standardavviket til X, betegnet med $sd(X)$:

$$\sigma = sd(X) = \sqrt{Var(X)} .$$

For binomisk og hypergeometrisk fordeling gjelder følgende:

Binomisk fordeling: $E(X) = np$ (antall forsøk multiplisert med suksess sannsynlighet)
$Var(X) = np(1-p)$

Hypergeometrisk fordeling : $E(X) = np$

$$Var(X) = np(1\text{-}p)\frac{N-n}{N-1} \approx np(1-p)\left(1-\frac{n}{N}\right)$$

Legg merke til at når *n/N* er liten så er $np(1-p)\left(1-\frac{n}{N}\right) \approx np(1\text{-}p)$, den binomiske variansen. Det betyr at når *n/N* er liten så er trekking med og uten tilbakelegging omtrent det samme, sannsynlighetsteoretisk .

1.3 Estimeringsteori i utvalgsundersøkelser

Enkelt tilfeldig utvalg

Betrakt enkelt tilfeldig utvalg. La oss se på estimering av populasjonens middelverdi for en variabel *y*, $\mu = t / N$, hvor *t er* totalsummen av alle *y*-verdiene i populasjonen. Vi bruker betegnelsen μ siden $\mu = E(X)$ hvor X er verdien til y ved tilfeldig trekking av en enhet (oppgave 1.3).

En naturlig *estimator* er gjennomsnittsverdien i utvalget. La oss betegne den med $\bar{y}_s = \sum_{i \in s} y_i \, /n$, hvor *s* er utvalget. Uttrykket $\sum_{i \in s} y_i$ står for summen av y-verdiene i utvalget *s.* Vi skal studere egenskapene til estimatoren. Betrakt eksempel 1.4 igjen.

Eksempel 1.4. Her er $\sum_{i \in s} y_i = 720$, og $\bar{y}_s = 720/6 = 120$. Som nevnt tidligere kalles verdien 120 av estimatoren for *estimatet*.

Det er to aspekter ved en estimator $\hat{\mu}$ som er viktige. For det første så bør estimatoren gi omtrent korrekt verdi μ ved gjentatte forsøk, dvs. ved gjentatte utvalg av størrelse *n* så bør gjennomsnitts-verdien av $\hat{\mu}$ bli omtrent lik μ. Presist uttrykt :

$$E(\hat{\mu}) = \mu .$$

Vi sier at $\hat{\mu}$ er *forventningsrett.* Sagt på en annen måte, så betyr dette at fordelingen til den tilfeldige variable $\hat{\mu}$ er sentrert rundt μ. Denne fordelingen er fordelingen av $\hat{\mu}$´s verdier over uendelig mange gjentatte utvalg (engelsk: sampel), og kalles derfor også *samplingfordelingen* til $\hat{\mu}$. Vi skal senere vise at $\bar{y}_s$ er forventningsrett.

I tillegg til at estimatoren «treffer målet» μ gjennomsnittlig ved mange gjentatte utvalg , så er det viktig at *variasjonen* av estimatorens verdier ved gjentatte utvalg ikke er for stor slik at vi kan ha tillit til at i et gitt utvalg så blir forskjellen mellom estimatorens verdi, *estimatet*, og μ ikke for stor. Denne variasjonen kan måles ved standardavviket eller ekvivalent variansen til $\hat{\mu}$, $Var(\hat{\mu})$.

Vi kan dermed oppsummere kravene til en estimator $\hat{\mu}$ for μ :

(I) Forventningsretthet, iallfall tilnærmet : $E(\hat{\mu}) = \mu$

(II) Liten $Var(\hat{\mu})$.

Anta vi er interessert i å estimere andel i populasjonen med et visst kjennetegn (egenskap) A, f.eks. andel sysselsatte. Dette tilsvarer en *y*-variabel som er lik 1 hvis enheten har kjennetegnet A og 0 ellers. Da blir $\mu = p$, andelen med kjennetegn A i populasjonen. La X = antall med kjennetegn A i utvalget. Da er X hypergeometrisk fordelt, og en estimator for p er $\hat{p} = X/n$. Vi ser at $\hat{p} = \bar{y}_s$. Vi skal nå undersøke egenskapene til denne estimatoren. Fra kapittel 1.2 :

$$E(X) = np$$

$$Var(X) = np(1-p)\frac{N-n}{N-1}.$$

Dette gir at

$$E(\hat{p}) = \frac{1}{n}E(X) = p \; : \; \hat{p} \text{ er forventningsrett.}$$

$$Var(\hat{p}) = \frac{1}{n^2}Var(X) = \frac{p(1-p)}{n}\cdot\frac{N-n}{N-1}.$$

Siden p er ukjent er $Var(\hat{p})$ også ukjent. For å få et mål på usikkerheten til estimatoren så kan vi estimere $Var(\hat{p})$ å sette inn $\hat{p}$ for p :

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

Det estimerte standardavviket til $\hat{p}$ kalles *standardfeilen* til $\hat{p}$, og betegnes med $SE(\hat{p})$, (*SE* for det engelske uttrykket «standard error»),

$$SE(\hat{p}) = \sqrt{\hat{Var}(\hat{p})} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}\cdot\frac{N-n}{N-1}} \; .$$

Estimeringsfeilen vi begår er forskjellen mellom $\hat{p}$ og p, uten å ta hensyn til fortegnet , $|\hat{p}-p|$. Denne er selvfølgelig ukjent, men vi kan gi et anslag på hvor stor den maksimalt kan forventes å være, nemlig $2 \cdot SE(\hat{p})$, kalt *feilmarginen.*

Eksempel 1.6. Politisk meningsmåling. Anta vi tar et enkelt tilfeldig utvalg på 1500 personer fra den voksne befolkningen i Norge, dvs. alder minst 18 år, og spør hvilket parti de ville stemme på i det neste Stortingsvalget. Anta 490 svarte Arbeiderpartiet og 295 svarte Høyre. De estimerte befolkningsandelene for de to partiene blir da

$$\hat{p}_A = 490 / 1500 = 0{,}327, \text{ dvs. } 32{,}7\%$$

$$\hat{p}_H = 295 / 1500 = 0{,}197, \text{ dvs. } 19{,}7\% \; .$$

Vi har at for begge partiene så er tilnærmet $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Feilmarginene på estimatene ovenfor blir da $2 \cdot SE(\hat{p}_A) = 0{,}024$ (2,4%) og $2 \cdot SE(\hat{p}_H) = 0{,}021$ (2,1%). (Med $N = 3.000.000$ så er faktoren $\sqrt{\frac{N-n}{N-1}} = 0{,}99975$ og de eksakte standardfeilene vil være de samme til 4 desimaler.)

Det er ikke bare populasjonens middeltall μ som kan være av interesse å estimere. Vi skal senere se at vi også trenger å estimere *populasjonens varians,* som er gjennomsnittet av alle $(y_i - \mu)^2$:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 \quad .$$

Vi bruker σ^2 som notasjon siden dette kan betraktes som $Var(X)$, hvor X er verdien på y ved tilfeldig trekking av en enhet (se oppgave 1.3). Det kan vises at en forventningsrett estimator for σ^2 er gitt ved :

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

Her står uttrykket $\frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y}_s)^2$ for summen av alle $(y_i - \bar{y}_s)^2$ i utvalget s delt på $(n - 1)$.

Når $y_i = 1$ hvis kjennetegn A og 0 ellers, så blir $\sigma^2 = p(1-p)$, hvor p er andel med A i populasjonen (se oppgave 1.3). En naturlig estimator er, som før nevnt, $\hat{p}(1-\hat{p})$. Den er imidlertid ikke eksakt forventingsrett for $p(1-p)$ selv om $\hat{p}$ er forventningsrett for p. For å finne den forventningsrette $\hat{\sigma}^2$ ser vi først at $\sum_{i \in s}(y_i - \bar{y}_s)^2$ summerer verdien $(1-\hat{p})^2$ x ganger, og verdien $(0-\hat{p})^2 = \hat{p}^2$ $(n - x)$ ganger:

$$\sum_{i \in s}(y_i - \bar{y}_s)^2 = x(1-\hat{p})^2 + (n-x)\hat{p}^2 = n\hat{p}(1-\hat{p})^2 + n(1-\hat{p})\hat{p}^2$$

$$= n\hat{p}(1-\hat{p})(1-\hat{p}+\hat{p}) = n\hat{p}(1-\hat{p}) .$$

Dermed : Den forventningsrette estimatoren for $\sigma^2 = p(1-p)$ blir:

$$\hat{\sigma}^2 = \frac{N - 1}{N} \cdot \frac{n}{n-1}\hat{p}(1-\hat{p}) \approx \hat{p}(1-\hat{p}) .$$

Dvs., den intuitive estimatoren $\hat{p}(1-\hat{p})$ er tilnærmet forventningsrett. Vi ser også at

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

Generelle sannsynlighetsutvalg

Middeltallet μ og variansen σ^2 er eksempler på numeriske beskrivelser av populasjonen som kalles *populasjonsparametre* eller bare *parametre.* La nå generelt θ være en populasjonsparameter som vi ønsker å estimere. Anta vi tar et sannsynlighetsutvalg, men ikke nødvendigvis enkelt tilfeldig. La $\hat{\theta}$

være en estimator for θ basert på utvalget. Nøyaktig som i det spesielle tilfelle ovenfor så er naturlige krav til $\hat{\theta}$:

(I) Forventningsretthet, iallfall tilnærmet : $E(\hat{\theta}) = \theta$

(II) Liten $Var(\hat{\theta})$.

Igjen defineres feilmarginen som $2 \cdot SE(\hat{\theta})$, hvor $SE(\hat{\theta})$ er (estimert) $\sqrt{\hat{V}ar(\hat{\theta})}$.

Det kan vises at generelt for stor *n* (vanligvis er $n \geq 100$ nok, ofte holder det med mindre *n*) så vil

$$\mathrm{P}(-2 \leq \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}ar(\hat{\theta})}} \leq 2) \approx 0{,}95. \qquad (*)$$

Dette følger av Sentralgrenseteoremet som sier at fordelingen til $\hat{\theta}$ følger den normale fordelings-kurven for store *n,* (under visse betingelser på $\hat{\theta}$ og utvalgsplanen som er oppfylt i de tilfellene vi skal se på). (*) forklarer definisjonen av feilmargin, siden

$$\mathrm{P}(|\hat{\theta} - \theta| \leq 2\,SE(\hat{\theta}))$$

$$= \mathrm{P}(-2SE(\hat{\theta}) \leq \hat{\theta} - \theta \leq 2SE(\hat{\theta}))$$

$$= \mathrm{P}(-2 \leq \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}ar(\hat{\theta})}} \leq 2) \approx 0{,}95.$$

Dvs., at sannsynligheten for at estimeringsfeilen, $|\hat{\theta} - \theta|$, er høyst lik feilmarginen er omtrent 0,95. Samtidig gir egenskapen (*) oss anledning til å beregne et *95% konfidensintervall.* Idéen er å lage et intervall ($\hat{\theta}_A$, $\hat{\theta}_B$), basert på data i utvalget, som høyst sannsynlig inneholder den sanne θ.

DEFINISJON. ($\hat{\theta}_A$, $\hat{\theta}_B$) *er et 95% konfidensintervall for* θ *hvis*

$$\mathrm{P}(\hat{\theta}_A \leq \theta \leq \hat{\theta}_B) = 0{,}95\ .$$

Deknings-sannsynligheten 0,95 kalles *konfidensnivået.*

Fra (*): $0{,}95 \approx \mathrm{P}(-2 \leq \frac{\hat{\theta} - \theta}{\sqrt{\hat{V}ar(\hat{\theta})}} \leq 2)$ som kan vises å være lik $\mathrm{P}(\hat{\theta} - 2\sqrt{\hat{V}ar(\hat{\theta})} \leq \theta \leq \hat{\theta} + 2\sqrt{\hat{V}ar(\hat{\theta})})$

Dermed : Et tilnærmet 95% konfidensintervall for θ er gitt ved :

$$(\hat{\theta} - 2\sqrt{\hat{V}ar(\hat{\theta})}\ ,\ \ \hat{\theta} + 2\sqrt{\hat{V}ar(\hat{\theta})}\ ) = \hat{\theta} \pm 2\sqrt{\hat{V}ar(\hat{\theta})}\ .$$

Fortolkning av deknings-sannsynligheten 0,95 : Etter at utvalget er observert og intervallet *beregnet* , er da sannsynligheten 0,95 for at θ ligger i intervallet ? Svaret er *nei* ! For hvert konkret tilfelle er θ enten i intervallet eller ikke. Sannsynligheten er 1 eller 0. Derimot, ved mange gjentatte utvalg så vil omtrent 95% av de beregnede konfidensintervallene inneholde den ukjente θ.

Eksempel 1.6, forts. 95% konfidensintervaller for partiandelene p_A og p_H blir:

$$p_A : \hat{p}_A \pm 2SE(\hat{p}_A) = 0{,}327 \pm 0{,}024 = (0{,}303\ ,\ 0{,}351)$$

Vi kan anslå med 95% sikkerhet at Abeiderpartiets oppslutning er mellom 30,3% og 35,1%

$$p_H : \hat{p}_H \pm 2SE(\hat{p}_H) = 0{,}197 \pm 0{,}021 = (0{,}176 , 0{,}218)$$

Vi kan anslå med 95% sikkerhet at Høyres oppslutning er mellom 17,6% og 21,8% .

Det er også mulig å bruke andre konfidensnivåer, for eksempel 0,90. Vi har, igjen fra normaltilnærmingen, at $\hat{\theta} \pm 1{,}65\sqrt{\hat{V}ar(\hat{\theta})}$ er et 90% konfidensintervall.

Eksempel 1.6, forts. 90% konfidensintervaller for p_A og p_H :

$$p_A : 0{,}327 \pm 0{,}020 = (0{,}307 , 0{,}347)$$
$$p_H : 0{,}197 \pm 0{,}017 = (0{,}180 , 0{,}214) .$$

1.4 Utvalgsplan

Utgangspunktet er en populasjon bestående av N enheter, hvor N er kjent. Alle enhetene i poulasjonen kan identifiseres og er nummerert fra 1 til N . Populasjonen betegnes med $U = \{1,2, \ldots , N\}$. I en undersøkelse om populasjonen vil vi være interessert i mange variable. Vi skal nå konsentrere oss om hvordan vi utfører statistiske analyser separat for hver variabel, basert på et utvalg fra populasjonen. Som tidligere lar vi y betegne en variabel , og y_i verdien av y for enhet i, $i = 1, \ldots , N$. Tallene y_i betraktes som gitte verdier, konstanter, og ikke som verdier av tilfeldige variable.

Vi skal se på problemet med å estimere populasjonstotalen t eller gjennomsnittet $\mu = t/N$, ved å trekke et utvalg s av enheter fra U uten tilbakelegging. Utvalget er en delmengde av U. La n betegne størrelsen på s. Utvalget kan betegnes med

$$s = \{i_1 , i_2 , \ldots , i_n \} ,$$

der indeksene $i_1 < i_2 < \ldots < i_n$ betegner de enhetene som er trukket ut til utvalget. Som illustrasjon, anta vi trekker et enkelt tilfeldig utvalg med $n = 3$ ved å trekke en og en enhet, med resultat: 10 , 3 , 8. Da består s av enhetene 3,8,10, $s = \{3,8,10 \}$, og $i_1 = 3$, $i_2 = 8$, $i_3 = 10$.

Utvalget s sies å være et sannsynlighetsutvalg hvis vi trekker ut enheter med visse sannsynligheter. Dvs. for ethvert mulig utvalg s er det knyttet en kjent sannsynlighet $p(s)$ = P(utvalget s trekkes), sannsynligheten for at det uttrukne utvalget blir akkurat dette utvalget s .

Eksempel 1.7 . Vi trekker et enkelt tilfeldig utvalg på 2 personer fra en populasjon på 4 personer, $U = \{1,2,3,4\}$. Det er 6 utvalg bestående av 2 personer; $\{i,j\}$ for $i = 1,2,3$ og $j > i$. Det betyr at $p(\{i,j\}) = 1/6$ for for $i = 1,2,3$ og $j > i$. For alle de andre s, ialt 9 utvalg , så er $p(s) = 0$.

Samlingen av alle $p(s)$ for alle mulige s kalles *utvalgsplanen.* For enhver utvalgsplan så vil summen av alle $p(s)$ være lik 1. For eksempel, et enkelt tilfeldig utvalg beskrives ved følgende utvalgsplan :

$$p(s) = \begin{cases} \dfrac{1}{\text{antall utvalg med } n \text{ enheter}} & \text{hvis } s \text{ har } n \text{ enheter} \\ 0 & \text{hvis antall enheter i } s \text{ ikke er lik } n \end{cases}$$

Det stokastiske elementet i utvalgsteorien kan beskrives ved indikatorvariablene for inklusjon i utvalget :

$I_k = 1$ hvis enhet k trekkes ut til utvalget
$= 0$ ellers.

Trekksannsynligheten π_k for enhet *k* er sannsynligheten for at enhet *k* trekkes ut . Dvs.,

$$\pi_k = P(I_k = 1) .$$

For enhver utvalgsplan hvor utvalgsstørrelsen *n* er bestemt på forhånd så has at summen av alle trekksannsynlighetene i hele populasjonen er lik *n*, $\pi_1 + \pi_2 + \ldots + \pi_N = n$.

Dette ses av følgende betraktning:

$$n = I_1 + I_2 + \ldots + I_N \quad \Rightarrow n = E(n) = E(I_1) + E(I_2) + \ldots + E(I_N) = \pi_1 + \pi_2 + \ldots + \pi_N .$$

Som en illustrasjon på begrepene vi har innført skal vi se på et enkelt eksempel hvor det ikke er rimelig å trekke et enkelt tilfeldig utvalg.

Eksempel 1.8. Vi har en populasjon på 4 bedrifter. Variabelen av interesse er omsetningen i løpet av et år. Som typisk er tilfelle skal vi anta at det er stor forskjell på bedriftene. Anta, for illustrasjonens skyld, at vi vet at omsetningen for et gitt år for de fire bedriftene 1,2,3,4 er 100, 200, 300 og 1000 (millioner) kroner henholdsvis. Vi trekker et utvalg på 2 bedrifter for å estimere den totale omsetningen på følgende måte. Bedrift 4 (med den største omsetningen) *skal* være med, og den andre bedriften trekkes fra bedriftene 1,2,3 med sannsynligheter :

bedrift 1 : 0,2 bedrift 2 : 0,3 bedrift 3 : 0,5 .

Utvalgsplanen *p(s)* er da gitt ved :

$$p(\{1,4\}) = 0,2$$
$$p(\{2,4\}) = 0,3$$
$$p(\{3,4\}) = 0,5$$
$$p(s) = 0 \text{ ellers .}$$

Det er ialt 12 utvalg *s* som har sannsynlighet 0 for å bli valgt. De individuelle trekksannsynlighetene beregnet fra utvalgsplanen blir:

$$\pi_1 = p(\{1,4\}) = 0,2$$
$$\pi_2 = p(\{2,4\}) = 0,3$$
$$\pi_3 = p(\{3,4\}) = 0,5$$
$$\pi_4 = p(\{1,4\}) + p(\{2,4\}) + p(\{3,4\}) = 1,0 .$$

Vi ser at trekksannsynlighetene summerer seg til utvalgsstørrelsen 2.

Planlegging av utvalgsstørrelse n

Utvalgets størrelse har en avgjørende innflytelse på kostnadene til en undersøkelse. Bestemmelse av *n* er nær knyttet sammen med formålet for undersøkelsen. Noen utvalgsundersøkelser klarer seg med førre enn 1000 enheter i utvalget, mens andre krever så mye som 24000 (Arbeidskraftundersøkelsen). Det er hovedsakelig tre forhold å ta hensyn til:

(a) *Ønsket nøyaktighet på resultatene*. I praksis er det vanskelig å si noe på forhånd, men det er viktig å forsøke å si noe om dette.

(b) *Homogenitet i populasjonen*. Trenger mindre utvalg hvis det er liten variasjon i populasjonen.

(c) *Oppsplittinger av utvalget for estimering i delpopulasjoner.*

Det er ofte (c) som setter de største kravene.

Vi skal nå se litt nærmere på (a). Disse betraktingene kan også brukes på (c) som anslag på størrelsen til de nødvendige delutvalg.

Anta vi skal estimere en populasjonsandel p med enkelt tilfeldig utvalg. Med liten utvalgsandel n/N så er 95% konfidensintervall for p gitt ved, med $\hat{p}$ lik observert andel i utvalget:

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,.$$

La oss si vi har bestemt oss for et nøyaktighetskrav på ± 5% :

$$2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0{,}05$$
$$\Rightarrow n = 400 \cdot 4\hat{p}(1-\hat{p}) \le 400.$$

Estimatet $\hat{p}$ er ukjent i planleggingsfasen. Vi bruker derfor enten det konservative anslaget 400, eller en planleggingsverdi p_0 som gir

$$n = 1600 p_0 (1 - p_0).$$

Med et generelt nøyaktighetskrav d (istedenfor 0,05) :

$$n = \frac{4 p_0 (1 - p_0)}{d^2}.$$

Oppgaver

1.1 Vi skal se på yrkesaktiviteten i kommunen Etnedal, basert på Folke- og boligtellingen 1990, FoB90. Yrkesaktivitet registreres ikke for personer under 16 år. Man fant at av 1558 innbyggere totalt var 1330 16 år eller eldre. Av disse 1330 personene var igjen 797 yrkesaktive.

Med statistikkpakken SAS ble 5 enkle tilfeldige 10% utvalg (dvs. 133 personer) trukket, og antall yrkesaktive i utvalgene ble notert. Vi skal sammenligne de faktiske resultatene med de teoretiske verdiene for forventning og varians for antall yrkesaktive. Hvert av de 5 utvalgene er trukket fra hele populasjonen på 1330 personer. Resultatene ble :

Utvalg nr.	1	2	3	4	5
Antall yrkesaktive	71	81	77	78	83

På grunnlag av disse data finner vi følgende estimater for forventning, varians og standardavvik til antall yrkesaktive.

	Estimat
Forventning	78
Varians	21
Standardavvik	4,58

(a) Finn de teoretiske verdiene for forventning, varians og standardavvik for antall yrkesaktive i et enkelt tilfeldig utvalg på 133 personer fra Etnedal

(b) Sammenlign resultatene i (a) med estimatene i tabellen.

1.2 Vi viser til oppgave 1.1. Fyll inn i tabellen nedenfor de teoretiske verdiene for antall yrkesaktive basert på hypergeometisk og binomisk fordeling, og sammenlign.

	Hypergeometrisk fordeling	Binomisk fordeling
Forventning		
Varians		
Standardavvik		

1.3* Betrakt en populasjon på N enheter med y - verdier (y_1 , y_2 , . . .,y_N) . La X være verdien til y ved tilfeldig trekking av en enhet. La $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$ og $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$.

(a) Vis at $E(X) = \mu$ og $Var(X) = \sigma^2$.

(b) La $y_i = 1$ hvis enhet i har kjennetegn A, og 0 ellers. La p være populasjonsandelen av A , $p = \mu$.

Vis at $\sigma^2 = p(1-p)$.

1.4 Et enkelt tilfeldig utvalg på 400 personer skal utspørres om de noen gang har deltatt i en intervju-undersøkelse (utenom denne undersøkelsen). Av svarene går det fram at 58 hadde en eller fler ganger deltatt i en intervjuundersøkelse. Lag et 95% konfidensintervall for andelen i befolkningen som har deltatt i intervjuundersøkelser.

2. Basis estimatorer for utvalgsundersøkelser

2.1 Horvitz-Thompson estimatoren

Vi skal nå betrakte estimering av totalen t for en generell utvalgsplan. Den eneste forutsetningen er at alle enhetene i populasjonen har en sjanse for å bli trukket ut, dvs. at $\pi_k > 0$ for alle k. En estimator som alltid er forventningsrett er Horvitz-Thompson estimatoren som er lik summen av alle y_i/π_i - verdiene i utvalget:

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

F.eks., hvis $s = \{3,8,10\}$, $\hat{t}_{HT} = \frac{y_3}{\pi_3} + \frac{y_8}{\pi_8} + \frac{y_{10}}{\pi_{10}}$.

Denne estimatoren er mye brukt i statistikkproduksjonen i Statistisk sentralbyrå. Det kan vises at variansen til estimatoren er lik summen, for alle par (i,j) i populasjonen, av uttrykkene $(\pi_i \pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$:

$$Var(\hat{t}_{HT}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\pi_i \pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

hvis størrelsen n på s er bestemt på forhånd.

Her er π_{ij} = P(enhetene i og j er trukket ut til utvalget) = P($I_i = 1$ og $I_j = 1$) . F.eks., hvis $N = 3$:

$$Var(\hat{t}_{HT}) = (\pi_1\pi_2 - \pi_{12})\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2 + (\pi_1\pi_3 - \pi_{13})\left(\frac{y_1}{\pi_1} - \frac{y_3}{\pi_3}\right)^2 + (\pi_2\pi_3 - \pi_{23})\left(\frac{y_2}{\pi_2} - \frac{y_3}{\pi_3}\right)^2$$

Var$(\hat{t}_{HT})$ må estimeres for å finne standardfeilen , $SE(\hat{t}_{HT}) = \sqrt{\hat{V}ar(\hat{t}_{HT})}$. Anta utvalgsplanen har fast størrelse på s. Det er vist i Appendiks B at en forventningsrett estimator (såfremt alle $\pi_{ij} > 0$) er gitt ved :

$$\hat{V}ar(\hat{t}_{HT}) = \sum_{i \in s} \sum_{\substack{j \in s \\ j>i}} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \ .$$

F.eks., hvis $s = \{3,8,10\}$,

$$\hat{V}ar(\hat{t}_{HT}) = \frac{\pi_3\pi_8 - \pi_{38}}{\pi_{38}}\left(\frac{y_3}{\pi_3} - \frac{y_8}{\pi_8}\right)^2 + \frac{\pi_3\pi_{10} - \pi_{3,10}}{\pi_{3,10}}\left(\frac{y_3}{\pi_3} - \frac{y_{10}}{\pi_{10}}\right)^2 + \frac{\pi_8\pi_{10} - \pi_{8,10}}{\pi_{8,10}}\left(\frac{y_8}{\pi_8} - \frac{y_{10}}{\pi_{10}}\right)^2 \ .$$

Vi kan nå beregne et 95% konfidensintervall for t, for store n og $N >> n$. Det er gitt ved

$$\hat{t}_{HT} \ \pm 2\sqrt{\hat{V}ar(\hat{t}_{HT})} \ .$$

Variansestimatet kan alltid beregnes siden $\pi_{ij} > 0$ for alle *i,j* i utvalget *s* nødvendigvis. Men hvis ikke *alle* $\pi_{ij} > 0$ bør ikke variansestimatet brukes. Det kan da gi meget feilaktige anslag. Eksempel 1.8 vil illustrere dette. Hvis variansestimering anses å være viktig , bør man velge utvalgsplaner hvor alle par (*i,j*) har en positiv sannsynlighet for å bli trukket ut. Ellers må man prøve å finne estimatorer som *overestimerer* $Var(\hat{t}_{HT})$. En variansestimator $\hat{V}$ overestimerer $Var(\hat{t}_{HT})$ hvis $E(\hat{V}) > Var(\hat{t}_{HT})$. Da vil konfidensintervallet basert på $\hat{V}$, $\hat{t}_{HT} \pm 2\sqrt{\hat{V}}$, ha et konfidensnivå på minst 95%. Vi sier intervallet er konservativt. Kan man også utlede en estimator som *underestimerer* $Var(\hat{t}_{HT})$ er det mulig å finne ut hvor mye overestimering vi har. En type utvalgsundersøkelser hvor dette må gjøres er i såkalt systematisk sampling eller intervalltrekking . Som et eksempel på intervalltrekking, anta vi skal trekke et 5% utvalg fra populasjonen. Fra en liste over populasjonens enheter, trekkeregisteret, velges først tilfeldig en enhet fra de første 20 på liste, og deretter hver 20. enhet på listen.

En variansestimator bør alltid gi positive verdier. En svakhet ved $\hat{Var}(\hat{t}_{HT})$ er at det finnes utvalgsplaner og utvalgsverdier ($y_i : i \in s$) slik at estimatet blir negativt. Men for mange av de vanligste utvalgsplanene vil $\hat{Var}(\hat{t}_{HT})$ være positiv.

Eksempel 1.8, forts. Populasjonen består av 4 bedrifter med *y*-verdier 100,200,300 og 1000, og trekksannsynligheter $\pi_1 = 0{,}2$, $\pi_2 = 0{,}3$, $\pi_3 = 0{,}5$, $\pi_4 = 1{,}0$. Sann totalverdi er $t = 1600$. Anta vi trekker utvalget {2,4}. Horvitz-Thompson estimatet blir da :

$$\hat{t}_{HT} = \frac{y_2}{\pi_2} + \frac{y_4}{\pi_4} = \frac{200}{0{,}3} + 1000 = 1666{,}67.$$

I dette tilfellet er det enkelt å utlede $E(\hat{t}_{HT})$ og $Var(\hat{t}_{HT})$ direkte. Det er kun 3 utvalg som er mulige :

{1,4} med sannsynlighet 0,2 , {2,4} med sannsynlighet 0,3 og {3,4} med sannsynlighet 0,5. $\hat{t}_{HT}$ har følgende mulige verdier :

$$s = \{1,4\}: \hat{t}_{HT} = (y_1 / 0{,}2) + y_4 = 1500$$
$$s = \{2,4\} : \hat{t}_{HT} = 5000/3 = 1666{,}67$$
$$s = \{3,4\} : \hat{t}_{HT} = (y_3 / 0{,}5) + y_4 = 1600.$$

Vi ser : $E(\hat{t}_{HT}) = 1500 \cdot (0{,}2) + 1666{,}67 \cdot (0{,}3) + 1600 \cdot (0{,}50)$

$$= 300 + 500 + 800 = 1600 = t .$$

$$Var(\hat{t}_{HT}) = (1500 - 1600)^2(0{,}2) + (5000/3 - 1600)^2(0{,}3) + (1600 - 1600)^2(0{,}5)$$
$$= 2000 + 4000/3 = 3333{,}33 .$$
$$SE(\hat{t}_{HT}) = \sqrt{Var(\hat{t}_{HT})} = 57{,}74 .$$

Vi kan også utlede variansen ved å bruke formelen som ble gitt:

$$Var(\hat{t}_{HT}) = \sum_{i=1}^{3} \sum_{j=i+1}^{4} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 .$$

Vi har at $\pi_i\pi_4 = \pi_{i4}$ for i = 1,2,3. Disse leddene faller derfor bort. Også, alle andre $\pi_{ij} = 0$. Dette gir:

$$Var(\hat{t}_{HT}) = \pi_1\pi_2\left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2 + \pi_1\pi_3\left(\frac{y_1}{\pi_1} - \frac{y_3}{\pi_3}\right)^2 + \pi_2\pi_3\left(\frac{y_2}{\pi_2} - \frac{y_3}{\pi_3}\right)^2$$

$$= (0{,}06)(500 - \tfrac{2000}{3})^2 + (0{,}10)(500 - 600)^2 + (0{,}15)(600 - \tfrac{2000}{3})^2$$

$$= 1666{,}667 + 1000 + 666{,}667 = 3333{,}33 \text{ , som ovenfor !}$$

Variansestimatet med $s = \{i,4\}$ er lik

$$\hat{Var}(\hat{t}_{HT}) = \sum_{i\in s}\sum_{\substack{j\in s \\ j>i}} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

$$= \frac{\pi_i\pi_4 - \pi_{i4}}{\pi_{i4}}\left(\frac{y_i}{\pi_i} - \frac{y_4}{\pi_4}\right)^2 = 0 \ !$$

Dette følger fordi $\pi_i\pi_4 = \pi_{i4}$ for i = 1,2,3. Så variansestimatet er uten mening. Dette illustrerer det vi nevnte tidligere at estimatet kan brukes bare når alle $\pi_{ij} > 0$.

La oss se hvordan det ville være å bruke $\bar{y}_s$ til å estimere *t/N*, dvs. $N\bar{y}_s$ som estimator for *t.* Denne kalles ekspansjonsestimatoren, $\hat{t}_e = N\bar{y}_s$. Med den utvalgsplanen vi har valgt fungerer denne estimatoren meget dårlig: Mulige verdier av estimatoren er 2200, 2400 og 2600 for s = {1,4},{2,4} og {3,4}. Dette gir at

$$E(\hat{t}_e) = 2200(0{,}2) + 2400(0{,}3) + 2600(0{,}5) = 2460 = 1{,}5375t.$$

Estimatoren er meget skjev ! Ekspansjonsestimatoren fungerer ofte dårlig når trekksannsynlighetene er veldig forskjellige.

Et mer nærliggende spørsmål vil være hvordan $\hat{t}_e$ vil fungere i forhold til $\hat{t}_{HT}$ hvis vi trekker et enkelt tilfeldig utvalg på 2 bedrifter. $\hat{t}_e$ er Horvitz - Thompson estimatoren i enkelt tilfeldig utvalg. Dette følger fra det faktum at da er $\pi_k = n/N$ for alle enhetene i populasjonen. Det ses lett ved å bruke egenskapen at totalsummen av trekksannsynlighetene er lik *n.* I enkelt tilfeldig utvalg har vi at alle π_k er like, og siden summen er lik *n,* så må $\pi_k = n/N$. (Også vist i Appendiks A, kap. 2.1 ved hypergeometrisk fordeling). Dermed,

$$\hat{t}_{HT} = \sum_{i\in s}\frac{y_i}{\frac{n}{N}} = \sum_{i\in s}\frac{N}{n}y_i = \frac{N}{n}\sum_{i\in s}y_i = N\bar{y}_s \ .$$

Det vi egentlig sammenligner er de to utvalgsplanene. I enkelt tilfeldig trekking er det 6 like sannsynlige utvalg, og $\hat{t}_e$ har følgende mulige verdier: 600,800,1000,2200,2400,2600. Dette gir at $E(\hat{t}_e) = 1600 = t$, og $Var(\hat{t}_e) = 666.666{,}67$, og $SE(\hat{t}_e) = \sqrt{Var(\hat{t}_e)} = 816{,}50$ (oppgave 2.3). I forhold til $SE(\hat{t}_{HT}) = \sqrt{Var(\hat{t}_{HT})} = 57{,}74$ ser vi at $SE(\hat{t}_e)$ er omtrent 14 ganger større. Det er enormt mye å vinne ved å sikre seg at den største bedriften alltid er med i utvalget.

Eksempel 1.8 viste hvor mye man kan vinne på å ha forskjellige trekksannsynligheter og samtidig bruke Horvitz-Thompson estimatoren. Ser vi på formelen for variansen til estimatoren,

$$Var(\hat{t}_{HT}) = \sum_{i=1}^{N}\sum_{j=i+1}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 ,$$

finner vi at variansen blir liten hvis vi bestemmer trekksannsynlighetene slik at y_i/π_i blir omtrent like. Det forutsetter at vi har en viss kjennskap til y-verdienes størrelse. F.eks., enheter med små y- verdier bør ha liten trekksannsynlighet og enheter med store y- verdier bør ha høy trekksannsynlighet. Selv om y- verdiene er ukjente under planleggingen av en undersøkelse kan det være at vi har andre mål på størrelse fra trekkeregisteret for populasjonen som sier noe om størrelseorden på y- verdiene. I eksempel 1.8 kunne det være at vi kjente antall ansatte i hver enkelt bedrift på forhånd.

Det er mest for populasjoner med høyst varierende y-verdier at det kan lønne seg å velge en utvalgsplan med forskjellige trekksannsynligheter. I personundersøkelser er vanligvis dette ikke tilfelle og utvalgsplanene Statistisk sentralbyrå bruker i slike undersøkelser har vanligvis $\pi_k = n/N$ for alle personer i populasjonen (men de er ikke enkle tilfeldige utvalg). Vi skal senere beskrive den generelle utvalgsplanen for besøksundersøkelser i Statistisk sentralbyrå.

Hvis det er liten sammenheng mellom trekksannsynlighetene og y-verdiene kan vi risikere at $Var(\hat{t}_{HT})$ blir meget stor. Da bør man *ikke* bruke Horvitz-Thompson estimatoren, selv om vi har forskjellige trekksannsynligheter. Fra variansformelen ser vi at variansen blir stor hvis enheter med små y_i har store π_i og omvendt. Vi skal nå illustrere hvor ille det kan bli med et enkelt eksempel. Dette er en forenklet versjon av Basu's elefanteksempel.

Eksempel 2.1. Betrakt en populasjon på 3 elefanter. Elefantene skal transporteres med båt , og man trenger et estimat for totalvekten. Veiing av en elefant er ingen enkel affære, så det besluttes å veie kun en elefant, og basere anslaget for totalvekten på vekten til denne ene utvalgte elefanten. Fra tidligere vet elefanteieren at elefant *2* har en vekt y_2 som er nær gjennomsnittsvekten til de 3 elefantene. Det er derfor ønskelig å velge denne elefanten og bruke $3y_2$ som estimat. For å få en forventningsrett estimator må imidlertid alle trekksannsynlighetene være positive. Det besluttes å la $\pi_2 = 0{,}90$ og $\pi_1 = \pi_3 = 0{,}05$. Anta at vektene til de tre elefantene *1,2,3* er (i tonn) 1, 2, 4. Horvitz-Thompson estimatoren gir følgende verdier :

$$\hat{t}_{HT} = y_i/\pi_i \quad \text{hvis } s = \{i\}$$

$$= \begin{cases} 20 & \text{hvis } s = \{1\} \\ 2{,}22 & \text{hvis } s = \{2\} \\ 80 & \text{hvis } s = \{3\} \end{cases}$$

Alle mulige estimater er langt unna den sanne $t = 7$, og fullstendig «håpløse». Horvitz-Thompson estimatoren kan ikke brukes som anslag for totalvekten, selv om $E(\hat{t}_{HT}) = 7 = t$.

$$(E(\hat{t}_{HT}) = 0{,}05\cdot 20 + 0{,}9\cdot 2{,}22 + 0{,}05\cdot 80 = 7.)$$

Det er $Var(\hat{t}_{HT})$ som er problemet. Vi finner at $Var(\hat{t}_{HT}) = (20-7)^2(0{,}05) + (2{,}22-7)^2(0{,}90) + (80-7)^2(0{,}05) = 295{,}46$ og $SE(\hat{t}_{HT}) = \sqrt{Var(\hat{t}_{HT})} = 17{,}2$!

La oss nå se hvordan ekspansjonsestimatoren fungerer her. Den er her gitt ved:

$$\hat{t}_e = N\bar{y}_s = 3y_i \quad \text{hvis } s = \{i\}.$$

De mulige verdiene av $\hat{t}_e$ er lik 3,6,12. Herav ses at :

$$E(\hat{t}_e) = 3 \cdot 0{,}05 + 6 \cdot 0{,}90 + 12 \cdot 0{,}05 = 6{,}15$$

$$Var(\hat{t}_e) = (3\text{-}6{,}15)^2\,(0{,}05) + (6\text{-}6{,}15)^2(0{,}90) + (12\text{-}6{,}15)^2(0{,}05) = 2{,}2275$$

$$SE(\hat{t}_e) = \sqrt{Var(\hat{t}_e)} = 1{,}49\ .$$

Selvom $\hat{t}_e$ ikke er forventningsrett så er den langt å foretrekke framfor $\hat{t}_{HT}$. Det er høyst sannsynlig at elefant *2* blir valgt og estimatene blir da $\hat{t}_e = 6$ og $\hat{t}_{HT} = 1{,}11 \cdot y_1 = 2{,}22$. Det er klart at $\hat{t}_{HT}$ er fullstendig urimelig.

2.2 Rate-estimatoren

La oss nå anta at vi har kjent tilleggsinformasjon for hele populasjonen , $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$. Alle x_i er positive. F.eks., x_i kan være et mål på størrelsen til enheten. I eksempel 1.8 kan x_i være antall ansatte. La $X_\circ$ være totalsummen av alle x_i'ene i populasjonen, dvs. $X_\circ = x_1 + x_2 + \ldots + x_N$. *Rate- estimatoren* for populasjonstotalen *t* av *y*-variabelen er lik forholdet mellom utvalgssummene av *y*- og *x*-variabelen multiplisert med $X_\circ$, og kan uttrykkes slik:

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

hvor $\bar{x}_s = \sum_{i \in s} x_i / n$ er gjennomsnittet av *x*-verdiene i utvalget. Vi ser at

$$\hat{t}_R = \frac{X_\circ}{N\bar{x}_s}\hat{t}_e$$

hvor $\hat{t}_e = N\bar{y}_s$.

Det vanlige er å bruke rate-estimatoren i forbindelse med enkelt tilfeldig utvalg hvor Horvitz-Thompson estimatoren er $\hat{t}_e$, men den kan også benyttes ved andre utvalgsplaner. Vi ser at $\hat{t}_R$ justerer $\hat{t}_e$ i de tilfeller hvor *x*-verdiene i utvalget er for små eller for store. Dette vil være rimelig å gjøre hvis det er en positiv sammenheng mellom x_i og y_i'ene. I neste kapittel skal vi sammenligne $\hat{t}_R$ og $\hat{t}_e$ nærmere når utvalget trekkes enkelt tilfeldig.

Eksempel 2.2. Anta at det totale dyrkede areal potetåker *t* skal estimeres for en populasjon på *N* gårder. Det totale dyrkede areal $X_\circ$ antas kjent. Med et utvalg på *n* gårder er da y_i = dyrket areal potetåker for gård *i* og x_i = total dyrket areal for *i*'te gård i utvalget.

Eksempel 1.8, forts. Populasjonen består av 4 bedrifter med *y*-verdier 100,200,300 og 1000. Tilleggsinformasjonen er x_i = antall ansatte i bedrift *i.* Anta $\boldsymbol{x}$ = (20,30,50,200), slik at $X_\circ$ = 300. Det er kun 3 utvalg som er mulige : {1,4} med sannsynlighet 0,2 , {2,4} med sannsynlighet 0,3 og {3,4} med sannsynlighet 0,5. $\hat{t}_R$ har følgende mulige verdier :

$$s = \{1,4\}: \ \hat{t}_R = 300 \cdot \frac{1100}{220} = 1500 \ \ (\hat{t}_{HT} = 1500 \ , \ \hat{t}_e = 2200)$$

$$s = \{2,4\} : \hat{t}_R = 300 \cdot \frac{1200}{230} = 1565{,}22 \ \ (\hat{t}_{HT} = 1666{,}67 \ , \ \hat{t}_e = 2400)$$

$$s = \{3,4\} : \hat{t}_R = 300 \cdot \frac{1300}{250} = 1560 \ \ (\hat{t}_{HT} = 1600 \ , \ \hat{t}_e = 2600)$$

Sann totalverdi er t = 1600 så vi ser at $\hat{t}_R$ er litt skjev, $E(\hat{t}_R)$ = 1549,6. Variansen er imidlertid meget liten , $Var(\hat{t}_R)$ = 619,3 og $SE(\hat{t}_R)$ = 24,9, sammenlignet med $SE(\hat{t}_{HT})$ = 57,74.

Hvis vi tar et enkelt tilfeldig utvalg på 2 bedrifter så blir $E(\hat{t}_R)$ = 1669,1 og $SE(\hat{t}_R)$ = 137,2 (oppgave 2.3). Fra tidligere har vi at da er $SE(\hat{t}_e)$ = 816,5. Rate-estimatoren er en kraftig forbedring av ekspansjonestimatoren. Fremdeles er imidlertid enkelt tilfeldig utvalg en mye dårligere utvalgsplan enn den opprinnelige utvalgsplanen.

Oppgaver

2.1* Horvitz-Thompson estimatoren kan uttrykkes på følgende form:

$$\hat{t}_{HT} = \sum_{k=1}^{N} I_k \frac{y_k}{\pi_k} \ .$$

Bruk dette til å vise at denne estimatoren er forventningsrett.

2.2 Betrakt eksempel 1.8 . Anta utvalgsplanen har samme opplegg som før med den endring at bedriftene 1,2,3 trekkes som den andre bedriften i utvalget med sannsynligheter

bedrift 1: 0,5 bedrift 2 : 0,3 bedrift 3 : 0,2 .

(a) Skriv ned utvalgsplanen.

(b) Finn $E(\hat{t}_{HT})$ og $Var(\hat{t}_{HT})$.

(c) Sammenlign med den opprinnelige utvalgsplanen. Hvilken vil du velge ?

2.3 Betrakt eksempel 1.8, men nå med et enkelt tilfeldig utvalg på 2 bedrifter.

(a) Vis at $E(\hat{t}_e) = 1600$, $Var(\hat{t}_e) = 666.666{,}67$, $SE(\hat{t}_e)$ = 816,5 .

(b) Vis at $E(\hat{t}_R) = 1669{,}1$, $Var(\hat{t}_R) = 18810{,}07$ $SE(\hat{t}_R)$ = 137,15.

2.4 Betrakt eksempel 2.1. Anta nå at en elefant trekkes tilfeldig.

(a) Finn $E(\hat{t}_e), Var(\hat{t}_e)$ og $SE(\hat{t}_e)$, hvor nå $\hat{t}_e = 3y_i$, hvis $s = \{i\}$.

(b) Hvilken utvalgsplan er best av denne og den i eksempel 2.1 ?

2.5 I mange samfunnsundersøkelser består populasjonen av husholdninger. Anta vi har en populasjon med 3 husholdninger. Husholdning 1 består av 3 personer med en samlet inntekt på kr. 400 000, mens de to andre er en-person husholdninger med inntekt på 150 000 for husholdning 2 og 200 000 for husholdning 3. Vi ønsker å estimere den samlede inntekten (som vi vet er kr. 750 000), på grunnlag av et utvalg på 2 husholdninger. Utvalget trekkes ved å trekke 2 personer rent tilfeldig, og består av de tilhørende husholdningene. Ved andre trekning fjernes husholdningen som tilhørte personen som ble trukket første gang. Det vil si at vi trekker husholdninger uten tilbakelegging. F.eks. hvis personen som ble trukket første gangen tilhørte husholdning 1, så var det på annen trekning kun mulig å trekke en av de to personene i husholdning 2 eller 3.

(a) La π_1, π_2, π_3 være trekksannsynlighetene for husholdningene 1,2,3. Vis at

$$\pi_1 = \tfrac{9}{10}, \pi_2 = \pi_3 = \tfrac{11}{20}.$$

(Hint: P(trekke husholdning 1 første gang) = $\frac{3}{5}$, og P(trekke husholdning 1 andre gang) = $\frac{2}{5} \cdot \frac{3}{4}$.)

(b) Anta utvalget består av husholdningene 1 og 3. Beregn Horvitz-Thompson estimatoren, rate-estimatoren og ekspansjonsestimatoren, og sammenlign. For rate-estimatoren, la x_i = antall personer i husholdning *i.*

(c) De andre mulige utvalgene er {1,2} og {2,3}. Beregn for disse to utvalgene verdiene av de tre estimatorene i (b).

(d) Vis at utvalgsplanen er gitt ved :

$p(\{1,2\}) = 9/20$

$p(\{1,3\}) = 9/20$

$p(\{2,3\}) = 2/20$

(e) Finn forventning og varians til de tre estimatorene. Hvilken estimator vil du foretrekke?

3. Enkelt tilfeldig utvalg

Vi har ved flere anledninger betraktet utvalgsplanen som gir enkelt tilfeldig utvalg . Vi skal i dette kapittel oppsummere tidligere utledede egenskaper til estimatorer for totalen t . Samtidig skal vi anvende teorien i kapittel 2 på denne utvalgsplanen.

Utvalgsplanen er gitt ved at alle utvalg s med størrelse n har samme sannsynlighet for å bli trukket. Trekksannsynlighetene $\pi_k = n/N$ for alle enhetene i populasjonen. Som vist i forrige kapittel er Horvitz-Thompson estimatoren derfor lik ekspansjonsestimatoren :

$$\hat{t}_{HT} = \hat{t}_e = N\bar{y}_s \ ,$$

hvor $\bar{y}_s = \sum_{i \in s} y_i / n$ er gjennomsnittet i utvalget s. Vi kan uttrykke estimatoren på følgende måte :

$$\hat{t}_e = \frac{N}{n}\sum_{i \in s} y_i = \sum_{i \in s} \frac{N}{n} y_i \ .$$

Dvs. vi kan regne ut estimatet ved å «blåse opp» hver observert y_i med *vekten N/n.* Vi kaller dette *enkel oppblåsing* siden hver enhet i utvalget har samme vekt. En annen måte å se dette på er at vi tenker oss at hver enhet i utvalget representerer N/n enheter i populasjonen. For eksempel, anta vi tar et 5% utvalg, dvs. $n/N = 0{,}05$. Da er $N/n = 20$, og hver enhet i utvalget representerer 20 enheter i populasjonen.

Fra teorien for Horvitz-Thompson estimatoren vet vi at $\hat{t}_e$ er forventningsrett, $E(\hat{t}_e) = t$. Vi kan også finne $Var(\hat{t}_e)$ ved å bruke den generelle formelen for $Var(\hat{t}_{HT})$. Vi trenger å finne π_{ij} = P(Enhetene i og j trekkes ut) = P($I_i\, I_j = 1$). Alle π_{ij} er like.

I appendiks B blir det vist at hver enkelt π_{ij} er lik

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)}.$$

Populasjonens varians ble tidligere definert til å være :

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 \ \ , \text{ med } \mu = t/N.$$

Det kan vises (se appendiks B) at

$$\text{Var}(\hat{t}_e) = N^2 \frac{\sigma^2}{n}\frac{N-n}{(N-1)},$$

og den forventningsrette estimatoren

$$\hat{Var}(\hat{t}_e) = N^2 \frac{\hat{\sigma}^2}{n} \cdot \frac{N-n}{N-1}$$

hvor

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

og vi ser at $\hat{\sigma}^2$ er forventningsrett for σ^2 .

Et 95% konfidensintervall for *t* er tilnærmet lik, for store *n* og N >> *n* :

$$\hat{t}_e \pm 2N\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \approx \hat{t}_e \pm 2N\frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1-f}$$

hvor $f = n/N$ er *utvalgsandelen* .

La oss nå samle sammen resultatene fra kapittel 1.3 for estimering av en populasjonsandel *p* av enheter med et visst kjennetegn A. Dette er nå et spesialtilfelle av resultatene for en generell *y*-variabel, med y_i = 1/0 hvis enhet *i* har/har ikke kjennetegn A. La X = antall med kjennetegn A i utvalget. En estimator for *p* er $\hat{p} = X/n$. Vi ser at $\hat{p} = \bar{y}_s$ og $\hat{t}_e = N\hat{p}$. I tillegg, $\sigma^2 = p(1\text{-}p)$ og

$$Var(N\hat{p}) = N^2\frac{p(1-p)}{n}\cdot\frac{N-n}{N-1}.$$

$\hat{\sigma}^2 = \frac{N\ -\ 1}{N}\cdot\frac{n}{n-1}\hat{p}(1-\hat{p}) \approx \hat{p}(1-\hat{p})$ som gir at forventningsrett estimator for $Var(N\hat{p})$ er gitt ved:

$$\hat{Var}(N\hat{p}) = N^2\frac{\hat{p}(1-\hat{p})}{n-1}\cdot\frac{N-n}{N}$$

$$= N^2\frac{\hat{p}(1-\hat{p})}{n-1}(1-f) \approx N^2\frac{\hat{p}(1-\hat{p})}{n}(1-f)$$

Et 95% konfidensintervall for antallet med kjennetegn A i populasjonen blir nå

$$N\hat{p} \pm 2N\sqrt{\frac{\hat{p}(1-\hat{p})}{n}(1-f)} \quad .$$

Sammenligning : Rate-estimatoren og ekspansjonsestimatoren

Vi har kjent tilleggsinformasjon for hele populasjonen , $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$. Alle x_i er positive. Rate-estimatoren ble definert i kapittel 2.2, med $X_\circ$ lik totalsummen av alle *x*-verdiene :

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

Vi skal nå sammenligne rate-estimatoren og ekspansjonsestimatoren. La $f = n/N$ og la *R* være forholdet mellom totalsummene i populasjonen for *y* og *x*, $R = t / X_\circ$. Følgende resultater holder for rate-estimatoren :

$$E(\hat{t}_R) \approx t \quad \text{for store } n$$

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

hvor $\sum_{i=1}^{N}(y_i - Rx_i)^2 = (y_1 - Rx_1)^2 + (y_2 - Rx_2)^2 + \ldots + (y_N - Rx_N)^2$.

Fra tidligere ,

$$\text{Var}(\hat{t}_e) = N^2 \frac{\sigma^2}{n} \frac{N-n}{(N-1)} \approx N^2 \frac{1-f}{n} \sigma^2 = N^2 \frac{1-f}{n} \cdot \frac{1}{N} \sum_{i=1}^{N} (y_i - \mu)^2 \ .$$

Dermed har vi at

$$Var(\hat{t}_R) < Var(\hat{t}_e) \Leftrightarrow \sum_{i=1}^{N} (y_i - Rx_i)^2 < \sum_{i=1}^{N} (y_i - \mu)^2 \ .$$

I Appendiks C gis en mer inngående vurdering av forholdet mellom variansene. La $\overline{X} = X_\circ / N$.

$Var(\hat{t}_R)$ kan estimeres ved :

$$\hat{V}ar(\hat{t}_R) = \left(\frac{\overline{X}}{\overline{x}_s} \right)^2 \cdot N^2 \cdot \frac{1-f}{n} \cdot \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{R}x_i)^2 \quad \text{hvor } \hat{R} = \frac{\overline{y}_s}{\overline{x}_s}.$$

Her betyr $\sum_{i \in s} (y_i - \hat{R}x_i)^2$ summen av alle kvadrater av $(y - \hat{R}x)$ -verdiene i utvalget. Variansestimatoren til $Var(\hat{t}_e)$ er , med $\hat{\sigma}^2 = \frac{N-1}{N} \cdot \frac{1}{n-1} \sum_{i \in s} (y_i - \overline{y}_s)^2$, lik :

$$\hat{V}ar(\hat{t}_e) = N^2 \frac{\hat{\sigma}^2}{n} \cdot \frac{N-n}{N-1} = N^2 \frac{1-f}{n} \cdot \frac{1}{n-1} \sum_{i \in s} (y_i - \overline{y}_s)^2 \ .$$

I en gitt situasjon kan vi derfor avgjøre om vi skal benytte $\hat{t}_R$ eller $\hat{t}_e$ ved å sammenligne

$$\left(\frac{\overline{X}}{\overline{x}_s} \right)^2 \sum_{i \in s} (y_i - \hat{R}x_i)^2 \quad \text{og} \quad \sum_{i \in s} (y_i - \overline{y}_s)^2 \ .$$

Et 95% konfidensintervall for t basert på rate-estimatoren er gitt ved:

$$\hat{t}_R \pm 2N \frac{\overline{X}}{\overline{x}_s} \sqrt{\frac{1-f}{n} \cdot \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{R}x_i)^2} \ .$$

Oppgaver

3.1 Betrakt eksempel 1.8. Anta at et enkelt tilfeldig utvalg resulterte i bedriftene 1,3 med (y,x)-verdiene (100,20) og (300,50).

(a) Estimer $SE(\hat{t}_R)$ og $SE(\hat{t}_e)$. Hvilken av de to estimatorene vil du benytte?

(b) Beregn estimatene basert på $\hat{t}_R$ og $\hat{t}_e$.

(c) Beregn 95% konfidensintervaller for totalen t, basert på $\hat{t}_R$ og $\hat{t}_e$. Sammenlign med den kjente t.

4. Stratifisering

4.1. Stratifisert utvalgsplan

I enkelt tilfeldig utvalg har vi betraktet hele populasjonen som én samling av enheter . Når vi skal trekke et landsomfattende utvalg for Norge så er dette en lite hensiktsmessig måte å trekke utvalget på. Det er ikke bare landstall som skal publiseres, men også tall for fylker eller regioner, eller for menn og kvinner hver for seg, eller for forskjellige aldersgrupper, eller for forskjellige yrkesgrupper , eller for forskjellige næringer i bedriftsundersøkelser. Dette setter krav til størrelsene på utvalget i disse forskjellige gruppene. For å kunne utarbeide statistikk for disse gruppene må vi sikre oss et utvalg der de enkelte gruppene sikres god nok representasjon i utvalget. Dette oppnås ved å *stratifisere* utvalget, dvs. inndele populasjonen i delpopulasjoner, såkalte *strata*, og deretter trekke utvalg innen hvert stratum. Vi kaller denne utvalgsplanen for *stratifisert enkelt tilfeldig utvalg* hvis utvalgene fra hvert stratum er enkle tilfeldige utvalg. Stratifisering vil også medføre mer presise estimater enn den vanlige ekspansjonsestimatoren ved at den «retter opp» lite representative utvalg, og er en alternativ utnyttelse av tilleggsinformasjon i forhold til rate-estimatoren.

Strataene betegnes med U_1 , U_2 , . . . , U_H , og delutvalgene betegnes med s_1 , s_2 , . . . , s_H . Strataoppdelingen gjøres ved hjelp av såkalte *stratifiseringsvariable* fra registerinformasjon om populasjon. Det kan være geografisk område, alder , kjønn etc.. La N_h være størrelsen på stratum U_h , og n_h størrelsen på s_h , n_h 'ene er bestemt på forhånd . Det betyr at

$$N = \sum_{h=1}^{H} N_h \quad \text{og} \quad n = \sum_{h=1}^{H} n_h \quad .$$

La t_h være *y*- totalen i stratum U_h. Vi kan nå uttrykke totalen *t* som summen av stratum-totalene t_h ,

$$t = \sum_{h=1}^{H} t_h \, .$$

Ved planlegging av et stratifisert utvalg er det hovedsakelig tre ting som må bestemmes: For det første *hvilke* stratifiseringsvariable en skal bruke, for det andre *hvordan* en skal konstruere strataene, og for det tredje hvordan den totale utvalgsstørrelsen skal *fordeles* på strataene.

Stratifisert enkelt tilfeldig utvalg

Vi skal betrakte stratifisert enkelt tilfeldig utvalg. Da er ekspansjonsestimatoren for t_h , basert på s_h , gitt ved :

$$\hat{t}_h = N_h \bar{y}_{s_h} \qquad \text{hvor} \quad \bar{y}_{s_h} = \frac{1}{n_h} \sum_{i \in s_h} y_i$$

er gjennomsnittet i delutvalg s_h . Stratifiseringsestimatoren for totalen blir da summen av $\hat{t}_h$:

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

Vi ser at $\hat{t}_h$ kan beregnes ved å multiplisere hver *y*-observasjon i s_h med N_h/n_h. Det betyr at hver enhet i delutvalget s_h får vekten N_h/n_h . La σ_h^2 være variansen i stratum U_h ;

$$\sigma_h^2 = \frac{1}{N_h}\sum_{i\in U_h}(y_i - \overline{Y}_h)^2 \ , \qquad \overline{Y}_h = t_h / N_h \ .$$

Stratumestimatorene $\hat{t}_h$ er uavhengige med

$$E(\hat{t}_h) = t_h$$

$$Var(\hat{t}_h) = N_h^2 \cdot \frac{\sigma_h^2}{n_h} \cdot \frac{N_h - n_h}{N_h - 1} \ .$$

Dette gir at :

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

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

$Var(\hat{t}_{st})$ kan estimeres ved å estimere stratumvariansene σ_h^2 med :

$$\hat{\sigma}_h^2 = \frac{N_h - 1}{N_h} \cdot \frac{1}{n_h - 1}\sum_{i=s_h}(y_i - \overline{y}_{s_h})^2$$

slik at

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

Stratifiseringsestimatoren er identisk med Horvitz-Thompson estimatoren. Dette følger av at trekksannsynlighetene er gitt ved:

$\pi_i = n_h/N_h$ hvis enhet *i* er i stratum U_h ,

siden vi trekker et enkelt tilfeldig utvalg fra hvert stratum. Det betyr at hver observasjon i s_h skal multipliseres med vekten $1/\pi_i = N_h/n_h$ som gir oss stratifiseringsestimatoren.

Et viktig problem i stratifisering er bestemmelse av størrelsen på delutvalgene i de enkelte strata. Dvs., gitt den totale utvalgsstørrelsen *n* og gitt stratum-inndelingen, hvordan skal de *n* utvalgsenhetene *allokeres* til de forskjellige strata. Hvis vi er kun opptatt av å estimere populasjonstotalen *t* så bør n_h velges slik at $Var(\hat{t}_{st})$ blir minst mulig. Det kan vises at den optimale allokering for dette formålet er gitt ved:

$$n_h = n \cdot \frac{N_h\sigma_h\sqrt{\frac{N_h}{N_h-1}}}{\sum_{h=1}^{H} N_h\sigma_h\sqrt{\frac{N_h}{N_h-1}}} \approx n \cdot \frac{N_h\sigma_h}{\sum_{h=1}^{H} N_h\sigma_h} \ .$$

Vi ser at optimal allokering krever kjennskap til stratumvariansene σ_h^2. Dette vil aldri være tilfelle i praksis. Men ved gjentatte undersøkelser så kan det være mulig å få noen anslag for σ_h .

En vanlig form for stratifisering er å la antallet i utvalget fra hvert stratum være proporsjonalt med stratumstørrelsen. Dette kalles proporsjonal allokering eller representativ stratifisering. Dette betyr at

$$n_h = cN_h \quad \text{for alle } h, \text{ for en konstant } c.$$

Ved å summere over alla strata ser vi at vi må ha $n = cN$. Dermed : $c = n/N$ og

$$n_h = \frac{N_h}{N} n \qquad \text{eller } \frac{n_h}{N_h} = \frac{n}{N} \ .$$

Vi ser at trekksannsynlighetene er lik $\pi_i = n/N$ for alle enhetene i populasjonen. Dvs., utvalget er selvveiende. Stratifiseringsestimatoren blir nå lik ekspansjonsestimatoren :

$$\hat{t}_{st} = \sum_{h=1}^{H} \sum_{i \in s_h} \frac{N_h}{n_h} y_i = \frac{N}{n} \sum_{h=1}^{H} \sum_{i \in s_h} y_i$$

$$= N \cdot \frac{1}{n} \sum_{i \in s} y_i = N\bar{y}_s ,$$

siden utvalgssummen av y-verdiene kan fås ved først å beregne de observerte stratumtotalene, $\sum_{i \in s_h} y_i$,

og deretter summere disse.

Sammenlignet med optimal allokering ser vi at hvis stratumvariansene er omtrent like store så vil proporsjonal allokering gi tilnærmet minimum varians for estimatoren. I tilfellet proporsjonal allokering er variansen gitt ved:

$$Var(\hat{t}_{st}) = N^2 \frac{1-f}{n} \sum_{h=1}^{H} W_h \frac{N_h}{N_h - 1} \sigma_h^2 \quad , \quad W_h = N_h/N \ , f = n/N.$$

$$\approx N^2 \frac{1-f}{n} \sum_{h=1}^{H} W_h \sigma_h^2 \ .$$

Ved proporsjonal allokering så vi at $\hat{t}_{st} = \hat{t}_e = N\bar{y}_s$. Men legg merke til at $Var(\hat{t}_{st})$ ikke er lik $Var(\hat{t}_e)$ ved enkelt tilfeldig utvalg. Betegn den med $Var_{ETU}(\hat{t}_e)$. Fra kapittel 3 has at

$$Var_{ETU}(\hat{t}_e) = N^2 \cdot \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} \approx N^2 \frac{1-f}{n} \sigma^2 .$$

Det kan vises at (oppgave 4.2):

$$\sigma^2 = \sum_{h=1}^{H} W_h \sigma_h^2 + \sum_{h=1}^{H} W_h (\bar{Y}_h - \bar{Y})^2 .$$

Her bruker vi betegnelsen $\bar{Y} = t / N$.

Dermed ser vi at *uansett* valg av stratuminndeling så vil proporsjonal allokering alltid medføre mer presis estimering, $Var(\hat{t}_{st}) < Var_{ETU}(\hat{t}_e)$. Jo mer homogene strata vi har jo større variasjon vil det være mellom stratumgjennomsnittene $\bar{Y}_h$ og dermed større reduksjon av usikkerheten i estimatet. Det betyr at størst variansreduksjon oppnås hvis populasjonen er meget inhomogen, f.eks. en

populasjon av bedrifter. I det teoretiske tilfellet at alle $\overline{Y}_h$ er identisk like så vil vi tilnærmet ha samme varians i de to tilfellene. (Eksakt vil da $Var_{ETU}(\hat{t}_e)$ faktisk være ubetydelig mindre enn $Var(\hat{t}_{st})$.)

Hvis vi også ønsker å publisere estimater for stratumstørrelsene t_h så kan det medføre at for små strata så må n_h økes for at estimatoren $\hat{t}_h$ ikke skal bli for usikker. Samtidig må da n_h reduseres tilsvarende for store strata.

Eksempel 4.1. Populasjonen består av 8 land fra det amerikanske kontinentet. Vi skal estimere det totale antall innbyggere i denne populasjonen i 1983. Vi kjenner folketallene for 1980. For å kunne illustrere estimeringsegenskaper oppgir vi også 1983-tallene. I en vanlig undersøkelse ville de selvfølgelig vært ukjente på forhånd. Populasjonen av land med tilhørende 1980 og 1983 befolkningsstørrelser er gitt i tabellen nedenfor:

Land	1980 folketall (mill.)	1983 folketall (mill.)
1-Canada	24,0	24,9
2-USA	227,7	233,9
3-Mexico	69,3	75,1
4-Argentina	28,2	29,6
5-Brasil	121,3	129,6
6-Chile	11,1	11,6
7-Uruguay	2,9	2,9
8-Cuba	9,7	10,0

Vi danner nå to strata ved å bruke 1980-folketallet som stratifiseringsvariabel. Stratum 1 består av landene Mexico, Brasil og USA, mens resten danner stratum 2. Et enkelt tilfeldig utvalg på 2 land fra hvert stratum trekkes. Vi skal i første omgang sammenligne stratifiseringsestimatoren i denne utvalgsplanen og ekspansjonsestimatoren i enkelt tilfeldig utvalg med $n = 4$. Begge estimatorene er forventningsrette i sine respektive utvalgsplaner. For å beregne $Var_{ETU}(\hat{t}_e)$ og $Var(\hat{t}_{st})$ trenger vi følgende størrelser :

$$\sigma_1^2 = 4340{,}69 \quad , \quad \overline{Y}_1 = 146{,}2$$
$$\sigma_2^2 = 98{,}188 \quad , \quad \overline{Y}_2 = 15{,}8$$
$$\sigma^2 = 5674{,}48 \, .$$

Dette gir :

$$Var_{ETU}(\hat{t}_e) = N^2 \cdot \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} = 51880{,}9 \qquad SE_{ETU}(\hat{t}_e) = 227{,}8$$

$$Var(\hat{t}_{st}) = \sum_{h=1}^{2} N_h^2 \cdot \frac{\sigma_h^2}{n_h} \cdot \frac{N_h - n_h}{N_h - 1} = 10687{,}1 \qquad SE(\hat{t}_{st}) = 103{,}4 \, .$$

Usikkerheten reduseres med ca. 55% ved denne stratifiseringen. Det finnes bedre stratifiseringer i denne situasjonen. F.eks., la USA,Brasil og Mexico være egne strata med de resterende fem land i ett stratum. Trekk «tilfeldig» et land fra hvert stratum. Da er $SE(\hat{t}_{st}) = 49{,}5$ (oppgave 4.3). Ulempen med denne utvalgsplanen er at $Var(\hat{t}_{st})$ ikke kan estimeres.

I denne situasjonen er det ikke optimalt å bruke en stratifisert estimator selv om den er betydelig bedre enn ekspansjonsestimatoren. Med 1980-folketallene som tilleggsinformasjon er rate-estimatoren

$$\hat{t}_R = \frac{X_\circ}{N\bar{x}_s}\hat{t}_e$$

atskillig mer presis enn også $\hat{t}_{st}$. Med $R = t / X_\circ$ har vi

$$Var(\hat{t}_R) \approx N^2 \frac{1-f}{n} \cdot \frac{1}{N}\sum_{i=1}^{N}(y_i - Rx_i)^2 \quad = 33{,}974.$$

Dette gir at $SE(\hat{t}_R)$ = 5,8 ! Rate-estmatoren er suverent best, fordi sammenhengen mellom 1980-tallene og 1983-tallene er meget stor. For å illustrere hvordan de tre estimatorene fungerer i praksis vises 5 trekkinger av enkelt tilfeldig utvalg og tilhørende estimater fra $\hat{t}_R$ og $\hat{t}_e$, og 5 trekkinger av det stratifiserte utvalg med $\hat{t}_{st}$ -verdier. Sann *t*-verdi er 517,6.

ETU	$\hat{t}_e$	$\hat{t}_R$	Strat. etu	$\hat{t}_{st}$
2,3,7,8	643,8	513,8	(2,3) , (4,7)	544,8
1,3,5,7	465,0	528,3	(3,5) , (6,8)	361,1
1,3,6,8	243,2	526,7	(2,3) , (1,6)	554,8
3,5,7,8	435,2	529,2	(2,5) , (7,8)	577,5
5,6,7,8	308,2	525,2	(3,5) , (4,6)	410,1
gjennomsnitt	**419,1**	**524,6**		**489,6**
standardavvik	**155,0**	**6,3**		**97,3**

For at eksemplet ikke skal gi et galt inntrykk bør det presiseres at det har vist seg i praksis i Statistisk sentralbyrå at stratifisering ofte gir samme variansreduksjon som rate-estimatoren. Siden stratifisering ikke medfører skjevhet slik tilfellet kan være med rate-estimatoren så er vanligvis stratifisering å foretrekke.

4.2. Etterstratifisering

Stratifisering reduserer usikkerheten til estimatoren i forhold til enkelt tilfeldig utvalg. I mange tilfeller ønsker man å stratifisere etter variable som ikke er tilgjengelig for stratifisering på forhånd. Dvs., hvor det i praksis kan det være vanskelig eller umulig på forhånd å avgjøre hvilket stratum en enhet tilhører. For eksempel, anta at en ønsker å stratifisere etter alder. Det vil da vanligvis være mulig å få tilgang på aldersfordelingen i populasjonen, dvs. N_h = antall i aldersgruppe *h* er kjent for alle *h*, men uten et godt register som gir alderen til hver person i befolkningen vil stratifisering ikke være mulig. I slike tilfeller kan vi stratifisere *etter* at data er innsamlet, såkalt *etterstratifisering.* Deretter danner vi den vanlige stratifiserte estimatoren, nå kalt den etterstratifiserte estimatoren, $\hat{t}_{est}$:

$$\hat{t}_{est} = \sum_{h=1}^{H} N_h \bar{y}_{s_h}$$

hvor nå $\bar{y}_{s_h}$ er observert gjennomsnitt fra etterstratum *h* og N_h er størrelsen på etterstratum *h* .

Hvis vårt utvalg trekkes enkelt tilfeldig så kan det vises at vi *tilnærmet* har :

$$E(\hat{t}_{est}) = t$$

$$Var(\hat{t}_{est}) = N^2 \frac{1-f}{n} \sum_{h=1}^{H} W_h \sigma_h^2$$

Dvs., etterstratifisering i enkelt tilfeldig utvalg tilsvarer essensielt stratifisering med proporsjonal allokering. Enkelt tilfeldig utvalg med etterstratifisering er ofte atskillig bedre enn enkelt tilfeldig utvalg uten etterstratifisering. Etterstratifisering brukes blant annet i AKU (arbeidskraftundersøkelsen). Utvalgsplanen i AKU er imidlertid ikke enkel tilfeldig, men stratifisert etter fylke.

En viktig begrunnelse for etterstratifisering i en vanlig undersøkelse er at totalen for mange variable skal estimeres. Da kan det være slik at man bør bruke forskjellig stratifisering for forskjellige variable. Det er altså ikke mulig å trekke et stratifisert utvalg som tar hensyn til alle variable. Stratifisering, som f.eks. i AKU, har da heller det formål å produsere statistikk for delområder av populasjonen. Deretter kan det være aktuelt med forskjellig etterstratifisering for forskjellige *y*-variable. For eksempel, noen variable kan forklares bra ved aldersgruppering, andre ved yrkesgruppering, etc.

Etterstratumstørrelsene n_h bør være av en viss minimumsstørrelse for å unngå for ustabile stratumestimater $\bar{y}_{s_h}$. Hvis alle n_h er minst 20 så er det vanligvis tilstrekkelig.

En viktig anvendelse av etterstratifisering er å redusere skjevheten forårsaket av frafall i det planlagte utvalget. Hvis frafallet kan forklares hovedsakelig av en stratifiseringsvariabel, så vil den etterstratifiserte estimatoren rette opp mesteparten av frafallsskjevheten. Ved å bruke denne estimatoren antar vi implisitt at svardelen i hvert utvalgstratum er representativt for frafallsdelen.

Eksempel 4.2. Anta vi skal estimere arbeidsledigheten i aldersgruppen 18-30. I aldersgruppen 18-24 er det 10000 personer og i aldersgruppen 25-30 er det 5000 personer. Vi tar et enkelt tilfeldig utvalg på 500 personer. Det viser seg at 300 var i aldersgruppen 18-24 og 200 i aldersgruppen 25-30. Ved innsamling av data fikk man svar fra 200 i den yngste aldersgruppen og 180 i den andre gruppen. Resultatet av undersøkelsen ble :

	18-24 år	**25-30 år**
Arbeidsledige	50	10
Ikke arbeidsledige	150	170
Frafall	100	20

Ekspansjonsestimatet for det totale antall ledige *t* er lik $15000 \cdot \frac{60}{380} = 2368$. Hvis vi etterstratifiserer etter aldersgruppe ser vi at frafallsandelene er 33,3% for 18-24 gruppen og 10% for 25-30 gruppen. Det etterstratifiserte estimatet er gitt ved:

$$\hat{t}_{st} = 10000 \cdot \tfrac{50}{200} + 5000 \cdot \tfrac{10}{180} = 2778 .$$

Siden vi hadde størst frafall i den yngste aldersgruppen som samtidig har høyest ledighet så vil ekspansjonsestimatoren, som ikke tar hensyn til frafallsmønsteret, underestimere totalen.

I kapittel 7 skal vi behandle frafallsproblemer generelt, og vil der også komme tilbake til etterstratifiseringens rolle i forbindelse med frafall.

Oppgaver

4.1 Vi ønsker å estimere gjennomsnittslønn for en populasjon på 6 personer ved å ta et utvalg på 2 personer. For å illustrere de forskjellige estimeringsstrategiene skal vi anta at inntektene i populasjonen for personene 1-6 er gitt ved (i tusen) 250, 220, 160, 180, 140, 140 henholdsvis, med et gjennomsnitt lik 1090/6 = 181,7.

(a) Vi trekker et enkelt tilfeldig utvalg, og bruker estimatet $\bar{y}_s$. De mulig utvalg og estimater er da:

Utvalg	1,2	1,3	1,4	1,5	1,6	2,3	2,4	2,5	2,6	3,4	3,5	3,6	4,5	4,6	5,6
Estimat	235	205	215	195	195	190	200	180	180	170	150	150	160	160	140

Siden det er lik trekkesannsynlighet for alle utvalgene, kan vi finne forventningen til estimatoren ved å ta gjennomsnittet av alle estimatene. Dette inkluderer å gjenta et estimat hvis flere utvalg gir det samme estimatet. Tilsvarende finner man variansen til estimatoren.

Beregn forventning og varians til $\bar{y}_s$.

(b) For å minske variansen til estimatoren stratifiserer vi utvalget med hensyn til alder, idet vi antar at det er en sammenheng mellom alder og inntekt. Alderen til personene 1-6 i populasjonen er 37, 62, 42, 26, 18, og 21 henholdsvis. Populasjonen deles inn i to strata; stratum 1 = «over 30 år», stratum 2 = «under 30 år». Vi trekker et stratifisert enkelt tilfeldig utvalg med 1 person fra hvert stratum. Siden antallet i utvalget fra hvert stratum er proporsjonalt med stratumstørrelsen så er stratifiseringsestimatoren også lik $\bar{y}_s$.

Lag en liste over de mulige utvalgene (9 ialt) og estimatene, og finn forventning og varians til $\bar{y}_s$. Sammenlign med egenskapene til $\bar{y}_s$ i enkelt tilfeldig utvalg. Hvilken utvalgsplan er å foretrekke ?

4.2* Betrakt stratifisert populasjon med populasjonsvarians σ^2 , stratumvarianser σ_h^2 og stratumgjennomsnitt $\overline{Y}_h$. La $W_h = N_h/N$, og $\overline{Y} = t / N$. Vis at

$$\sigma^2 = \sum_{h=1}^{H} W_h \sigma_h^2 + \sum_{h=1}^{H} W_h (\overline{Y}_h - \overline{Y})^2 .$$

4.3 Betrakt eksempel 4.1. La USA, Brasil og Mexico være egne strata, med de resterende fem land samlet i et fjerde stratum. Anta vi trekker «tilfeldig» et land fra hvert stratum. Vis at $SE(\hat{t}_{st}) = 49{,}5$.

4.4 Viser til oppgave 1.1. Vi skal estimere antall yrkesaktive i Etnedal(som vi vet er 797), ved å trekke et stratifisert enkelt tilfeldig utvalg. Stratifiseringen gjøres etter alder. Vi har tre strata:

U_1 = 16- 18 år, U_2 = 19-65 år, U_3 = over 65 år.

Fra FoB90 :

Stratum	Antall yrkesaktive	Totalt
16-18 år	28	59
19-65 år	729	916
over 65 år	40	355

(a) Vi skal trekke et stratifisert utvalg på 133 personer, hvor antallet som trekkes fra hvert stratum skal være proporsjonalt med stratumstørrelsen. Bestem antallet som skal trekkes fra hvert stratum.

Vi gjennomfører en trekning med følgende resultat :

Stratum	antall yrkesaktive
16-18 år	3
19-65 år	70
over 65 år	5

(b) Beregn estimatet, og dets teoretiske forventning og standardfeil.

(c) Bestem standardfeilen til $\hat{t}_e = N\bar{y}_s$ ved et enkelt tilfeldig utvalg på 133 personer. Hvilken utvalgsplan er å foretrekke ?

4.5 Viser til oppgavene 1.1 og 4.4. Vi skal estimere andel yrkesaktive i Etnedal for de to aldersgruppene 16-24 år og over 24 år (dvs. 25 år og eldre). Delpopulasjonene ser slik ut, i virkeligheten :

	16-24 år	Over 24 år
Antall yrkesaktive	138	659
Totalt	197	1133

(a) Vi stratifiserer etter de to aldersgruppene, og trekker et 10% utvalg proporsjonalt med størrelsen på strataene. Dvs. 20 personer fra aldersgruppen 16- 24, og 113 personer over 24 år. Vi estimerer andelen yrkesaktive i de to gruppene på vanlig måte. Finn standardfeil for de to estimatorene.

(b) Et vanlig krav til tall som publiseres er at standardfeilen ikke er over en viss prosent av estimatet. Anta at kravet er at standardfeilen ikke skal overskride 10% av estimatet. Anta estimatene er lik 0,65 og 0,60 for 16-24 år og over 24 år. Er våre estimater pålitelige nok til å bli publisert?

(c) Anta at vi istedenfor å trekke proporsjonalt med stratastørrelsene trekker like mange fra hvert stratum, dvs. 66 fra hvert stratum. Tilfredsstiller våre estimater 0,65 og 0,60 nå kravet fra (b) ?

4.6 Igjen skal vi estimere antall yrkesaktive. Anta vi trakk et 10% enkelt tilfeldig utvalg fra befolkningen over 15 år og fikk 100 svar, dvs. et frafall på 33. I utvalget på 133 personer var det 75 menn og 58 kvinner, med 45 svar fra menn og 55 svar fra kvinner , dvs. frafallet var betydelig større hos menn enn hos kvinner. Av de 45 menn svarte 32 at de var yrkesaktive, og blant de 55 kvinnene svarte 27 at de var yrkesaktive. Av de 1330 personene over 15 år var det 705 menn.

(a) Bestem det vanlige estimatet for antall yrkesaktive i befolkningen uten å bruke det vi vet om svarfordelingen mellom menn og kvinner.

(b) Etterstratifiser etter kjønn og bruk $\hat{t}_{est}$.

4.7 Vi skal se på en situasjon hvor vi på forhånd bare kjenner totaltallene og ikke verdiene for enkeltpersoner, slik at etterstratifisering kan utføres, men ikke en stratifisert utvalgsplan. Det gjelder valg hvor hver persons stemmegivning er hemmelig mens totaltallene er offentlige.

Anta at vi holdt en meningsmåling i Etnedal i 1994 om holdningene til norsk EU medlemskap. Av de 150 som ble spurt svarte 105 at de var motstandere, 20 at de var tilhengere og 25 var usikre/visste ikke.

Det er kjent at holdningen til norsk EU medlemskap kan ha sterk sammenheng med partipreferanse. Av den grunn ble også personene i utvalget spurt om hvilket parti de stemte på ved Stortingsvalget i 1993, med følgende resultat:

Parti	**AP**	**FrP**	**H**	**KrF**	**SP**	**SV**	**Andre**	**Stemte ikke**
EU-tilhenger	9	2	5	2	1	0	0	1
EU-motstander	10	0	1	1	85	7	0	1
Vet ikke/usikker	5	1	3	1	3	1	1	10

Stemmegivningen i Etnedal ved Stortingsvalget i 1993 :

Parti	**AP**	**FrP**	**H**	**KrF**	**SP**	**SV**	**Andre**	**Stemte ikke**
Prosent	23,6	2,5	4,9	6,8	28,8	6,3	3,2	23,9

Legg merke til at tabellen inneholder hele befolkningen med stemmerett (ialt 1273 personer), og ikke bare de som stemte. Dette er gjort fordi vi ønsker å si noe om EU-synet til hele befolkningen.

(a) Beregn estimater både med og uten etterstratifisering etter partitilhørighet, for antall EU-tilhengere, motstandere og tvilere, både som totaltall og prosenter.

(b) Hvilket av estimatene i (a) ville du i praksis stole mest på?

(c) Når et parti er i framgang er det en kjent effekt i forbindelse med meningsmålinger at noen flere, enn de som faktisk stemte på partiet, svarer at de stemte på dette partiet ved forrige valg. Anta at det i vår situasjon ble en slik overrepresentasjon av personer som sa de hadde stemt SP, og at det store flertallet av disse «nye» SP personene var EU-motstandere. Ville dette påvirke vårt estimat, og isåfall på hvilken måte ?

(d) I forbindelse med meningsmålinger som blir tatt over telefon er det kjent at man gjerne får en overrepresentasjon av kvinner i utvalget. Vil dette kunne være et problem, og hvordan kan man isåfall løse dette problemet?

4.8 Etterstratifisering kan brukes til å forbedre registerdata. Anta vi har en populasjon på 1500 personer, og vi ønsker å bestemme antallet som er gift. Vi har to datakilder, et register som ikke er oppdatert og et enkelt tilfeldig utvalg på 150 personer. I registeret er 800 registrert som gifte, mens i utvalget fant man at 50 av de 150 for tiden er gift. Observasjonene fordeler seg som vist i tabellen

	Gift ifølge register	Ikke gift ifølge register
Gift i utvalg	40	10
Ikke gift i utvalg	20	80

To mulige metoder :

1) Bruk kun utvalget, siden registeret er gammelt. Det gir et estimat på 1/3 gifte, dvs. 500 gifte totalt.

2) Bruk kun registeret, siden utvalget er lite (og tilsynelatende skjevt). Estimatet blir da 800.

Vi skal nå se hvordan etterstratifisering (etter registerinndelingen) kombinerer registerinformasjon med de observerte data.

(a) Beregn den etterstratifiserte estimatoren for antall gifte i populasjonen, og sammenlign med anslagene ovenfor.

(b) Hvilket estimat er mest pålitelig? Begrunn svaret.

5. Ett- og flertrinnsutvalg

De typer utvalgsplaner vi har sett på så langt er såkalte ett-trinnsutvalg, dvs. utvalget av enheter trekkes direkte i ett trinn. Det gjelder enkelt tilfeldig utvalg, intervalltrekking, stratifisert utvalg, og forskjellige utvalgsplaner med en form for proporsjonal trekking (med hensyn til enhetenes størrelse). Av økonomiske og praktiske hensyn er det ofte nødvendig å modifisere disse utvalgsplanene. Det kan gjøres enten ved at utvalget trekkes i flere trinn eller ved at enhetene trekkes indirekte i grupper i ett trinn. To årsaker til at man ofte ikke kan bruke direkte enhetstrekking er:

(i) Det eksisterer ikke et register over alle enhetene i populasjonen, og det vil være umulig eller meget kostbart å opprette et slikt register. Samtidig kan man muligens identifisere grupper av enheter.

(ii) Enhetene i populasjonen er spredt over et stort område. Direkte trekking av enhetene i utvalget vil da resultere i et utvalg som også er spredt over et stort område. Ved besøksundersøkelser vil dette medføre meget høye reisekostnader.

Det finnes mange forskjellige utvalgsplaner som kan anvendes i situasjoner hvor det er umulig eller upraktisk med direkte enhetstrekking. Vi skal konsentrere oss i dette kapitlet hovedsakelig om to typer av slike utvalgsplaner, *klyngeutvalg* og *to-trinnsutvalg*.

I klyngetrekking grupperes populasjonen i delpopulasjoner kalt klynger. Et utvalg av klynger trekkes og *alle* enhetene i de uttrukne klyngene blir med i utvalget.

Eksempel 5.1. Anta vi ønsker å trekke et utvalg av elever i niende klasse for å undersøke alkoholforbruk. Da kan man trekke et utvalg av niende-klasser og deretter dele ut spørreskjema til alle elevene i de klassene som ble trukket ut. Her er populasjonen alle elever i niende klasse og niende-klassene er klynger.

Ved to-trinnsutvalg grupperes igjen populasjonen i delpopulasjoner, nå kalt *primære utvalgsenheter (PUE).* Trekking av enheter foregår i to trinn:

(I) Trinn 1 består i å trekke et utvalg av PUE .

(II) Trinn 2: For hver PUE som er trukket ut på trinn 1 trekkes et utvalg av enheter, nå også kalt *sekundære utvalgsenheter* (*SUE*).

For eksempel, hvis en skal trekke et landsomfattende utvalg av personer, kan dette gjøres ved at man først trekker et utvalg av kommuner, og deretter trekker utvalg av personer innen de uttrukne kommunene. Dette gjøres i SSB's generelle utvalgsplan for besøksundersøkelser, som beskrives i kap.6.

I forbindelse med besøksundersøkelser har ovenstående to-trinns utvalgplan den fordel at den reduserer innsamlingskostnadene vesentlig. Et utvalg av kommuner trekkes med sjeldne mellomrom, f.eks. hvert tiende år. De utvalg som skal brukes i forbindelse med besøksundersøkelser trekkes deretter innen de uttrukne kommuner. Ved å ansette en eller flere intervjuere innen de uttrukne kommuner reduserer en det geografiske området en bestemt intervjuer må dekke, og reduserer derved reisekostnadene.

Som i tilfellet med stratifisering står planleggeren fritt når det gjelder å velge hvorledes populasjonen skal inndeles i PUE, og hvilke metoder som skal brukes ved trekking av PUE og deretter enheter innen de uttrukne PUE.

Eksempel 5.2 Anta, som i eksempel 5.1, at vi ønsker å trekke et utvalg av elever i niende klasse. Hvis vi trekker et utvalg av elever fra hver klasse som ble trukket ut, er dette et eksempel på to-trinnsutvalg med klassene som PUE og elevene som SUE. En alternativ utvalgsplan for denne type av undersøkelser er følgende: (a) Fra et register over alle skolene i landet trekkes først et utvalg av skoler;

(b) Fra hver skole i utvalget trekkes et utvalg av niende-klasser, og alle elevene i de uttrukne klassene blir med i utvalget. Dette er et eksempel på to-trinns klyngeutvalg med skoler som PUE og klasser som SUE. Hvis vi isteden trekker et utvalg innen hver uttrukne klasse har vi et eksempel på tre-trinnsutvalg.

Vi skal se på problemet med å estimere populasjonstotalen t for en variabel y. Dette er ikke lenger generelt ekvivalent med å estimere populasjonens gejnnomsnitt t/N siden N kan være ukjent i denne type undersøkelser.

5.1 Klyngeutvalg

Populasjonen er inndelt i K klynger ; U_1 , . . . , U_K . La N_i være antall enheter i klynge U_i . Klyngeutvalg trekkes nå på følgende måte :

1. Et sannsynlighetsutvalg s_I av klynger trekkes. Størrelsen på utvalget betegnes med n_I .

2. Det endelige utvalget s består av alle enhetene i de utvalgte klyngene.

Et eksempel på klyngeutvalg i Statistisk sentralbyrå er AKU (arbeidskraftundersøkelsen) hvor en først trekker et utvalg familier og deretter intervuer alle familiemedlemmer.

Vi merker oss at størrelsen n på utvalget s ikke er gitt på forhånd men er selv en tilfeldig variabel siden n avhenger av hvilket utvalg s_I som trekkes . Vi har at n er lik summen av N_i for enhetene i utvalget s_I, som vi kan uttrykke

$$n = \sum_{i \in s_I} N_i \quad .$$

La π_{Ii} være trekksannsynligheten for klynge U_i . Observasjonen for hver klynge U_i er da summen t_i av alle y-verdiene for enhetene i klyngen ;

$$t_i = \sum_{k \in U_i} y_k \ .$$

Trekksannsynlighetene for enhetene er gitt ved :

$$\pi_k = \pi_{Ii} \text{ hvis enhet } k \text{ ligger i klynge } U_i \ .$$

Det betyr at alle y-verdier i en utvalgt klynge får samme vekt $1/\pi_{Ii}$ med Horvitz-Thompson estimatoren. Den delen av totalsummen av alle (y_k/π_k) for alle uttrukne enheter blir dermed lik t_i/π_{Ii}. Horvitz-Thompson estimatoren, som vi nå betegner med $\hat{t}_{Kl}$, blir dermed :

$$\hat{t}_{Kl} = \sum_{i \in s_I} \frac{t_i}{\pi_{Ii}} \ .$$

Enkelt tilfeldig klyngeutvalg

Et enkelt tilfeldig utvalg på n_I klynger trekkes blant de K klyngene. Da er trekksannsynlighetene lik

$\pi_{Ii} = n_I/K$, og Horvitz-Thompson estimatoren blir lik ekspansjonsestimatoren basert på klyngetotalene t_i i utvalget s_I :

$$\hat{t}_{Kl} = K \cdot \frac{1}{n_I} \sum_{i \in s_I} t_i = K \cdot \bar{t}_{s_I} \ .$$

Denne er forventningsrett siden Horvitz-Thompson estimatoren alltid er forventingsrett (når alle enhetene har positive trekksannsynligheter). Variansen er gitt ved

$$Var(K \cdot \bar{t}_{s_I}) = K^2 \frac{\sigma_t^2}{n_I} \cdot \frac{K - n_I}{K - 1} \approx K^2 \frac{\sigma_t^2}{n_I} \cdot (1 - f_I)$$

hvor $f_I = n_I / K$ og $\sigma_t^2 = \frac{1}{K} \sum_{i=1}^{K} (t_i - \bar{t}_K)^2$, $\bar{t}_K = \sum_{i=1}^{K} t_i / K.$

En estimator for $Var(K\bar{t}_{s_I})$ oppnås ved å estimere σ_t^2 med $\hat{\sigma}_t^2 = \frac{K-1}{K} \cdot \frac{1}{n_I - 1} \sum_{i \in s_I} (t_i - \bar{t}_{s_I})^2$.

Måten en konstruerer klyngene på har en stor innflytelse på variansen til estimatoren. Det gjelder å få σ_t^2 så liten som mulig og det oppnås ved å velge klyngene mest mulig *heterogene*, dvs. med stor spredning på *y*-verdiene. Da vil mest mulig av den totale y-variasjonen ligge innenfor klyngene og t_i -verdiene blir mest mulig like. Legg merke til at det motsatte var tilfellet med en stratifisert utvalgsplan. Vanligvis vil klyngeutvalg medføre større varians på estimatoren enn enkelt tilfeldig utvalg med ekspansjonsestimatoren, men samtidig vil kostnadene ved undersøkelsen reduseres.

Eksempel 5.3. Anta vi har en populasjon på fire husholdninger gruppert i to klynger med inntektene 200.000 og 300.000 i klynge 1, og inntektene 400.000 og 600.000 i klynge 2. En ønsker å trekke et utvalg på 2 husholdninger for å estimere den totale inntekten *t (= 1.500.000)* ved å trekke tilfeldig en klynge. Her er det to mulige utvalg med estimater lik verdier av $2 \cdot t_i$ for i =1,2 : 1.000.000 og 2.000.000. Variansen til estimatoren er lik = $25 \cdot 10^{10}$, med $SE(2 \cdot t_i) = 500.000$. Ekspansjonsestimatoren i enkelt tilfeldig utvalg med $n = 2$ gir $SE(\hat{t}_e) = 341.565$. Mer heterogene klynger oppnår vi ved å la klynge 1 bestå av husholdningene med inntektene 200.000 og 600.000. Da blir $SE(2 \cdot t_i) = 100.000$.

5.2 To-trinnsutvalg

Variansen til Horvitz-Thompson estimatoren i klyngeutvalg kan alltid reduseres ved å trekke ut flere klynger. Men det kan medføre for høye kostnader. En måte å kontrollere kostnadene på og samtidig øke antall utvalgte klynger er å trekke et utvalg fra hver uttrukne klynge. Klyngetotalene t_i må da estimeres fra utvalgene. Hvis variasjonen innen klyngene er liten så vil estimatorene $\hat{t}_i$ ha liten varians. Klyngene kalles nå primære utvalgsenheter (PUE) og vi har to-trinnstrekking. Det vil ofte lønne seg å bruke to-trinnsutvalg istedenfor klyngeutvalg.

Populasjonen er inndelt i K primære utvalgsenheter ; U_1 , . . . , U_K . PUE U_i består av N_i enheter, også kalt sekundære utvalgsenheter (SUE). Den generelle to-trinn utvalgsplanen kan beskrives på følgende måte :

Trinn 1. Et sannsynlighetsutvalg s_I av PUE trekkes. Størrelsen på utvalget betegnes med n_I .

Trinn 2. For hver $i \in s_I$ trekkes et utvalg s_i av n_i SUE .

Den totale utvalgsstørrelsen blir da summen av utvalgsstørrelsene n_i for de utvalgene som trekkes på trinn 2, dvs.:

$$n = \sum_{i \in s_I} n_i \quad .$$

La π_{Ii} være trekksannsynligheten for primær utvalgsenhet U_i . For 2.trinnstrekking, la $\pi_{k|i}$ være den betingede trekksannsynligheten for enhet k i PUE U_i gitt at U_i er trukket ut på første trinn.

Trekksannsynligheten for enheten k i U_i er gitt ved :

π_k = P(U_i trekkes på trinn 1 $\cap$ enhet k trekkes på trinn 2)

= P(U_i trekkes på trinn 1)P(enhet k trekkes på trinn 2 $\mid$ U_i trekkes på trinn 1)

$= \pi_{Ii}\, \pi_{k|i}$.

Det betyr at y-verdiene i utvalget s_i får vektene $(1/\pi_{Ii}\pi_{k|i})$ med Horvitz-Thompson estimatoren. Den delen av totalsummen av alle (y_k/π_k), for alle uttrukne enheter, blir dermed lik

$$\sum_{k \in s_i} \frac{y_k}{\pi_{Ii}\pi_{k|i}} = \frac{1}{\pi_{Ii}} \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} .$$

Horvitz-Thompson estimatoren for PUE totalen t_i basert på 2.trinnsutvalget s_i er gitt ved

$$\hat{t}_{i,HT} = \sum_{k \in s_i} \frac{y_k}{\pi_{k|i}} \quad .$$

Horvitz-Thompson estimatoren for 2-trinnsutvalg betegnes nå med $\hat{t}_{2T}$ og er dermed lik

$$\hat{t}_{2T} = \sum_{i \in s_I} \frac{\hat{t}_{i,HT}}{\pi_{Ii}} \ ,$$

sammenlignet med «klynge-estimatoren» $\hat{t}_{Kl} = \sum_{i \in s_I} \frac{t_i}{\pi_{Ii}}$.

Den vanligste måten å velge trekkemetodene i trinn 1 og trinn 2 på er å sørge for at alle π_k er like, dvs. at

$$\pi_{Ii}\, \pi_{k|i} = n/N.$$

Slike utvalg kalles *selvveiende*. Da blir Horvitz-Thompson estimatoren lik ekspansjonestimatoren.

Eksempel 5.4. Populasjonen består av ansatte i 4 bedrifter. Bedriftene er PUE og de ansatte er SUE. Bedriftene 1,2,3,4 har 10,20,100 og 120 ansatte, og N = 250. Vi skal se på tre forskjellige to-trinnsutvalgsplaner :

- Utvalgsplan 1: n = 10 .

  Trinn 1 : Trekk en bedrift rent tilfeldig.

  Trinn 2 : Enkelt tilfeldig utvalg på 10 ansatte.

- Utvalgsplan 2: n = 10. Selvveiende utvalg på 10 ansatte (alle π_k = 10/250 = 0,04) .

Trinn 1 : Velg en bedrift

Trinn 2: Enkelt tilfeldig utvalg på 10 ansatte.

- Utvalgsplan 3 : 10% selvveiende utvalg, n = 25 (alle π_k = 25/250 = 0,1).

Trinn 1 : Stratifiser bedriftene i to strata, med stratum 1 = bedriftene 1,2. Fra hvert stratum trekkes en bedrift proporsjonalt med bedriftens størrelse.

Trinn 2 : Enkelt tilfeldig utvalg fra hver valgte bedrift.

Utvalgsplan 1 : π_{Ii} = ¼ for i = 1,2,3,4, og $\pi_{k|i}$ = $10/N_i$ for $k \in U_i$, slik at enhetenes trekksannsynligheter blir

$$\pi_k = \frac{10}{4N_i} \text{ hvis enhet } k \text{ er i PUE } U_i$$

= 1/4 for bedrift 1
= 1/8 for bedrift 2
= 1/40 for bedrift 3
= 1/48 for bedrift 4.

To-trinnsestimatoren for en variabel y blir, når bedrift i velges på trinn 1 :

$$\hat{t}_{2T} = 4\sum_{k \in s_i} \frac{N_i}{10} y_k = 4N_i \bar{y}_{s_i} .$$

Denne estimatoren virker ikke særlig rimelig.

Utvalgsplan 2 : Vi må bestemme hvordan trekkingen skal skje på første trinn. Det er bestemt at π_k = n/N og $\pi_{k|i}$ = $10/N_i$ slik at $\pi_{Ii}\frac{10}{N_i} = \frac{10}{N}$. Det betyr at π_{Ii} = N_i/N ; bedriften trekkes proporsjonalt med bedriftens størrelse. Estimatoren blir nå lik ekspansjonsestimatoren; $\hat{t}_{2T} = N\bar{y}_{s_i}$.

Utvalgsplan 3 : Problemet er her å bestemme utvalgsstørrelsene på trinn 2 fra hvert stratum. I oppgave 5.1 skal det vises at for i = 1,2 så er n_i = 3, og dermed n_i = 22 for i = 3,4.

To-trinnsestimatoren er igjen lik ekspansjonsestimatoren ; hvis bedriftene i og j trekkes ut på 1.trinn kan vi uttrykke den på følgende form :

$$\hat{t}_{2T} = N \cdot \frac{1}{n}\left(\sum_{k \in s_i} y_k + \sum_{k \in s_j} y_k\right) .$$

Et generelt uttrykk for $Var(\hat{t}_{2T})$:

$$Var(\hat{t}_{2T}) = Var\left(\sum_{i \in s_I} \frac{t_i}{\pi_{Ii}} \right) + \sum_{i=1}^{K} \frac{Var(\hat{t}_{i,HT})}{\pi_{Ii}} \ .$$

Den første komponenten uttrykker usikkerheten ved at vi trekker et *utvalg* av PUE, og er varians-komponenten på trinn 1. Vi kjenner den igjen som variansen til Horvitz-Thompson estimatoren med $t_i, i \in s_I$ som observasjoner. Den måler hvor godt $t_i, i \in s_I$ kan estimere totalen t . Den andre komponenten er 2.trinnsvariansen, og er et uttrykk for hvor godt vi klarer å estimere hver enkelt PUE total t_i .

To-trinnstrekking vil vanligvis gi større varians enn tilsvarende ett-trinnstrekking med samme trekksannsynligheter π_k. Vi skal foreta noen konkrete sammenligninger i forbindelse med SSB's generelle utvalgsplan for besøksundersøkelser. Ofte er de praktiske og økonomiske fordelene, imidlertid, så store at man kan leve med den økte usikkerheten.

Oppgaver

5.1 Betrakt utvalgsplan 3 i eksempel 5.4. Vis at utvalgsstørrelsene på trinn 2 er gitt ved:

For i = 1,2 : n_i = 3
For i = 3,4 : n_i = 22 .

6. Statistisk sentralbyrås generelle utvalgsplan

Dette er en generell utvalgsplan for besøksundersøkelser; både for husholdningsundersøkelser og personundersøkelser. Den brukes for følgende besøksundersøkelser i Statistisk sentralbyrå:

- omnibus
- helseundersøkelsen
- levekårsundersøkelsen
- forbruksundersøkelsen
- prisinnsamling
- kultur- og medieundersøkelsen .

Den generelle utvalgsplanen ble tidligere brukt også ved telefon- og postale undersøkelser. Nå trekkes de fleste utvalg til slike undersøkelser i ett trinn. Den opprinnelige versjonen har vært i bruk siden 1975, men har gjennomgått en del justeringer blant annet på grunn av kommune-endringer. I 1995 ble det foretatt en større revidering, blant annet for å kunne gi fylkesbasert statistikk. Utvalgsplanen er en to-trinnstrekking med et stratifisert utvalg på trinn 1 og enkelt tilfeldig utvalg på trinn 2. Landet er først delt i primære utvalgsenheter som stort sett består av kommuner. Disse kalles nå *primære utvalgsområder* . De primære utvalgsområdene er deretter stratifisert etter kommunetype og størrelse. Innen hvert stratum trekkes ett primært utvalgsområde proporsjonalt etter størrelse (antall innbyggere). På 2.trinn trekkes personer enkelt tilfeldig innen hvert utvalgte primærområde. Vi skal nå gi en detaljert beskrivelse av utvalgsplanen og starter med å angi de viktigste forutsetningene:

Forutsetninger for utvalgsplanen :

(i) Utvalgsplanen skal gi grunnlag for undersøkelser av varierende art .

(ii) Antall intervjuere som brukes i forbindelse med en undersøkelse skal være ca. 135.
(beregnet ut fra årlig arbeidsmengde)

(iii) Det viktigste er å kunne publisere tall for hele landet, men for større undersøkelser bør utvalget være slik at man gi tall for mindre geografiske områder, spesielt fylkestall.

(iv) Utvalget skal være selvveiende hvis det er mulig.

6.1 Valg og trekking av primære utvalgsområder

Økonomiske og praktiske hensyn tilsier at utvalget trekkes i to trinn. Ved konstruksjonen av de primære utvalgsområdene (PU) på trinn 1 var følgende momenter avgjørende:

(i) Størrelsen på PU samt fordelingen av intervjuobjektene(IO) på intervjuerne er de viktigste valg.

(ii) Ut fra ønsket om minst mulig varians bør en velge mange små PU. Ut fra et økonomisk og praktisk synspunkt er det derimot en fordel å konsentrere intervjuerne så mye som mulig.

(iii) PU må være identifiserbare i de registre som som er tilgjengelige.

Ut fra en samlet vurdering ble det bestemt å bruke kommuner som primære utvalgsområder. Små kommuner slås sammen med andre kommuner, slik at det (vanligvis) er minst 3000 personer i hver PU. Dette for å forhindre at området tømmes for IO på kort tid. Dessuten ble alle PU delt i 3 områder. I mindre undersøkelser trekkes utvalget i tre trinn. Først utvalgsområde, deretter del av utvalgsområdet og tilslutt enhetene til undersøkelsen.

På trinn 1 trekkes et stratifisert utvalg av to grunner :

(a) redusering av estimeringsvariansen (også kalt *utvalgsvariansen*)

(b) produsering av regional statistikk .

Stratifiseringen

Punkt (a) betyr at strataene bør være mest mulig homogene. De mest brukte stratifiseringsvariablene er innbyggerantall og geografisk beliggenhet. Stratifiseringen bestemmer de minste områder en kan gi tall for. En hovedgrunn for revideringen i 1995 var å kunne gi fylkestall. Hvert fylke er derfor stratifisert for seg. Ut fra ønsket om å redusere estimeringsvariansen så bør en lage så mange strata som mulig, og trekke ett primært utvalgsområde fra hvert stratum. En ulempe ved denne formen for stratifisert trekking er at det blir vanskelig å estimere $Var(\hat{t}_{2T})$. På tross av det valgte en å lage flest mulig strata og trekke ett PU innen hvert stratum.

De primære utvalgsområdene ble inndelt i 109 strata. Det minste stratum omfatter 10 242 personer og det største 477 781. Gjennomsnittlig stratumstørrelse er 39 677 (innbyggertallene er pr. 1.1.94). Byer med flere enn 30 000 innbyggere, samt noen få i tillegg, er stratifisert slik at de hver for seg utgjør det eneste PU i stratumet. Disse PU er da trukket ut med sikkerhet. Ellers, fra hvert stratum ble en PU trukket proporsjonalt med størrelse. Det uttrukne PU skal normalt dekkes av én intervjuer. Trekkingen av PU ble foretatt i desember 1994 for en tiårsperiode. Intervjuerne ble ansatt etter desember 1994 til å dekke de uttrukne PU. Siden de fleste intervjuerne har ansvar for kun en PU blir reisekostnadene kraftig redusert i forhold til direkte trekking fra hele populasjonen.

Som et eksempel på stratifiseringen i et fylke , skal vi se på Østfold som er delt inn i 6 strata (kommuner med «/» mellom er slått sammen til en PU). Tallene i parentes er innbyggertall pr. 1.1.94 og antall intervjuere til å betjene stratumet.

Stratum 1: Fredrikstad/Hvaler (68 207, 2)

Stratum 2: Sarpsborg (46 381, 2)

Stratum 3: Halden (25 908, 1)

Stratum 4 : Moss (25 071, 1)

Stratum 5 : Spydeberg, Askim, Råde, Rygge, Våler, Hobøl (43 619, 1)

Stratum 6 : Trøgstad, Eidsberg, Skiptvedt, Rakkestad, Aremark/Marker/Rømskog (29 526, 1)

Strataene 1-4 består alle av et PU, og dermed er disse trukket ut. Fra stratum 5 ble Rygge trukket ut med trekksannsynlighet π_{1i} = 12189/43619 = 0,28, og fra stratum 6 ble Eidsberg trukket ut med trekksannsynlighet π_{1i} = 9156/29526 = 0,31. Rygge er nest størst (Askim har 12 826 innbyggere) og Eidsberg er størst innenfor sine respektive strata.

6.2 Utvalgsplanen

Trinn 1 : Innenfor hvert stratum trekkes ett primært utvalgsområde proporsjonalt med størrelse. Hvis N_h er det totale innbyggertallet i stratum h, og $N_{h,i}$ er innbyggertallet for PU i i stratum h så trekkes PU i med sannsynlighet $\pi_{Ii} = N_{h,i}/N_h$. La s_I være utvalget av PU. I en tiårsperiode holdes s_I fast. Det nåværende utvalg av PU ble som nevnt trukket i desember 1994 hvor $N_{h,i}$ og N_h er folketallene pr. 1.1.94.

Trinn 2 : For hver PU $i \in s_I$: Trekker et enkelt tilfeldig utvalg s_i på n_i personer. Utvalgsstørrelsen n_i bestemmes slik at det totale utvalg på n personer blir selvveiende.

Trekksannsynligheten på 2.trinn er

$$\pi_{k|i} = n_i / N_{h,i}$$

for stratum h. Dermed has at

$$\pi_k = \pi_{Ii}\pi_{k|i} = n_i / N_h \ .$$

Selvveiende utvalg betyr at $\pi_k = n/N$. Det gir at

$$\frac{n_i}{N_h} = \frac{n}{N} \quad \text{og} \quad n_i = \frac{n}{N} N_h \tag{6.1}$$

proporsjonal med stratumstørrelsen. Det betyr at enhetene i de mindre utvalgte PU får høyere trekksannsynligheter på 2. trinn.

Kommentar

Trinn 1 er foretatt i desember 1994. Hvis vi skal utføre en undersøkelse senere, f.eks. i 1999, så vil innbyggertallene ha endret seg noe. I tillegg er vi vanligvis interessert i den delen av befolkningen som er over 15 år. Dermed blir trekksannsynlighetene på 2. trinn noe annerledes, og bestemmelsen av n_i kan bli endret litt. Som regel vil imidlertid (6.1) være en god nok tilnærming til å kunne brukes i praksis. For å se det, la $N^{99}, N_h^{99}, N_{h,i}^{99}$ være henholdsvis totalen, stratumstørrelsene og PU-størrelsene for befolkningen av interesse i 1999. Trekksannsynlighetene på 2. trinn blir da :

$$\pi_{k|i} = n_i \ / \ N_{h,i}^{99} \ .$$

Et selvveiende utvalg betyr dermed at :

$$\frac{N_{h,i}}{N_h} \cdot \frac{n_i}{N_{h,i}^{99}} = \frac{n}{N^{99}}, \quad \text{dvs.} \quad n_i = \frac{n}{N^{99}} \cdot N_h \frac{N_{h,i}^{99}}{N_{h,i}} = n \frac{N_h}{N} \cdot \Big[\frac{N_{h,i}^{99} \ / \ N_{h,i}}{N^{99} \ / \ N} \Big] . \tag{6.2}$$

Vanligvis så has at $\dfrac{N_{h,i}^{99} \ / \ N_{h,i}}{N^{99} \ / \ N} \approx 1$, slik at (6.1) kan benyttes.

F.eks., hvis $N_{h,i}^{99} \ / \ N_{h,i} = c$ for alle i,h så er for alle h

$$N_h^{99} = \sum_i N_{h,i}^{99} = c\sum_i N_{h,i} = cN_h \ ,$$

og dermed $N^{99} = cN$, dvs. $N^{99} / N = c$ og

$$n_i = \frac{n}{N} N_h \ .$$

Man kan også bruke

$$n_i = n\frac{N_h^{99}}{N^{99}} \tag{6.3}$$

som en tilnærming til (6.2), siden fra (6.2)

$$n_i = n\frac{N_h^{99}}{N^{99}} \cdot [\frac{N_{h,i}^{99} / N_{h,i}}{N_h^{99} / N_h}] \quad \text{og} \quad \frac{N_{h,i}^{99} / N_{h,i}}{N_h^{99} / N_h} \approx 1 \ .$$

Eksempel 6.1. La oss si at vi har en undersøkelse hvor vi skal trekke et landsutvalg på 5000 personer over 15 år (16 år og eldre). Pr. 1.1.99 besto denne populasjonen av 3 511 082 personer. Betrakt stratum 6 i Østfold ovenfor. Pr. 1.1.99 hadde stratum 6 i alt 29 933 innbyggere. Av disse var 24 132 over 15 år . Det uttrukne PU, Eidsberg hadde 7552 innbyggere over 15 år pr. 1.1.99. Det totale innbyggertallet i Norge 1.1.94 var 4 324 815. Vi har dermed følgende tall:

1999-undersøkelsen	$N^{99} = 3\ 511\ 082$	$N_h^{99} = 24\ 132$	$N_{h,i}^{99} = 7\ 552$
1994-tall	$N = 4\ 324\ 815$	$N_h = 29\ 526$	$N_{h,i} = 9\ 156$

Tilnærmingen (6.1) gir:

$$n_i = 5000\frac{29526}{4324815} = 34{,}136 = 34.$$

Forholdet $\frac{N_{h,i}^{99} / N_{h,i}}{N^{99} / N} = \frac{0{,}8248}{0{,}8118} = 1{,}016$. Dermed blir eksakt bestemmelse lik:

$$n_i = 34{,}136 \cdot 1{,}016 = 34{,}68 = 35,$$

mens tilnærmingen (6.3) gir

$$n_i = 5000\frac{24132}{3511082} = 34{,}37 = 34.$$

Siden begge valg av n_i kun tilnærmet gir et selvveiet utvalg, så kan vi like godt bruke (6.1) eller (6.3) til å bestemme utvalgsstørrelsene innen hvert valgte PU.

Estimering av populasjonstotalen

Siden vi har et selvveiende utvalg blir to-trinnsestimatoren lik ekspansjonsestimatoren ;

$$\hat{t}_{2T} = N \cdot \bar{y}_s \ .$$

Her betegner *s* det endelige utvalget av personer. Vi skal sammenligne SSB's utvalgsplan med den tilsvarende selvveiende ett-trinns stratifiserte utvalgsplan, dvs. hvor det tas et enkelt tilfeldig utvalg

fra hvert av de 109 strata. Med basis i (6.1), så blir utvalgsstørrelsene i strataene igjen lik $n_h = \frac{n}{N} N_h$. Også for ett-trinnsplanen blir estimatoren lik ekspansjonsestimatoren ; $\hat{t}_{st} = N\bar{y}_s$.

Anta vi ønsker å estimere andelen *p* i befolkningen med et gitt kjennetegn A, for eksempel andelen sysselsatte. Andelene er da estimert ved $\hat{t}_{2T} / N$ og $\hat{t}_{st} / N$, dvs. $\bar{y}_s = \hat{p}$ = andelen med kjennetegn A i utvalget. La :

m_h = antall PU i stratum *h,*

p_h = andel med kjennetegn A i stratum *h*,

$p_{h,i}$ = andel med kjennetegn A i PU *i* i stratum *h.*

Anta $n_i / N_{h,i}$ er små. Da er variansene for de to utvalgsplanene tilnærmet gitt ved :

$$Var_{SSB}(\hat{p}) = V_1 + V_2$$

hvor

$$V_1 = \frac{1}{N^2}\sum_{h=1}^{109} N_h v_h, \quad \text{hvor } v_h = \sum_{i=1}^{m_h} N_{h,i}(p_{h,i} - p_h)^2$$

$$V_2 = \frac{1}{n}\cdot\frac{1}{N}\sum_{h=1}^{109} w_h, \quad \text{hvor } w_h = \sum_{i=1}^{m_h} N_{h,i}\, p_{h,i}(1 - p_{h,i}).$$

$$Var_{STRAT}(\hat{p}) = \frac{1}{n}\cdot\frac{1}{N}\sum_{h=1}^{109} N_h p_h (1 - p_h).$$

Det kan vises at $V_2 \le Var_{STRAT}(\hat{p})$. Vanligvis vil imidlertid $Var_{SSB}(\hat{p}) > Var_{STRAT}(\hat{p})$. V_1 er bidraget fra 1.trinn-trekkingen. Den er mindre jo mindre variasjonen i andelene $p_{h,i}$ er innen strataene, dvs. når strata er homogene med hensyn på andel med kjennetegn A.

Vi kan sammenligne variansene ved å se på *designeffekten* $Var_{SSB}(\hat{p}) / Var_{STRAT}(\hat{p})$. Som en illustrasjon skal vi se på estimering av andel sysselsatte for forskjellige næringer i Østfold når den totale utvalgsstørrelsen for *hele* landet er 12 000 (ca. 870 for Østfold). I hvert tilfelle er da populasjonen den delen av befolkningen i Øsfold i den gitte næringen.

	Næringsgruppe	$Var_{SSB}(\hat{p}) / Var_{STRAT}(\hat{p})$
1	jordbruk, skogbruk, fiske og fangst	1,752
2	olje og bergverk	1,018
3	industri	1,380
4	kraft og forsyning	1,184
5	bygg og anlegg	1,333
6	varehandel, hotell, restaurant	1,683
7	transport, lager, post, televerket	1,186
8	finans, forretn., m.m.	1,066
9	offentlig og privat tjenesteyting	1,860
10	uoppgitt selvstendige	1,086
11	uoppgitt arbeidstakere	2,439

Tabellen illustrerer en av de to hovedegenskapene ved to-trinnstrekking :

Fordel: redusering av reiseutgifter

Ulempe : økt estimeringsvarians.

Det er viktig å danne homogene strata. Da vil den økte estimeringsvariansen, i forhold til et-trinns stratifisert utvalg, være liten og av mindre betydning enn kostnadsreduksjonen man oppnår med to-trinnstrekking.

Det kan vises at $Var_{STRAT}(\hat{p}) \leq \frac{p(1-p)}{n}$, som tilnærmet er variansen til $\hat{p}$ under enkelt tilfeldig utvalg. Statistisk sentralbyrå har brukt $1{,}5 \cdot p(1\text{-}p)/n$ som et tilnærmet uttrykk for $Var_{SSB}(\hat{p})$.

Hvis man er interessert i husholdningsvariable så kan man la utvalget bestå av de husholdningene som personene i *s* tilhører. I et selvveiende utvalg av personer så vil husholdninger av forskjellig størrelse ha ulike trekksannsynligheter. Sannsynligheten for å trekke en husholdning med *j* personer er tilnærmet lik $j \cdot (n/N)$. Horvitz-Thompson estimatoren kan da justeres med disse vektene.

Oppgaver

6.1 Stratum 4 i Oppland fylke består av følgende primære utvalgsområder, med innbyggertall:

PU	Antall innbyggere pr. 1.1.94	Antall innbyggere minst 16 år, pr. 1.1.99
Nord-Aurdal	6 601	5 306
Nordre Land/Etnedal	8 474	6 865
Sør-Aurdal	3 550	2 738
Vestre Slidre/Vang	4 296	3 165
Øystre Slidre	3 100	2 444
Totalt	*26 021*	*20 518*

(a) På trinn 1 skal et PU trekkes proporsjonalt med størrelse. Beregn trekksannsynlighetene for alle PU. (Ved trekkingen i SSB i desember 1994 ble Nordre Land/Etnedal valgt.)

(b) Anta vi skal trekke et landsutvalg på 10 000 personer (alder minst 16 år) med SSB's utvalgsplan. Det totale antall personer i Norge i denne befolkningen pr. 1.1.99 er 3 511 082, mens det totale innbyggertallet var 4 324 815 pr. 1.1.94. Bestem antallet personer som skal trekkes fra Nordre Land/ Etnedal ved (6.1), (6.2) og (6.3).

(c) Basert på utvalgsstørrelsene bestemt i punkt b), beregn trekksannsynligheten for personene i Nordre Land/Etnedal.

(d) Anta utvalgsstørrelsen innen hvert stratum er bestemt ved (6.1). Hva er trekksannsynligheten for en person fra Øystre Slidre ?

6.2 Anta at vi en tid etter at de primære utvalgsområdene ble trukket, får en betydelig flytting innen de enkelte strataene fra landlige primære utvalgsområder til de mer tettbygde.

(a) Hva vil en slik flytting bety for utvalgsundersøkelsene basert på SSB's plan, når (6.1) benyttes?

(b) Nevn noen praktiske problemer med å stadig gjenta trekningen av de primære utvalgsområdene.

7. Frafall og Imputering

Den kanskje viktigste feilkilden i utvalgsundersøkelser, i tillegg til utvalgsfeil, er *frafall*, dvs. manglende opplysninger fra enkelte enheter i utvalget. Man skiller mellom *enhetsfrafall* og *partielt frafall.* Enhetsfrafall betyr at en enhet i det planlagte utvalget uteblir fra det endelige utvalget, mens partielt frafall betyr at en enhet gir opplysninger om noen av variablene i undersøkelsen, men ikke alle. Det største problemet med frafall er at det kan medføre at utvalget ikke blir representativt for populasjonen. Det vil dermed oppstå skjevheter i estimeringen hvis vi bruker vanlige metoder. Et eksempel er inntektsundersøkelser hvor frafallet vanligvis er større i de høyeste inntektsgruppene.

I alle utvalgsundersøkelser av en viss størrelse må man regne med at vi får frafall. I de senere årene har det vært en tendens, i person- og husholdningsundersøkelser spesielt, at frafallsandelen har økt. Det ser ut til å være to hovedårsaker: (i) ikke å treffe: befolkningen tilbringer mindre tid hjemme og er derfor vanskeligere å få tak i, og (ii) nekter: det ser ut til å være en økende skepsis mot slike undersøkelser. Når det gjelder (i) så vil gjenbesøk øke svarprosenten vesentlig.

Det er hovedsakelig tre typer av utvalgsundersøkelser, telefonbaserte, postale og besøksundersøkelser. Generelt er det sjelden at frafallet er mindre enn 10%. Spesielt i postale undersøkelser kan frafallet bli stort ; det er ikke uvanlig med frafallsprosenter på 70-80. Det er derfor viktig å trekke inn mulige skjevheter på grunn av frafall for å kunne foreta en realistisk statistisk analyse. Vi har i kap.4 sett et lite eksempel på hvordan etterstratifisering kan benyttes for å rette opp frafallsskjevheter.

7.1 Effekt av frafall

For å illustrere effekten av frafall ser vi på personundersøkelser og tenker vi oss at populasjonen er inndelt i to strata. Det ene stratum består av de personene som ville svart dersom de ble trukket ut til utvalget, og det andre stratum består av de personene som ikke ville svart hvis de ble trukket ut. Dette er selvsagt en idealisert situasjon, men den kan tjene til å belyse frafallsproblemet. La N_R være størrelsen på "responsstratumet" U_R , og tilsvarende er N_F størrelsen på "frafallsstratumet" U_F. Populasjonsgjennomsnittene i de respektive strata betegnes med $\overline{Y}_R$ og $\overline{Y}_F$, slik at det totale populasjonsgjennomsnittet $\overline{Y}(=t/N)$ er gitt ved

$$\overline{Y} = \frac{N_R\overline{Y}_R + N_F\overline{Y}_F}{N} = q_R\overline{Y}_R + (1-q_R)\overline{Y}_F,$$

hvor $q_R = N_R / N$ er forventet responsandel.

Anta vi trekker et enkelt tilfeldig utvalg på *n* personer for å estimere $\overline{Y}$, og at vi får svar fra n_r personer, med observert gjennomsnitt lik $\bar{y}_r$. Data er da kun fra responsstratumet, men $\bar{y}_r$ er ment å estimere $\overline{Y}$. La s_r være svarutvalget. Gitt n_r så vil s_r være et enkelt tilfeldig utvalg fra U_R. Det betyr at $E(\bar{y}_r) = \overline{Y}_R$, og *estimeringsskjevheten* blir dermed

$$E(\bar{y}_r) - \overline{Y} = \overline{Y}_R - \overline{Y} = \overline{Y}_R - q_R\overline{Y}_R - (1-q_R)\overline{Y}_F = (1-q_R)(\overline{Y}_R - \overline{Y}_F). \qquad (7.1)$$

Uten kjennskap til $\overline{Y}_F$ kan ikke $\bar{y}_r$ korrigeres for denne ukjente skjevheten. Det er to måter å angripe dette problemet på. Enten ved noen *modell*-antagelser om U_F (f..eks. at populasjonsgjennomsnittene $\overline{Y}_R$ og $\overline{Y}_F$ er like), eller en modell for sannsynligheten for at en person svarer som funksjon av *y* og tilgjengelig registerinformasjon (responsmodellering) . Den siste tilnærmingen er den mest vanlige, og vi skal se et eksempel senere (eksempel 7.2).

Tilsvarende utledningen av (7.1), gitt observert n_r, $Var(\bar{y}_r) = \frac{\sigma_R^2}{n_r} \cdot \frac{N_R - n_r}{N_R - 1} \approx \frac{\sigma_R^2}{n_r}$ når $n_r << N_R$.

Her er σ_R^2 populasjonsvariansen i responsstratum, dvs.

$$\sigma_R^2 = \frac{1}{N_R} \sum_{i \in U_R} (y_i - \bar{Y}_R)^2 .$$

Det betyr at , ubetinget over alle mulige verdier av n_r ,

$$Var(\bar{y}_r) \approx \sigma_R^2 E(\frac{1}{n_r}) \approx \frac{\sigma_R^2}{q_R n} ,$$

siden $E(n_r) = q_R n$. For estimatorer som ikke er forventningsrette er det vanlig å benytte bruttovariansen som et totalt mål på estimeringsusikkerhet,

$$\begin{aligned} E(\bar{y}_r - \bar{Y})^2 &= [E(\bar{y}_r) - \bar{Y}]^2 + Var(\bar{y}_r) \\ &\approx (1 - q_R)^2 (\bar{Y}_R - \bar{Y}_F)^2 + \frac{\sigma_R^2}{q_R n}. \end{aligned}$$

Anta *y* er en binær variabel, dvs. *y* = 1 hvis enheten har et gitt kjennetegn A, og 0 ellers. Da er $\bar{Y}_R, \bar{Y}_F$ andelene med kjennetegn A, og $(\bar{Y}_R - \bar{Y}_F)^2 \leq 1$, slik at bruttovariansen er høyst lik $(1 - q_R)^2 + \frac{\sigma_R^2}{q_R n}$.
Et forventet frafall på 30% vil da maksimalt gi et tillegg på bruttovariansen på 9%.

Legg merke til at skjevheten er uavhengig av *n*, dvs. skjevheten kan ikke reduseres ved å øke *n.* I tillegg har vi følgende mulige konsekvenser av frafall:

- Skjevheten øker med økende forventet frafallsandel (1-q_R).
- Skjevheten øker når forskjellen mellom respons- og frafallsgruppen, $\bar{Y}_R - \bar{Y}_F$, øker.
- Hvis det er ingen forskjell mellom respons- og frafallsgruppen, dvs. $\bar{Y}_R - \bar{Y}_F = 0$, så er effekten av frafall kun at forventet utvalgsstørrelse reduseres fra *n* til $q_R n$. I dette tilfellet sier vi at frafallsmekanismen er *ignorerbar.* Hvis man ønsker en viss størrelse på det endelige svarutvalget, f.eks. ca. 1000 og vet omtrent hva q_R er, så tar man et utvalg med n = 1000/ q_R. For eksempel, hvis forventet responsandel er 60% så blir n = 1000/0,6 = 1667.
- Det er vanligvis en urealistisk antagelse at $\bar{Y}_R = \bar{Y}_F$, men innen mindre delpopulasjoner er det ofte ikke urimelig å anta at svargruppen er tilnærmet representativ for frafallsgruppen, spesielt hvis den variabelen som brukes til å dele opp populasjonen er høyt korrelert med *y.* Dette gir opphav til etterstratifisering som en metode for å rette opp frafallsskjevheten. Man prøver da å bruke registervariable nær knyttet til *y* som etterstratifiseringsvariable.

7.2 Estimeringsmetoder for å redusere effekten av frafall

Behandling av frafallsproblemer skjer både ved å redusere *størrelsen* på frafallet, spesielt ved gjenbesøk, samt å redusere *effekten* av frafall ved å estimere skjevheten på grunn av frafall og korrigere de opprinnelige estimatorene som er beregnet på et fullstendig utvalg. Vi skal nå se på tre estimeringsmetoder som tar hensyn til frafall. Den første er etterstratifisering.

Etterstratifisering for frafall

Når vi skal etterstratifisere for å rette opp frafallsskjevhet er det ønskelig å stratifisere etter variable som deler opp populasjonen i homogene grupper med hensyn til *y*- variabelen, hvor samtidig svarandelene varierer mye. La H betegne antall etterstrata, og la N_h være antall enheter i etterstratum h. Vi tenker oss at hvert etterstratum er inndelt i en responsgruppe og en frafallsgruppe, og definerer:

q_h = svar-andelen i etterstratum h, $h = 1, \ldots, H$

W_h = andel som etterstratum h utgjør av hele populasjonen, $W_h = N_h/N$

$\overline{Y}_{rh}$ = gjennomsnitt i responsgruppen i etterstratum h

$\overline{Y}_{fh}$ = gjennomsnitt i frafallsgruppen i etterstratum h.

Anta vi har et enkelt tilfeldig utvalg,og la igjen $\overline{y}_r$ estimere $\overline{Y}$. Skjevheten er gitt ved, fra (7.1), $E(\overline{y}_r) - \overline{Y} = (1 - q_R)(\overline{Y}_R - \overline{Y}_F)$. Ved å benytte at

$$\overline{Y}_R = \frac{1}{q_R}(q_1 W_1 \overline{Y}_{r1} + q_2 W_2 \overline{Y}_{r2} + \ldots + q_H W_H \overline{Y}_{rH}) = \frac{1}{q_R}\sum_h q_h W_h \overline{Y}_{rh}$$

og $$\overline{Y}_F = \frac{1}{1-q_R}[(1-q_1)W_1\overline{Y}_{f1} + (1-q_2)W_2\overline{Y}_{f2} + \ldots + (1-q_H)W_H\overline{Y}_{fH}] = \frac{1}{1-q_R}\sum_h (1-q_h)W_h\overline{Y}_{fh}\ ,$$

kan det vises (oppgave 7.1) at vi kan uttrykke skjevheten på følgende form

$$E(\overline{y}_r) - \overline{Y} = \frac{1}{q_R}\sum_h \overline{Y}_{rh} W_h (q_h - q_R) + \sum_h (1-q_h) W_h (\overline{Y}_{rh} - \overline{Y}_{fh})\,. \qquad (7.2)$$

Den første komponenten er skjevheten på grunn av forskjellige forventede responsandeler i strataene, og kan estimeres, mens den andre komponenten ikke kan estimeres hvis respons- og frafallsgjennomsnittene er forskjellige (og ingen responsmodell er antatt) . Etterstratifisering kan bare estimere den første komponenten. Det gjelder derfor å velge etterstrata slik at mest mulig av skjevheten er i første komponent, dvs. slik at responsandelene varierer maksimalt samtidig med at $\overline{Y}_{rh} = \overline{Y}_{fh}$, iallfall tilnærmet, i alle strata. Da vil etterstratifisering gi tilnærmet forventningsrett estimator. Generelt vil den andre komponenten være den gjenværende del av skjevheten ved bruk av etterstratifisert estimator. Legg merke til at den totale skjevheten er *uavhengig* av valg av etterstrata, slik at summen av de to komponentene i (7.2) er den samme for alle mulige etterstratifiseringer. Den første komponenten er lik $\overline{Y}_R - \sum_h W_h \overline{Y}_{rh}$. La $\overline{y}_h = \overline{y}_{s_h}$ være observert gjennomsnitt fra etterstratum h. En forventningsrett estimator for den første komponenten er dermed:

$$\overline{y}_r - \sum_h W_h \overline{y}_h$$

og den korrigerte estimatoren blir dermed

$$\hat{\overline{y}}_{est} = \overline{y}_r - (\overline{y}_r - \sum_h W_h \overline{y}_h) = \sum_h W_h \overline{y}_h = \frac{1}{N}\sum_h N_h \overline{y}_h\ .$$

Dvs., for totalen $t = N\overline{Y}$, så blir etterstratifiseringsestimatoren

$$\hat{t}_{est} = \sum_h N_h \overline{y}_h$$

som vi kjenner igjen fra kapittel 4.2.

Eksempel 7.1. Estimering av antall husholdninger i Norge. Vi skal bruke data fra Statistisk sentralbyrås forbruksundersøkelse i 1992. Det er et selvveiende landsutvalg på ialt n = 1698 personer over 15 år, etter SSB's generelle utvalgsplan fra kap. 6. Problemet er å estimere antall husholdninger av forskjellige størrelser : 1, 2, 3, 4, ≥ 5, og det totale antall husholdninger. For en gitt husholdningstørrelse j så er variabelen av interesse for person i, y_i = 1 hvis husholdningstørrelsen er lik j og 0 ellers. Da er de totale antall personer i husholdninger av størrelse j lik totalen

$$t^{(j)} = \text{ summen av alle } y_i \text{ i populasjonen.}$$

Norge har et register over familier. Fra dette registeret vet vi *familie*størrelsen for alle personer som vi skal bruke til etterstratifisering. Resultatet av undersøkelsen ble :

Familie-størrelse	Husholdningsstørrelse					Total	Frafall	Frafalls-prosent
	1	*2*	*3*	*4*	*≥5*			
1	83	48	20	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, tallet 48 i celle (*1*,*2*) betyr at av 162 personer i svarutvalget som i følge registeret var aleneboende, tilhørte faktisk 48 personer en to-person husholdning. De fleste av disse var antagelig samboere uten å være gift.

Vi har 5 etterstrata etter familiestørrelse, og frafallsandelene er sterkt synkende når familiene blir større. Det vil være fordelaktig å etterstratifisere etter familiestørrelse, også fordi familie- og husholdningstørrelse er høyt korrelerte. La oss, imidlertid, først betrakte ekspansjonsestimatoren for antall husholdninger H_j av størrelse j. La N være antall personer i befolkningen, n_j antall personer i svarutvalget med husholdningsstørrelse j, og n_r antall personer i svarutvalget. Andel personer i utvalget med husholdningsstørrelse j er $n_j/\ n_r$ (= $\bar{y}_r$, gjennomsnittet av indikatorvariabelen i svarutvalget) slik at *antall personer* i befolkningen med husholdningsstørrelse j da blir estimert til $\hat{t}_e^{(j)} = N\frac{n_j}{n_r}$. Dermed blir estimatoren for H_j, antall *husholdninger* med størrelse j, lik

$$\hat{H}_e^{(j)} = \frac{1}{j} N \frac{n_j}{n_r}.$$

Pr. 1.1.93, N = 4.131.874 så f.eks. $\hat{H}_e^{(1)} = 4131874\frac{105}{1111} = 390.501$. Dette tallet underestimerer antall en-person husholdninger kraftig, siden frafallsprosenten blant en-person familier, og dermed høyst sannsynlig blant en-person husholdninger er mye større enn for de andre familiestørrelsene. La oss se på etterstratifiseringsestimatoren for antall en-person husholdninger spesielt og H_j generelt. La N_k være antall personer i Norge med familiestørrelse k, k = 1, 2, 3, 4, ≥ 5. Vi har N_1 = 793.869, N_2 = 816.880, N_3 = 784.581, N_4 = 1.066.016, $N_{\geq 5}$ = 670.528. Med n_{kj} lik antall personer i svarutvalget med familiestørrelse k og husholdningsstørrelse j, og m_k lik antall personer i svarutvalget med familiestør-

relse k så blir observert andel personer med husholdningstørrelse j i etterstratum k lik $\bar{y}_k = n_{kj} / m_k$ slik at etterstratifiseringsestimatoren for H_j er gitt ved

$$\hat{H}_{est}^{(j)} = \frac{1}{j} \sum_k N_k \frac{n_{kj}}{m_k}.$$

Antall en-person husholdninger estimeres da til :

$$\hat{H}_{est}^{(1)} = 793869 \frac{83}{162} + 816880 \frac{9}{230} + 784581 \frac{10}{212} + 1066016 \frac{2}{300} + 670528 \frac{1}{207} = 486\,055.$$

Dette er en økning i estimatet på ca. 25%. Fremdeles får vi, imidlertid, underestimering . Det viser seg, ved en nærmere analyse basert på responsmodellering, at frafall har mer sammenheng med faktisk husholdningsstørrelse enn familiestørrelse. Basert på en responsmodell som antar at sannsynligheten for respons avhenger av husholdningsstørrelse og bosted (by eller land) blir estimatet av antall en-person husholdninger lik 595.400 som, høyst sannsynlig, er ganske nær det faktiske antallet. For eksempel, en kvalitetsundersøkelse av folketellingen fra 1990 ga et estimat på 626.000.

Følgende tabell viser de andre estimatene ved ekspansjonestimatoren og etterstratifiseringsestimatoren, til nærmeste 1000. Vi har også inkludert estimatene basert på responsmodellen.

Tabell 7.1

Husholdningstørrelse	Ekspansjonsestimator	Etterstratifisert estimator	Modellbasert estimator
1	391.000	486.000	595.000
2	496.000	508.000	526.000
3	284.000	286.000	249.000
4	280.000	271.000	269.000
≥ 5	148.000	131.000	126.000
Totalt	1. 599.000	1.682.000	1.765.000

Justeringsceller

Hvis variablene som brukes til etterstratifisering er slik at stratastørrelsene N_h er ukjente så kan man bruke en lignende estimator såfremt størrelsene n_h på etterstrata i det planlagte utvalget s er kjente. Man kan da bruke $N \cdot (n_h/n)$ som anslag på N_h, og får estimatoren

$$\hat{t}_{vc} = N \sum_h \frac{n_h}{n} \bar{y}_h \ .$$

Dette kalles den veiede celle-estimatoren. La $\hat{q}_h$ være observert svarandel i stratum h. Vektene til y-verdiene i utvalget fra stratum h er lik $1/(n\hat{q}_h)$, slik at sammenlignet med $\hat{t}_e = N\bar{y}_r$, justerer $\hat{t}_{vc}$ vektene etter hvilket stratum observasjonene kommer fra (fra $1/n_r$ til $(1/(n\hat{q}_h))$, og strataene kalles derfor i denne sammenhengen for justeringsceller. Veiing av observasjonene etter justeringsceller er en av de mest vanlige metoder for å korrigere for frafall.

En metode basert på ett gjenbesøk

En direkte, enkel bruk av ett gjenbesøk er å betrakte data fra gjenbesøket som representativt for frafallsgruppen. Anta vi har enkelt tilfeldig utvalg . La $\bar{y}_1$ være observert gjennomsnitt ved første besøk, og $\bar{y}_2$ gjennomsnittet fra de enhetene som svarer på det andre besøket. Vi kan da betrakte $\bar{y}_2$ som et estimat for gjennomsnittet i hele frafallsgruppen i utvalget. La n_1, n_2 være antall observasjoner på første og andre besøk. Da vil $(n - n_1 - n_2)\bar{y}_2$ estimere totalen i utvalgets frafallsgruppe og estimatoren for $\bar{Y}$ blir dermed, med $p = n_1 / n$ (svarandelen på første besøk),

$$\bar{y}^* = \frac{n_1\bar{y}_1 + n_2\bar{y}_2 + (n - n_1 - n_2)\bar{y}_2}{n} = p\bar{y}_1 + (1-p)\bar{y}_2 \ . \qquad (7.3)$$

Vanligvis vil ikke gjenbesøket være representativt for frafall, men denne estimatoren vil antagelig være noe bedre enn det totale gjennomsnittet fra de to besøkene siden enhetene fra gjenbesøket forventes å være mer lik frafallsgruppen enn enhetene fra første besøk.

7.3 Imputering

I tillegg til skjevhetsproblemer, skaper frafall problemer med standard statistiske analyser, selv om frafallet skjer tilfeldig. En generell metode for å rette opp eventuelle skjevheter, og samtidig en hjelp til å benytte standard statistiske metoder, er å fylle inn for manglende data ved å predikere de savnede verdiene, både for partielt frafall og enhetsfrafall. Dette kalles *imputering* og er mye brukt i Statistisk sentralbyrås undersøkelser.

Illustrasjon. Anta vi har en undersøkelse med fire variable y_1, y_2, y_3, y_4 og registervariable **x**. Populasjonsstørrelsen er N =10 , og planlagt utvalgsstørrelse er lik n = 5, med resultat av trekkingen lik s = {2,3,6,8,9}. Det viste seg at enhet 9 ikke ble med i utvalget. Dessuten ble det partielt frafall for enhetene 3 og 6 med hensyn til y_3 og (y_2, y_3) henholdsvis. Vi kan representere data som en ufullstendig matrise , hvor **r** indikerer respons/frafall for de fire variablene.

Data

	x		**r**			observerte *y*-variable			
	$\mathbf{x}_2$	1	1	1	1	y_{21}	y_{22}	y_{23}	y_{24}
	$\mathbf{x}_3$	1	1	0	1	y_{31}	y_{32}	-	y_{34}
s	$\mathbf{x}_6$	1	0	0	1	y_{61}	-	-	y_{64}
	$\mathbf{x}_8$	1	1	1	1	y_{81}	y_{82}	y_{83}	y_{84}
	$\mathbf{x}_9$	0	0	0	0	-	-	-	-
	$\mathbf{x}_1$								
	$\mathbf{x}_4$								
utenfor s	$\mathbf{x}_5$								
	$\mathbf{x}_7$								
	$\mathbf{x}_{10}$								

Imputering betyr nå å fylle inn de tomme cellene i datamatrisen for (y_1, y_2, y_3, y_4) i hele utvalget s.

La oss si litt om bruk av ordet "imputere". Det står ikke nevnt i norske ordbøker, men kan vel betraktes som det motsatte av "amputere". På engelsk, fra Webster dictionary leser vi - impute: ascribe (legge til, tilskrive), attribute. I Kunnskapsforlagets engelsk-norsk ordbok - imputation : beskyldning, og impute: tilskrive en noe, legge en noe til last, om noe klanderverdig. For vår bruk av ordet er vel "tilskrive" nærmest. Det som faktisk gjøres er å estimere eller egentlig predikere de ukjente verdiene i frafallsgruppen. Statistisk sett er derfor imputering en spesiell form for prediksjon.

Vi skal nå betrakte imputering for èn variabel *y.* Det kan være forskjellige formål med imputering:

1. Estimater som justerer for mulig frafallsskjevhet
2. Estimert varians av estimator som reflekterer redusert utvalgsstørrelse og frafallseffekt
3. Produsere komplette data som tillater standard statistisk analyse (eksempelvis, data for offentlig bruk)
4. Vise analysens følsomhet overfor forskjellige modeller for frafall.

Når det gjelder punktene 2 og 3 er det viktig at imputeringsverdiene gjenspeiler *variasjonen* man kan forvente i frafallsgruppen, i tillegg til en eventuell skjevhet på grunn av frafall. La oss først, imidlertid, se på punkt 1. Dette problemet har vi angrepet tidligere ved etterstratifisering eller justeringsceller, og en estimator basert på gjenbesøk.

Imputeringsbasert estimering

Problemet er å estimere populasjonstotalen *t.* Anta $\hat{t}$ er den valgte estimator basert på hele utvalget *s*, f.eks. $\hat{t} = N\bar{y}_s$. La $\hat{t}_r$ være $\hat{t}$ basert på svarutvalget s_r, for *y.* Som vi har sett i kapitlene 7.1 og 7.2 så vil $\hat{t}_r$ være skjev hvis frafallet har sammenheng med underliggende *y*-verdi. Med imputeringsverdier fyllt inn i frafallsgruppen får vi et *konstruert* fullstendig utvalg. *Imputeringsestimatoren* $\hat{t}_I$ er da $\hat{t}$ basert på det konstruerte fullstendige utvalget. Den mest fruktbare måten å imputere på er ved bruk av responsmodellering, og vanligvis *populasjonsmodellering.* Responsmodellen kan være spekulativ og man bør foreta gjenbesøk for å sjekke om modellen er realistisk. Et eksempel er å anta at verdiene både i svar- og i frafallsgruppen følger en regresjonsmodell mot visse variable som *er* observert i frafallsgruppen. Imputerte verdier kan da være predikerte verdier fra den tilpassede regresjonsmodellen. Modellering av populasjonen har vi ikke vært inne på før, men er et vanlig utgangspunkt i moderne utvalgsteori. Det betyr at vi betrakter hele populasjonsvektoren $\mathbf{y} = (y_1,\ldots,y_N)$ som verdier av stokastiske variable, og gjør analysen gitt det utvalget vi faktisk får. La imputeringsverdiene i utvalgets frafallsgruppe betegnes med y_i^*. To vanlige typer av imputeringsteknikker under en slik modelleringsramme er :

(i) y_i^* = estimert forventet verdi for *y gitt* frafall , f.eks. predikerte verdier fra regresjonsmodell.

(ii) y_i^* trekkes tilfeldig fra estimert fordeling for *y gitt* frafall.

Vi skal nå, med et enkelt eksempel uten bruk av populasjonsmodell, illustrere hvordan responsmodellering og imputering kan brukes til å rette opp frafallsskjevhet.

Eksempel 7.2. Variabelen av interesse er binær, y_i = 1 hvis person er sysselsatt og 0 ellers. Responsmodellen antar at responsandelen blant de sysselsatte er dobbelt så stor som blant ikke-sysselsatte. Ved å innføre responsvariabelen R_i for *i*'te enhet; R_i = 1 hvis svar og 0 hvis frafall, så kan denne modellen formuleres som en betinget sannsynlighet for respons-variabelen gitt sysselsetting-status:

$$P(R_i = 1 | y_i) = \begin{cases} \psi \text{ hvis } y_i = 0 \\ 2\psi \text{ hvis } y_i = 1 \end{cases}$$

Her er ψ en ukjent modellparameter. La oss si at det planlagte utvalget har n = 80, og at vi får svar fra 40 personer hvorav 30 personer viste seg å være sysselsatte. Hvordan skal vi da estimere andel sysselsatte i befolkningen? I svarutvalget vil det være en overrepresentasjon av sysselsatte, siden en større del av disse vil svare enn blant de ikke-sysselsatte. Den observerte andelen, $\bar{y}_r$ = 30/40 = 0,75, overestimerer dermed populasjonsandelen. Men spørsmålet er nå hvor mye ned estimatet 0,75 bør senkes. Dette løser vi ved å imputere et antall sysselsatte, z, i frafallsgruppen, basert på responsmodellen. Dvs., z av frafallsenhetene får imputeringsverdien y_i^* = 1, mens resten får imputeringsverdi 0. På bakgrunn av responsmodellen vil vi forvente at

$$\frac{30}{30+z} = 2\frac{10}{10+(40-z)}$$

Dvs. vi bestemmer z slik at responsandelen blant de sysselsatte i utvalget er det dobbelte av responsandelen blant ikke-sysselsatte. Dette gir :

$$30(50\text{-}z) = 20(30+z) \text{ dvs. } 1500 - 30z = 600 + 20z, \text{ og dermed}$$

$$50z = 900 \text{ dvs. } z = 900/50 = 18.$$

Vi ser nå at responsandelen blant de sysselsatte er estimert til 30/48 = 0,625, mens responsandelen blant ikke-sysselsatte blir estimert lik 10/32 = 0,3125, dvs., responsparameteren er estimert til $\hat{\psi} = 0{,}3125$.

For estimering av andel sysselsatte i hele populasjonen så bruker vi nå andelen i det *kompletterte* utvalget som har imputert 18 sysselsatte i frafallsgruppen :

$$\bar{y}_{s,I} = 48/80 = 0{,}60;$$

vi har redusert andelestimatet fra 75% til 60%.

La f = 1- n_r/n være observert frafallsandel Det kan vises at generelt vil antall sysselsatte imputert i frafallsgruppen være lik:

$$(n-n_r)\bar{y}_r \frac{n-n_r-n_r(1-\bar{y}_r)}{n-n_r+(n-n_r)(1-\bar{y}_r)} = n_r\bar{y}_r \frac{\frac{f}{1-f}-(1-\bar{y}_r)}{1+(1-\bar{y}_r)}$$

med $\bar{y}_{s,I} = \bar{y}_r / (2-\bar{y}_r)$ $(< \bar{y}_r)$ og $\hat{\psi} = \frac{1}{2}\cdot\frac{n_r}{n}(2-\bar{y}_r)$.

Imputering for standard statistisk analyse

Vi betrakter, for enkelthets skyld, det tilfellet at frafallet er rent tilfeldig. Hvis vi ikke har noen kjent registerinformasjon knyttet til alle enhetene , så vil imputeringsteknikk (i) bety at $y_i^* = \bar{y}_r$, såkalt *middel* imputering og (ii) at y_i^* trekkes tilfeldig blant de observerte y-verdiene, med tilbakelegging. Den siste metoden kalles *hot-deck* imputering, og er, i forskjellige versjoner, mye brukt i Statistisk sentralbyrå og i de fleste lands offisielle statistiske byråer. Med registerinformasjon tilgjengelig er det vanlig å bruke hot-deck imputering innenfor etterstrata. Vi skal nå illustrere disse to metodene ved enkelt tilfeldig utvalg. Vi vil se at den første metoden, ikke uventet, er ubrukelig hvis formålet også er å gi komplette datasett som skal gjenspeile forventet variasjon i frafallsgruppen.

Betrakt estimering av $\bar{Y} = t/N$. La $\bar{y}_s$ være, som før, gjennomsnittet hvis hele utvalget *s* oberveres, og la $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i \in s}(y_i - \bar{y}_s)^2$. For store *n*, *N-n* er standard 95% konfidensintervall :

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

La $\bar{y}_s^*, \hat{\sigma}_*^2$ være lik $\bar{y}_s, \hat{\sigma}^2$ basert på det konstruerte fullstendige utvalget med observerte og imputerte verdier. Med frafall blir dermed standard konfidensintervall basert på det komplette utvalget:

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

For middel imputering og hot-deck imputering har vi følgende konfidensnivåer for KI*:

Frafallsandel	Middel imputering	Hot-deck imputering
0	0,95	0,95
0,1	0,922	0,925
0,2	0,883	0,896
0,3	0,830	0,864
0,4	0,760	0,826
0,5	0,673	0,785

Som ventet fungerer ikke middel imputering. Med denne imputeringsteknikken så blir $\bar{y}_s^* = \bar{y}_r$, det observerte gjennomsnittet, som er forventningsrett, dvs., ingen frafallsskjevhet i estimatoren. Imidlertid, $\hat{\sigma}_*^2 = \frac{n_r - 1}{n-1}\hat{\sigma}_r^2$, hvor $\hat{\sigma}_r^2$ er $\hat{\sigma}^2$ basert på svarutvalget. Siden $\hat{\sigma}_r^2$ er forventningsrett estimator for populasjonsvariansen σ^2, ser vi at $\hat{\sigma}_*^2$ underestimerer σ^2. I tillegg,

$$Var(\bar{y}_s^*) = \sigma^2(\frac{1}{n_r} - \frac{1}{N}) > \sigma^2(\frac{1}{n} - \frac{1}{N}),$$

dvs, vi har "feil" formel for standardfeilen i konfidensintervallet . Det er altså to årsaker til at konfidensintervallet basert på det kompletterte utvalget blir for kort . Den ene er at $\hat{\sigma}_*$ underestimerer σ med faktoren $\sqrt{n_r / n}$, og den andre årsaken er at KI* "later som" om utvalgsstørrelsen er *n* når den egentlig er n_r.

Poenget med hot-deck imputeringen er å prøve å unngå underestimeringen av variansen, dvs. av variasjonen i frafallsgruppen. Selv om denne imputeringsmetoden gir bedre konfidensnivåer så er den ikke tilfredsstillende, også her blir KI* for kort. Det kan vises at gitt de observerte data i s_r , så vil $E(\bar{y}_s^*) = \bar{y}_r$, og $E(\hat{\sigma}_*^2) \approx \hat{\sigma}_r^2$, slik at estimeringen er forventningsrett og variansestimatet er o.k. Problemet er, imidlertid, at også her vil $Var(\bar{y}_s^*) > \sigma^2(\frac{1}{n} - \frac{1}{N})$.

Et korrekt 95% konfidensintervall basert på $(\bar{y}_s^*, \hat{\sigma}_*)$ med hot-deck imputering kan utledes og er gitt ved

$$\bar{y}_s^* \pm 1{,}96\hat{\sigma}_* \sqrt{\tfrac{1}{n_r} - \tfrac{1}{N} + \tfrac{n-n_r}{n^2}(1 - \tfrac{1}{n_r})} \,,$$

men poenget er at man ikke kan foreta "vanlig" statistisk analyse på data-materiale som inneholder imputerte verdier, selv om de imputerte verdiene har samme varians som data. Det som mangler er et mål på usikkerheten i selve de imputerte verdier som estimater for de sanne verdiene i frafallsgruppen. Med *multippel imputering* er det mulig å inkludere et slikt mål for imputeringsusikkerheten, og samtidig *kombinere* standard analyser.

Multippel imputering betyr at for hver enhet i frafallsgruppen imputeres *flere* verdier. La *m* være antall imputerte verdier for hver enhet. Vi får da *m* konstruerte fullstendige utvalg. Anta vi foretar *m* hot-deck imputeringer, dvs. vi trekker *m* verdier for hver enhet i frafallsgruppen, fra de observerte data med tilbakelegging.

Vi har nå *m* standard analyser med gjennomsnitt og variansestimatene $\bar{y}_s^*(i), \hat{\sigma}_*^2(i)$, for $i = 1, \ldots, m$, som vi kan kombinere. La $\bar{\bar{y}}_s^*, \bar{\sigma}_*^2$ være gjennomsnittene. Et "direkte" standard konfidensintervall er

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

Det viser seg at det fremdeles er for kort, også når $m > 1$. Uttrykket $\bar{\sigma}_* \sqrt{\tfrac{1}{n} - \tfrac{1}{N}}$ måler variasjonen bare innen data-settene. I tillegg er det, som nevnt foran, nødvendig å ta med et mål på variasjonen mellom data-settene for å få med usikkerheten på grunn av imputeringen. Det er gitt ved :

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

Et forslag til kombinasjon av de *m* komplette data-settene for estimering av $Var(\bar{\bar{y}}_s^*)$, av D. Rubin :

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

med tilhørende 95% konfidensintervall :

$$\bar{\bar{y}}_s^* \pm 1{,}96\sqrt{V_*} \,.$$

Også dette intervallet er litt for kort som følgende tabell over konfidensnivåer viser, men det er en sterk forbedring fra vanlig hot-deck basert intervall ($m = 1$) til $m = 2$.

Frafallsandel	$m = 1$	$m = 2$	$m = \infty$
0	0,95	0,95	0,95
0,1	0,925	0,949	0,949
0,2	0,896	0,946	0,945
0,3	0,864	0,940	0,939
0,4	0,826	0,930	0,927
0,5	0,785	0,916	0,910

Rubin's kombinasjon krever egentlig en spesiell modellbasert imputering for å kunne gi korrekt konfidensnivå, men vi ser den fungerer bra også med hot-deck imputering for moderate frafallsandeler. Det er mulig å justere selve kombinasjonsestimatet for $Var(\bar{\bar{y}}_s^*)$ slik at konfidensnivået blir korrekt. Det oppnås ved å benytte, istedenfor V_*,

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

hvor f er frafallsandelen i utvalget. (Vist av Tonje Braaten i sin cand.scient. oppgave ved Universitetet i Tromsø, 1999).

Oppgaver

7.1* Betrakt etterstratifisering for frafall i kap. 7.2.

(a) Vis at $\bar{Y}_R = \frac{1}{q_R} \sum_h q_h W_h \bar{Y}_{rh}$ og $\bar{Y}_F = \frac{1}{1-q_R} \sum_h (1-q_h) W_h \bar{Y}_{fh}$.

(b) Vis skjevhetssuttrykket (7.2).

7.2 Kontrollberegn estimatene for H_2, H_3, H_4, $H_{\geq 5}$ i tabell 7.1, basert på ekspansjonsestimatoren og etterstratifisert estimator. For $H_{\geq 5}$ deler vi med j = 5,25 som et estimat på gjennomsnittlig husholdningsstørrelse for husholdninger på 5 eller flere personer. (Antall familier med 5 eller flere personer er 127.653 og har ialt 670.528 personer som gir en gjennomsnittlig familiestørrelse på 5,25.)

7.3 Anta vi tar ett gjenbesøk. Vi skal betrakte $\bar{y}_r$ og $\bar{y}^*$, gitt ved (7.3) i kap. 7.2, som imputeringsbaserte estimatorer for $\bar{Y}$. La $\bar{y}_s$ være basisestimatoren.

(a) Hvilke imputeringsverdier gir at imputeringsestimatoren $\bar{y}_s^* = \bar{y}_r$?

(b) Hvilke imputeringsverdier gir at imputeringsestimatoren $\bar{y}_s^* = \bar{y}^* = p\bar{y}_1 + (1-p)\bar{y}_2$?

7.4* Ved hot-deck imputering trekkes de imputerte verdiene y_i^* tilfeldig fra de n_r observerte y-verdiene, med tilbakelegging. Dvs. at hver observert y har sannsynlighet ($1/n_r$) for å trekkes ut som imputert verdi. Vis at, *gitt* data i svarutvalget s_r , så er

$$E(y_i^*) = \frac{1}{n_r} \sum_{k \in s_r} y_k = \bar{y}_r \quad \text{og} \quad Var(y_i^*) = \frac{1}{n_r} \sum_{k \in s_r} (y_k - \bar{y}_r)^2 .$$

7.5 Vi skal se på Valgundersøkelsen utført av Statistisk sentralbyrå i 1993. Det ble, bl.a., stilt spørsmål om de intervjuede personene stemte ved Stortingsvalgene i 1993 og 1989. I 1993 var det 3.259.957 stemmeberettigede i Norge. Det ble tatt et utvalg på 3000 personer etter SSB's utvalgsplan, og foretatt ialt 11 gjenbesøk. Vi skal bruke data etter to gjenbesøk. Da hadde ialt 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) skal vi etterstratifisere etter valgdeltakelse i 1989. Vi deler inn i tre etterstrata:

Stratum 1: De som deltok i valget i 1989, antall $N_1 = 2.510.669$

Stratum 2: De som ikke deltok i valget 1989, antall $N_2 = 508.288$

Stratum 3 : Nye velgere i 1993, antall $N_3 = 241.000$.

De 1403 personene i svarutvalget fordelte seg slik på disse etterstrata: Variabelen y indikerer om de stemte i 1993 ($y = 1$) eller ikke ($y = 0$).

Stratum	1		2		3	
y	0	1	0	1	0	1
antall	132	1060	58	57	23	73
totalt	1192		115		96	

(c) Finn etterstratifiseringsestimatet for andelen som stemte i 1993, og sammenlign med estimatet i punkt (a) og den sanne verdien 0,759. Kommentér.

(d) Etterstratifisering kan betraktes som en imputeringsbasert estimator. Hva er i dette tilfellet basisestimatoren (den valgte estimatoren basert på hele utvalget), og hvordan er imputeringen foretatt.

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

7.6 Vi skal estimere gjennomsnittslønn i en stor befolkning og beslutter å ta et enkelt tilfeldig utvalg på $n = 10$ personer. Det viser seg at kun 6 personer svarte, med følgende årslønner (i tusen): 300, 260, 310, 250, 190, 230. Vi antar at frafallet er rent tilfeldig. To sett av hot-deck imputeringer ble foretatt med følgende resultat:

Hot-deck 1 : 190, 260, 190, 230

Hot-deck 2 : 310, 190, 260, 230.

(a) Beregn standard 95% konfidensintervall for befolkningens gjennomsnittslønn, basert på konstruerte fullstendige utvalg med imputeringsverdiene fra hot-deck 1.

(b) Beregn standard 95% konfidensintervall for befolkningens gjennomsnittslønn, basert på konstruerte fullstendige utvalg med imputeringsverdiene fra hot-deck 2.

(c) Beregn standard 95% konfidensintervall for befolkningens gjennomsnittslønn, basert på svarutvalget.

(d) Bruk multippel imputering til å kombinere standardintervallene i punktene a) og b), både ved Rubin's kombinasjon og modifiseringen av denne. Sammenlign med de tre intervallene i punktene foran.

Appendiks A

1. Sannsynlighetsteori

Formålet med å samle inn data er å trekke konklusjoner om den større populasjonen som data er observert fra. Fundamentet for å kunne gjøre dette er sannsynlighetsteorien , som er en teori om *mekanismen* som generer data. I utvalgsundersøkelser er det trekking av utvalget som generer data.

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

Sannsynlighet er et mål på hvor usikkert det er at en spesiell *begivenhet* skal inntreffe, f.eks. at en gitt enhet *i* i populasjonen skal trekkes til utvalget.

Generelt kan man si at sannsynlighetsteori dreier seg om forsøk eller aktiviteter hvor man ikke kan forutsi på forhånd hva resultatet skal bli.

Disse aktivitetene skal vi kalle «stokastiske forsøk». Dvs.:

Stokastiske forsøk: En prosess eller aktivitet som har minst to mulige utfall.

- Eksempel. Trekking av et enkelt tilfeldig utvalg på *n* enheter fra en populasjon på *N* enheter.
- Eksempel. Måling av ozonlaget over Tromsø.

For å illustrere begreper i sannsynlighetsteorien så skal vi se på noen enklere eksempler .

Eksempel 1.1 : Stokastisk forsøk : Kast med en terning.

Eksempel 1.2 : Stokastisk forsøk : Trekke et enkelt tilfeldig utvalg på 2 personer fra en populasjon på 4 personer, $U = \{1,2,3,4\}$.

Eksempel 1.3 : Stokastisk forsøk : Trekke 2 personer tilfeldig med tilbakelegging fra populasjonen i eksempel 1.2 . Dvs. Vi trekker en og en person etter hverandre. En person som er trukket «legges tilbake» i populasjonen og kan trekkes på nytt.

Eksempel 1.4 : Stokastisk forsøk : Trekke en kule fra en urne med to røde og tre hvite kuler.

Eksempel 1.5 : Stokastisk forsøk: Trekke ut 10 personer i Sagene bydel (Oslo) til forbruks-undersøkelsen 1995, og notere antall personer det er tilsammen i de husholdningene disse tilhører.

Sannsynlighetsteori dreier seg om utfall av stokastiske forsøk. Før vi kommer til selve sannsynlighetsbegrepet trenger vi å innføre en del andre begreper (illustrert med eksemplene ovenfor).

1. ***Utfall*** : En mulig verdi av et stokastisk forsøk. Et stokastisk forsøk vil resultere i ett og bare ett utfall, noen ganger kalt enkeltutfall.

2. ***Utfallsrom S*** (sample space) : Samlingen av alle mulige utfall .

Eks. 1.1: $\boldsymbol{S} = \{1,2,3,4,5,6\}$

Eks. 1.2 : $\boldsymbol{S}$ = alle mulige utvalg på 2 forskjellige personer, dvs. $\boldsymbol{S} = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$

Eks. 1.3: $\boldsymbol{S}$ = alle mulige utvalg på 2, dvs.
$\boldsymbol{S}=\{(1,1),(1,2),(1,3),(1,4),(2,2),(2,3),(2,4),(3,3),(3,4),(4,4)\}$.

Eks. 1.4 : **S** = { R1,R2,H1,H2,H3}, der de røde og hvite kulene er nummererte.

Eks. 1.5 : $\boldsymbol{S}$ = { 10,11,12,...N} hvor N er antall innbyggere i Sagene bydel.

3. Begivenheter. En begivenhet A er en samling av mulige utfall, vanligvis karakterisert ved en felles egenskap. Dvs., A er en delmengde av ***S***. Vi sier at A inntreffer hvis et av de mulige utfall i A er resultatet av det stokastiske forsøket. Vi bruker store bokstaver A,B,C, ... for å betegne begivenheter. Illustrasjoner :

Eks. 1.1 : A = {like tall} = {2,4,6}.

Eks. 1.2 : A = {Enhet 2 er med i utvalget} = {(1,2),(2,3),(2,4)}

Eks. 1.3 : A = {Enhet 2 er med i utvalget} = {(1,2),(2,2),(2,3),(2,4)}

Eks. 1..4 : A = { Kulen er rød } = { R1,R2}

Eks. 1.5 : A = { Det totale antall personer i de uttrukne husholdningene ligger mellom 15 og 20} = {15,16,17,18,19,20}.

Begrepet sannsynlighet

Sannsynligheten til en begivenhet A , betegnet med P(A), skal angi sjansen for at A inntreffer. Vi skal definere P(A) og utlede en teori basert på rimelige egenskaper (kalt aksiomer) ved å tenke oss at det stokastiske forsøket kan gjentas mange ganger.

La ***S*** = {e_1 , ..., e_k} bestå av *k* (enkelt)utfall. A er en begivenhet. Anta forsøket gjentas *n* ganger, og la m(A) være antall ganger A inntreffer i løpet av disse *n* forsøkene. Andel (dvs. relativ antall) ganger A inntreffer blir da

$r_n(A) = m(A)/n$.

P(A) defineres som grenseverdien til $r_n(A)$ når *n* går mot uendelig. Det er umulig å bruke denne definisjonen til å *beregne* P(A), men den kan brukes til å stille opp aksiomer og utlede regneregler som kan brukes til å beregne sannsynligheter for kompliserte begivenheter.

Tilsvarende for enkeltutfallene : La $m(e_i)$ være antall ganger e_i er utfallet av de *n* forsøkene. Da er $P(e_i)$ grenseverdien til $r_n(e_i) = m(e_i)/n$.

Eksempel. Kast med mynt. Når vi sier at P(Kron) =1/2 så mener vi at i gjentatte forsøk så vil Kron inntreffe i 50% av kastene i det lange løp.

Egenskaper til P(e_i)

Andel ganger hver e_i inntreffer , $r_n(e_i)$, har følgende egenskaper :

(1) $0 \le r_n(e_i) \le 1$ for alle i.

(2) $r_n(A) = m(A)/n = \sum_{e_i \in A} m(e_i)/n = \sum_{e_i \in A} r_n(e_i)$

(3) $r_n(\boldsymbol{S}) = 1$

(Betegnelsen $e_i \in A$ betyr alle utfall i A).

Fra (1)-(3) ses at P(A) har følgende egenskaper :

(A1) $0 \le P(e_i) \le 1$

(A2) $P(A) = \sum_{e_i \in A} P(e_i)$

(A3) $P(\boldsymbol{S}) = \sum_{i=1}^{k} P(e_i) = 1$

Den viktigste av disse er (A2) : P(A) er summen av $P(e_i)$ for alle utfall e_i i A.

Et viktig spesialtilfelle :

Den uniforme sannsynlighetsmodellen: Anta alle utfall i $\boldsymbol{S}$ er like sannsynlige : $P(e_i) = 1/k$.
Da blir P(A) = m/k hvor m er antall utfall i A.

Eks. 1.1 : Kaster en terning . $\boldsymbol{S}$ = {1,2,3,4,5,6}, alle 6 utfall er like sannsynlige. A = {like tall} = {2,4,6}. P(A) = 3/6 = 0,5 .

Eks. 1.2 : Trekker et enkelt tilfeldig utvalg på 2 personer fra en populasjon på 4 personer, *U* = {1,2,3,4}. $\boldsymbol{S}$ = {(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)}. Et enkelt tilfeldig utvalg betyr at alle utvalg av størrelse 2 er like sannsynlige, dvs. alle 6 utfall i $\boldsymbol{S}$ er like sannsynlige. A = {Enhet 2 er med i utvalget} = {(1,2),(2,3),(2,4)}, og P(A) = 3/6 = 0,5 .

Eks. 1.3 : Trekker 2 personer tilfeldig med tilbakelegging fra populasjonen i eksempel 1.2 . **S** = {(1,1),(1,2),(1,3),(1,4),(2,2),(2,3),(2,4),(3,3),(3,4),(4,4)}, men disse utfallene er ikke like sannsynlige. For å få til det så må vi ta hensyn til hvilken rekkefølge personene ble valgt. La **S*** være utfallsrommet for disse utfallene. Dvs. la et utfall være (i,j) hvor i er den første enheten som blir trukket og j er den andre enheten. Da består **S*** av alle mulige kombinasjoner av (i,j), 16 ialt. Disse utfallene er like sannsynlige. Begivenheten A = {Enhet 2 er med i utvalget} blir nå {(1,2),(2,1), (2,2),(2,3),(3,2),(2,4),(4,2)} og dermed P(A) = 7/16 = 0,4375 .

Eks. 1.4 : Trekker en kule fra en urne med to røde og tre hvite kuler. **S** = { R1,R2,H1,H2,H3}, der de røde og hvite kulene er nummererte, og alle 5 utfall er like sannsynlige. A = { Kulen er rød } = {R1,R2} og P(A) = 2/5 = 0,4.

Eksempel 1.6 : En student blir gitt 4 utsagn som er sanne eller usanne. Anta studenten gjetter svaret. Se på begivenhetene A = {minst ett riktig svar} og B = {3 eller 4 riktige svar}. La for et gitt utsagn R bety riktig svar og G galt svar. Utfallsrommet $\boldsymbol{S}$ består da av alle sekvenser av R, G. Ialt k =16 utfall (16 = 2·2·2·2). Alle utfall er like sannsynlige. Det eneste utfall som ikke er i A er GGGG, og dermed P(A) = 15/16 = 0,9375. B består av enkeltutfallene RRRG,RRGR,RGRR,GRRR, og RRRR, svarende til utsagn 1,2,3,4. Antall utfall i B er 5 og dermed : P(B) = 5/16 = 0,3125. Hvis det ikke spiller noen rolle hvilke utsagn som er korrekte så kunne man beskrive resultatet av forsøket som *antall* riktige svar : $\boldsymbol{S}$ = {0,1,2,3,4}. Legg merke til at disse utfallene er ikke like sannsynlige. Vi har følgende verdier (se oppgave A1) : P(0) = 1/16, P(1) = 4/16, P(2) = 6/16, P(3) = 4/16, P(4) = 1/16.

Regneregler for kombinasjoner av begivenheter

Anta vi har to begivenheter A,B og vi er interessert i å finne sannsynligheten for at A eller B eller begge to inntreffer. Denne kombinasjonen av A og B betegnes med A∪B og kalles *unionen* av A og B. A∪B består av utfall som er med i A eller B eller begge. Vi kan også si at A∪B er begivenheten at minst en av begivenhetene A eller B inntreffer.

Snittet av A og B består av de utfall som er felles for A og B, og betegnes med A∩B. Dette er begivenheten at både A og B inntreffer samtidig.

Komplementet til en begivenhet A er begivenheten at A ikke inntreffer og betegnes med A^c. A^c består av de utfall som ikke er i A.

Eks. 1.1 : Kaster en terning . ***S*** = {1,2,3,4,5,6}. A = {like tall} = {2,4,6}. A^c = {odde tall} = {1,3,5}.

Eks. 1.2 : Trekker et enkelt tilfeldig utvalg på 2 personer fra en populasjon på 4 personer, U = {1,2,3,4}. **S** = {(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)}. La A = {enhet 2 er med i utvalget} og B= { enhet 3 er med i utvalget}. Da er $A \cap B$ = {2,3}, utvalget består av enhetene 2,3. $A \cup B$ = {2 eller 3 er med i utvalget} = {(1,2),(1,3),(2,3),(2,4),(3,4)}, og $(A \cup B)^c$ = {(1,4)}.

Eks. 1.4 : Trekker en kule fra en urne med to røde og tre hvite kuler. **S** = { R1,R2,H1,H2,H3}. La A = {Kulen er rød } og B = { Kulen er hvit}. Snittet $A \cap B$ blir en umulig begivenhet og betegnes med $\varnothing$, den tomme mengden. $A \cup B$ = { Kulen er rød eller hvit } = **S**.

To begivenheter A og B sies å være *disjunkte* hvis de ikke har noen felles utfall, skrives $A \cap B = \varnothing$.

Regneregel 1. $P(A^c) = 1 - P(A)$.

Regneregel 2. (Addisjonsregelen) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Regneregel 3. *Hvis A og B er disjunkte så er* $P(A \cup B) = P(A) + P(B)$.

Regneregel 4. *Hvis* $A_1, .., A_m$ *alle er disjunkte så er* $P(A_1 \cup A_2 \cup ... A_m) = P(A_1) + P(A_2) + ... + P(A_m)$

$A_1 \cup A_2 \cup ... A_m$ *er begivenheten at minst en av begivenhetene* $A_1,...,A_m$ *inntreffer.*

Bevis. Vi trenger følgende to egenskaper :

(A2) $P(A) = \sum_{e_i \in A} P(e_i)$

(A3) $P(\boldsymbol{S}) = \sum_{i=1}^{k} P(e_i) = 1$

3. $P(A \cup B) = \sum_{e_i \in A \cup B} P(e_i) = \sum_{e_i \in A} P(e_i) + \sum_{e_i \in B} P(e_i) = P(A) + P(B)$.

1. $A \cup A^c = \boldsymbol{S}$, og A, A^c er disjunkte slik at $1 = P(\boldsymbol{S}) = P(A) + P(A^c)$ og dette gir at $P(A^c) = 1 - P(A)$.

4. Følger på samme måte som regel 3.

2. Den vanskeligste å vise. La : $I = A \cap B^c$, $II = A \cap B$, og $III = A^c \cap B$.

Vi ser at $A \cup B = I \cup II \cup III$ og I, II ,III er disjunkte. Dermed : $P(A \cup B) = P(I) + P(II) + P(III)$.

Vi ser også: $P(A) = P(I) + P(II)$, $P(B) = P(III) + P(II)$, og $P(A \cap B) = P(II)$.

Dette gir at $P(A) + P(B) = P(I) + 2P(II) + P(III) = P(A \cup B) + P(II) = P(A \cup B) + P(A \cap B)$.

Dermed : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Eks. 1.1 : Kaster en terning . **S** = {1,2,3,4,5,6}. A = {like tall} = {2,4,6}. A^c = {odde tall} = {1,3,5}. $P(A^c) = 1 - P(A) = 1 - 0,5 = 0,5$.

Eks. 1.2 : Et enkelt tilfeldig utvalg på 2 personer fra en populasjon på 4 personer, U = {1,2,3,4}. **S** = {(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)}. La A = {enhet 2 er med i utvalget} og B= { enhet 3 er med i

utvalget}. Da er A∩B = {(2,3)}, utvalget består av enhetene 2,3. A∪B = {2 eller 3 er med i utvalget} ={(1,2),(1,3),(2,3),(2,4),(3,4)}. Vi ser direkte at P(A∪B) = 5/6. Dette kan også beregnes ved regneregel 2. P(A) = P(B) = 3/6 = 0,5, og P(A∩B) = 1/6 som gir at P(A∪B) = 1 - 1/6 = 5/6.

Eks. 1.4 : Trekker en kule fra en urne med to røde og tre hvite kuler. **S** = { R1,R2,H1,H2,H3}. La A = {Kulen er rød } og B = { Kulen er hvit}. A∪B = { Kulen er rød eller hvit } = **S**. P(B) = P(A^c) = 1 - P(A) = 3/5 = 0,6.

Betinget sannsynlighet og stokastisk uavhengighet

Betinget sannsynlighet

Vi har to begivenheter A og B. Anta vi vet at B har inntruffet. Det er da nødvendig å modifisere sannsynligheten for A ut fra denne viten. B blir på en måte det nye utfallsrommet. Denne modifiserte sannsynligheten kalles den betingede sannsynlighet for A gitt B. Betegnes med P(A|B). P(A|B) forteller oss sjansen for at A inntreffer når B har inntruffet.

Eksempel 1.2. Anta populasjonen på 4 personer består av 2 kvinner og 2 menn. Det stokastiske forsøket består i å velge to personer tilfeldig, etter hverandre- *uten tilbakelegging.* Se på følgende to begivenheter:

B = {Første person er kvinne}, A = {Den andre personen er mann}. Da er P(A|B) = 2/3 .

Fordi, gitt at B har inntruffet så består utfallsrommet ved 2. trekning av 2 menn og 1 kvinne.

La oss også finne P(A∩B) og P(B):

$$P(B) = 2/4 = 1/2.$$

For å finne P(A∩B) trenger vi å se på hele forsøket. La K_1,K_2 være de to kvinnene, og M_1,M_2 være de to mennene. Utfallsrommet ***S*** er ikke det samme som før for nå teller også rekkefølgen av trekkingene. ***S*** består av alle mulige parvise kombinasjoner av (K_1,K_2 ,M_1,M_2), 12 ialt. Dette følger fordi ved første trekning er det 4 muligheter, og ved andre trekning er det 3 mulige resultater for hvert resultat av første trekning. Herav 4·3 = 12 mulige utfall, alle like sannsynlige. A∩B består av utfallene : {K_1M_1,K_1M_2,K_2M_1,K_2M_2}, Dvs. A∩B består av 4 utfall, og vi får derfor at

$$P(A\cap B) = 4/12 = 1/3.$$

Legg merke til at $P(A\cap B)/P(B) = \frac{1/3}{1/2} = \frac{2}{3} = P(A|B)$. Dette er ingen tilfeldighet.

Eksempel 1.7. Anta vi har en befolkning inndelt i en krysstabell etter variablene røyking og blodtrykk. I tabellen oppgis andelen i de forskjellige gruppene.

	Røykere	**Ikke-røykere**	**Totalt**
Høyt blodtrykk	0,2	0,1	0,3
Normalt blodtrykk	0,2	0,5	0,7
Totalt	0,4	0,6	1,0

Det stokastiske forsøket består i å trekke én person rent tilfeldig fra denne populasjonen. Vi skal se på følgende begivenheter :

A = {person har høyt blodtrykk}, P(A) = 0,3.

B = { person røyker}.

Gitt B: 2.kolonne er ikke relevant. Andel med høyt blodtrykk blant røykere er 50%.

Dermed:

$$P(A|B) = 0{,}5 \ .$$

Også:

$$\frac{P(A \cap B)}{P(B)} = \frac{0{,}2}{0{,}4} = \frac{1}{2} = \ P(A|B) \ .$$

Dvs. her er P(A|B) andelen blant de med kjennetegn B som har kjennetegn A.

I begge eksemplene så has at P(A|B) = P(A∩B)/P(B). Dette gjelder generelt.

Generelt tilfelle: $\boldsymbol{S} = \{e_1, \ . \ . \ . \ e_k \}$ med like sannsynlige utfall. La *b* være antall utfall i B, *c* antall utfall i A∩B. Gitt B : B er utfallsrom med *b* mulige utfall, hvorav *c* er gunstig for A. Dermed fra uniform sannsynlighetsmodell:

$$P(A|B) = \frac{c}{b} = \frac{c/k}{b/k} = \frac{P(A \cap B)}{P(B)} \ .$$

Dette gjelder også når enkeltutfallene e_i har forskjellige sannsynligheter :

DEFINISJON: P(A|B) = P(A∩B)/P(B), hvis P(B) > 0.

Dette gir oss også:

Multiplikasjonsregelen : P(A∩B) =P(A|B)P(B).

Også: P(B|A) = P(A∩B)/P(A) slik at P(A∩B) = P(A) P(B|A), hvilket gir oss regneregel 5:

Regneregel 5 (multiplikasjonsregelen). P(A∩B) =P(A|B)P(B) = P(A) P(B|A).

Hvis man kan beregne P(A|B) eller P(B|A) direkte kan man finne P(A∩B).

Eks. 1.2, forts. : $P(A \cap B) = P(B)P(A|B) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}$, lettere enn før.

Stokastisk uavhengighet

Uavhengige begivenheter

A,B sies å være uavhengige begivenheter hvis informasjon om at B har inntruffet ikke forandrer sannsynligheten for A. Dvs., B gir ingen ekstra informasjon om A's hendelse. Presist:

Definisjon. A, B er uavhengige hvis P(A|B) = P(A). Ellers sier vi at A og B er avhengige.

Ekvivalente betingelser for uavhengighet :

(1) A , B er uavhengige ⇔ P(A ∩ B) = P(A|B)P(B) = P(A)P(B).

(2) A , B er uavhengige $\Leftrightarrow$ P(B|A) = P(B),

fordi :

$$P(\mathrm{B}|\mathrm{A}) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B) .$$

Eksempel 1.8.

(a) Trekking med tilbakelegging.

Populasjon: 6 personer, 4 yrkesaktive og 2 arbeidsledige (ikke yrkesaktive). Trekker to personer med tilbakelegging, rent tilfeldig. Det betyr at den personen som blir trukket først også kan bli valgt ut andre gangen. La

A= { 1. person er yrkesaktiv}

B = { 2. person er arbeidsledig}

P(B|A) = 2/6 = P(B). Dermed : A og B er uavhengige. Trekkingene er uavhengige når de foretas med tilbakelegging.

(b) Trekking uten tilbakelegging.

Se på samme begivenheter . Det er nå mer komplisert å beregne P(B).

$B = (A \cap B) \cup (A^c \cap B)$, A^c = { 1.person er arbeidsledig}.

$$P(B) = P(A \cap B) + P(A^c \cap B) = P(A)P(B|A) + P(A^c)P(B|A^c)$$

$$= \frac{4}{6} \cdot \frac{2}{5} + \frac{2}{6} \cdot \frac{1}{5} = \frac{10}{30} = \frac{1}{3} .$$

P(B|A) = 2/5 ≠ P(B). Dermed : A og B er avhengige . Trekkingene er avhengige når de foretas uten tilbakelegging.

2. Fordelinger, tilfeldige variable

2.1. Fordelinger- hypergeometrisk og binomisk

Enkelt tilfeldig utvalg- hypergeometrisk fordeling

Populasjonen består av N enheter. La n være størrelsen på utvalget, som er bestemt på forhånd. Vi trekker nå en og en enhet *uten tilbakelegging* , dvs. en person kan bare trekkes en gang. Ved første trekking lar vi alle enhetene i populasjonen ha samme sannsynlighet $1/N$ for å bli valgt. Etter at første enhet er valgt, velges neste slik at alle N-1 gjenværende enheter har samme sannsynlighet for å bli trukket ut. osv. til n enheter er trukket ut til utvalget.

La s betegne de enhetene som blir trukket til utvalget. Det kan vises at hvert utvalg s av størrelse n har samme sannsynlighet for å bli valgt, og dette er også den vanlige definisjonen på *enkelt tilfeldig utvalg :*

DEFINISJON. Vi har et *enkelt tilfeldig utvalg* av størrelse n, hvis alle utvalg s av størrelse n har samme sannsynlighet for å bli valgt.

Prosedyren ovenfor er *en* måte å implementere denne utvalgsplanen. Det finnes andre, mer tidsbesparende, trekkemetoder som også oppnår enkle tilfeldige utvalg. En slik metode er følgende:

a) La alle personene i populasjonen få tilordnet forskjellige tilfeldige tall mellom 0 og 1.

b) Sorter personene etter størrelsene på deres tilfeldige tall, og la de n første være utvalget.

Anta vi er interessert i å finne ut noe om antall i populasjonen med et visst kjennetegn (egenskap) A, f.eks. antall sysselsatte. Vi er da interessert i variabelen X = antall med kjennetegn A i utvalget. Hvis vi lar p være andelen med kjennetegn A i populasjonen, så kunne vi bruke X/n som et anslag for p. Vi sier da at X/n er en estimator for p. Vi skal nå studere fordelingsegenskapene til variabelen X. La M være antall med kjennetegn A i populasjonen, slik at $p = M/N$. Vi skal anta at $M > n$. X kan ta verdiene 1, ... n, med visse sannsynligheter.

Det kan vises at

$$\mathrm{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \quad \text{for } x = 1,2,....,n.$$

(Notasjon: Utelater multiplikasjonstegnet i uttrykk som $a \cdot b = ab$ og $\frac{a \cdot b}{c \cdot d} = \frac{ab}{cd}$.)

Her er $\binom{N}{k} = \frac{N!}{k!(N-k)!}$ den binomiske koeffisienten.

Fakultetetsfunksjonen for hele tall er gitt ved : $N! = N(N-1)(N-2) \cdot \cdot \cdot \cdot 2 \cdot 1$. Vi kan også få uttrykk av typen $\binom{M}{0} = \binom{M}{M} = \frac{M!}{0!\,M!} = \frac{1}{0!}$. Da defineres 0! = 1 slik at $\binom{M}{0} = \binom{M}{M} = 1$.

Dette kalles den *hypergeometriske* fordelingen, og vi sier at X er hypergeometrisk fordelt.

Dette er et eksempel på en sannsynlighetsfordeling: Forteller hvordan sannsynlighetene *fordeler* seg over alle mulige verdier av X. X er et eksempel på en *tilfeldig variabel.* Vi ser at X er en funksjon av utvalget som trekkes, dvs. X er en tallfunksjon av utfallet av et stokastisk forsøk. Alle slike funksjoner av utfall av stokastiske forsøk kalles tilfeldige variable.

Legg merke til at P($X = x$) = m/k hvor :

$$k = \text{antall utvalg med størrelse } n = \binom{N}{n}$$

m er antall utvalg med x A'er og n-x ikke A'er = $\binom{M}{x}\binom{N-M}{n-x}$.

Et eksempel er lottospillet.

Eksempel 2.1. Lottospillet. Populasjonen består av N = 34 tall , og vi velger M = 7 tall. Så trekkes tilfeldig n = 7 tall som «korrekte». Vi er interessert i kjennetegn A : « tallet er på vår lottokupong». La X = antall korrekte tall på vår kupong. X er da hypergeometrisk fordelt med N,M,n gitt ovenfor. Sannsynligheten for at vi med en rekke skal få k riktige tall er derfor:

$$P(X = k) = \frac{\binom{7}{k}\binom{27}{7-k}}{\binom{34}{7}} \quad , \quad k = 0,1,2,3,4,5,6,7 .$$

Vi får : $\binom{34}{7} = 5.379.616$.

k	7	6	5	4
P($X = k$)	0,00000019	0,000035	0,0014	0.0190

For eksempel, P($X \geq 4$) = 0,020.

Vi kan anvende den hypergeometriske fordelingen til å finne sannsynligheten for at en gitt enhet k i populasjonen blir med i utvalget. La π_k være denne *trekksannsynligheten* for enhet k. La kjennetegn A være «enhet k». Da er X = antall med kjennetegn A i utvalget = 1, hvis enhet k er med og 0 ellers. Her er M = 1, og dermed:

$$\pi_k = P(X = 1) = \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{(N-1)!n!(N-n)!}{(n-1)!(N-n)!N!} = \frac{n}{N} \quad .$$

Eksempel 1.4.,forts. Trekker en kule fra en urne med to røde og tre hvite kuler. A = «rød», og X er 1 hvis rød kule trekkes og 0 ellers. Her er M = 2, N = 5, og n = 1.

$$P(X = x) = \frac{\binom{2}{x}\binom{3}{1-x}}{\binom{5}{1}} \quad , \text{ for } x = 0,1.$$

P(X = 0) = $\binom{3}{1} / \binom{5}{1} = 3 / 5$, og P(X= 1) = 2/5.

Eksempel 2.2. En eksamen består av 5 oppgaver som skal trekkes tilfeldig fra en liste på 10 oppgaver. Hvor mange oppgaver må en student forberede for at sannsynligheten for å besvare minst 4 oppgaver korrekt er 0,75 eller mer. La

M = antall forberedte oppgaver blant de 10.

X = antall riktige svar = antall oppgaver som er forberedt i utvalget på 5.

Vi ønsker å bestemme M slik at P(X = 4 eller 5) = P($X \geq 4$) $\geq$ 0,75. X er hypergeometrisk fordelt med n = 5 og N = 10.

$$\mathrm{P}(X = x) = \frac{\binom{M}{x}\binom{10-M}{5-x}}{\binom{10}{5}} = \frac{\binom{M}{x}\binom{10-M}{5-x}}{252}$$

Med M = 8:

$$\mathrm{P}(X \geq 4) = \mathrm{P}(X = 4) + \mathrm{P}(X = 5)$$

$$= \frac{\binom{8}{4}\binom{2}{5-4}}{252} + \frac{\binom{8}{5}\binom{2}{5-5}}{252} = \frac{70 \cdot 2}{252} + \frac{56 \cdot 1}{252} = \frac{196}{252} = 0{,}778.$$

Trekking med tilbakelegging - binomisk fordeling

Populasjonen består av N enheter og n er størrelsen på utvalget. Vi trekker nå en og en enhet *med tilbakelegging* , dvs. en og samme person kan trekkes flere ganger. Ved hver trekking lar vi alle enhetene i populasjonen ha samme sannsynlighet 1/N for å bli valgt.

Vi er fremdeles interessert i å finne ut noe om antall i populasjonen med et visst kjennetegn (egenskap) A. Som under enkelt tilfeldig utvalg skal vi se på sannsynlighetsfordelingen til variabelen X = antall med kjennetegn A i utvalget. Som før, la M være antall med kjennetegn A i populasjonen, slik at $p = M/N$ er populasjonsandelen med egenskapen A. Trekningene av enhetene er nå uavhengige. Dette medfører at X får en annen sannsynlighetsfordeling. Den kalles den *binomiske* fordelingen og er gitt ved :

$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \ , \ x = 0,1, \ldots, n \, .$$

Eksempel 1.3, fortsettelse. Trekker 2 personer tilfeldig med tilbakelegging fra en populasjon på 4 personer. Vi fant tidligere at P(enhet 2 er med i utvalget) = $\frac{7}{16}$ = 0,4375. Dette kan også vises ved hjelp av den binomiske fordelingen ved å la A = «enhet 2». Da er P(enhet 2 er med i utvalget) = P(X = 1 eller 2), med n = 2 og p = 1/4. Dette gir at

$$\mathrm{P}(X = 1) = \binom{2}{1}(\tfrac{1}{4})^1(1-\tfrac{1}{4})^{2-1} = 2 \cdot \tfrac{1}{4} \cdot \tfrac{3}{4} = \tfrac{6}{16}$$

og

$$\mathrm{P}(X = 2) = \binom{2}{2}(\tfrac{1}{4})^2(1-\tfrac{1}{4})^0 = \tfrac{1}{16}$$

(kan også ses ved at P(X = 2) = P(A 1.gang)P(A 2.gang) = $p \cdot p = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$)

P(X = 1 eller 2) = P(X = 1) + P(X = 2) = 7/16 = 0,4375.

Mer generelt så forekommer den binomiske fordelingen i de såkalte binomiske forsøk som kan beskrives på følgende måte:

Binomiske forsøk

(1) Består av n uavhengige enkeltforsøk med to mulige utfall, begivenhet A eller A^c .

(2) p = P(A) er den samme for alle forsøkene.

A kalles gjerne «suksess», A^c «fiasko», for å ha en generell betegnelse .

La X være antall suksesser i de n forsøkene. Da er X binomisk fordelt.

Eksempel 2.3. Betrakt n gjentatte kast med en mynt. La X være antall kron og p = P(kron) = 1/2. Da er $P(X = x) = \binom{n}{x}(\frac{1}{2})^x(1-\frac{1}{2})^{n-x} = \binom{n}{x}(\frac{1}{2})^n$. F.eks., 3 myntkast gir at $P(X = x) = \binom{3}{x} / 8$.

2.2. Tilfeldige variable, fordelinger, forventning og varians

Tilfeldige variable og fordelinger

Som nevnt tidligere, en tilfeldig variabel X er generelt en tallfunksjon av utfallet av et stokastisk forsøk. Dvs. X anviser et tall $x = X(e)$ for ethvert mulig utfall e i utfallsrommet **S**. La x_1 , ...,x_k betegne de forskjellige verdiene X kan anta. Ofte vil de mulige verdiene være hele tall 0,1, osv. .

Sannsynlighetsfordelingen til X, vanligvis kalt fordelingen, er da samlingen av alle $p(x_i) = P(X = x_i)$ for i = 1, ..., k.

Vi har allerede sett flere eksempler på tilfeldige variable og fordelinger i kapittel 2.1. Et eksempel til er følgende :

Eksempel 2.4. Følgende prisendringer antas mulige : stort avslag, middels avslag, lite avslag, samme pris, moderat økning, og sterk økning. *Sannsynlighetsmodellen* for dette forsøket antar følgende sannsynligheter for enkeltutfallene:

P(stort avslag)	= 0,05
P(middels avslag) = P(sterk økning)	= 0,10
P(lite avslag) = P(moderat økning)	= 0,20
P(samme pris)	= 0,35

Vi definerer følgende tilfeldig variabel X : X angir forholdet mellom ny pris og gammel pris:

X(stort avslag)	= 0,60
X(middels avslag)	= 0,80
X(lite avslag)	= 0,95
X(samme pris)	= 1
X(moderat økning)	= 1,15
X(sterk økning)	= 1,35

Fordelingen til X :

verdi x	**0,60**	**0,80**	**0,95**	**1**	**1,15**	**1,35**
p(x)	0,05	0,10	0,20	0,35	0,20	0,10

Forventning og varians

Forventning

To sentrale egenskaper til en fordeling er forventningen og variansen. Forventningen er et mål for midtpunktet i fordelingen. Den forteller oss verdien av X i gjennomsnitt ved mange gjentatte observasjoner av X. Forventningen til X betegnes med $E(X)$ og defineres ved:

$$\mathrm{E(X)} = \sum_{i=1}^{k} x_i P(X = x_i) = \sum_{i=1}^{k} x_i p(x_i).$$

Dvs., $E(X)$ = sum av verdi·sannsynlighet, og betegnes vanligvis med den greske bokstaven μ (my).

Illustrasjon. Anta X har de mulige verdiene 1,2,...,k. Vi skal nå illustrere at $E(X)$ er tilnærmet lik gjennomsnittet $\bar{x} = \sum_{i=1}^{n} x_i / n$ av n observasjoner $x_1, ..., x_n$ av X. Merk at i denne sammenhengen har x_i 'ene en annen betydning enn i definisjonen ovenfor. La

n_i = antall observasjoner som tar verdien i.

Da er $\sum_{i=1}^{n} x_i = 1 \cdot n_1 + 2 \cdot n_2 + \ldots + k \cdot n_k$.

Dette gir at

$$\bar{x} = \frac{1 \cdot n_1 + 2 \cdot n_2 + \ldots + k \cdot n_k}{n} = 1 \cdot \frac{n_1}{n} + 2 \cdot \frac{n_2}{n} + \ldots + k \cdot \frac{n_k}{n}$$

$$\approx 1 \cdot \mathrm{P}(X = 1) + 2 \cdot \mathrm{P}(X = 2) + .. + k \cdot \mathrm{P}(X = k) = E(X).$$

Eksempel 2.5. Roulettespill. Et roulettehjul har 38 tall ; 00, 0, 1,2,...,36. Alle tall har samme sjanse for å vinne. Anta vi vedder kr.100 på odde tall (1,3,5,...,35). Det er ialt 18 odde tall. La X være gevinsten:

$$X = \begin{cases} 100 \text{ hvis odde tall (vi får } \textit{tilbake} \text{ kr. } 200) \\ -100 \text{ ellers (vi taper innsatsen)}. \end{cases}$$

La oss finne $E(X)$. Fordelingen til X er gitt ved:

$$\mathrm{P}(X = 100) = 18/38 = 9/19 \text{ og } \mathrm{P}(X = -100) = 20/38 = 10/19.$$

$E(\text{Gevinst}) = E(X) = 100 \cdot \frac{9}{19} + (-100) \cdot \frac{10}{19} = -\frac{100}{19} = -kr.$ 5,26. Det betyr at vi taper kr. 5,26 i gjennomsnitt pr. spill. Anta vi spiller 1000 ganger. Da er forventet tap lik kr.5,26·1000 = kr. 5260.

Eksempel 2.4. Fordelingen til prisendringen X :

verdi x	**0,60**	**0,80**	**0,95**	**1**	**1,15**	**1,35**
p(x)	0,05	0,10	0,20	0,35	0,20	0,10

Forventningen til prisendringen blir :

$$E(X) = 0{,}60 \cdot 0{,}05 + 0{,}80 \cdot 0{,}10 + 0{,}95 \cdot 0{,}20 + 1 \cdot 0{,}35 + 1{,}15 \cdot 0{,}20 + 1{,}35 \cdot 0{,}10 = 1{,}015.$$

Det betyr at forventet prisendring er en økning på 1,5% . Ved gjentatte mange målinger av X med denne fordelingen så blir gjennomsnittet av prisendringene omtrent 1,5%. Prisendringene vil imidlertid variere fra måling til måling, de er usikre. Et mål på denne usikkerheten og variasjonen er gitt ved *variansen* til X.

Varians

Variansen til X, *Var(X),* er et mål for spredningen i fordelingen rundt $\mu = E(X)$. *Var(X)* gir gjennomsnittlig verdi av $(X - \mu)^2$, og er definert ved:

$$Var(X) = E(X - \mu)^2 = \sum_{i=1}^{k} (x_i - \mu)^2 p(x_i) .$$

Notasjon : $\sigma^2 = Var(X)$.(σ er den greske bokstaven sigma)

Variansen måles i kvadratet av måle-enheten til *X*. For å uttrykke spredningen i samme enhet så brukes kvadratroten til *Var(X),* kalt standardavviket til *X*, betegnet med *sd(X)*:

$$\sigma = sd(X) = \sqrt{Var(X)} .$$

Eks. 2.4, forts.

$VAR(X) = (0,60 - 1,015)^2 \cdot 0,05 + (0,80 - 1,015)^2 \cdot 0,10 + \quad .. = 0,029025.$

$sd(X) = 0,17$.

Det er vanskelig å tolke *sd(X)* direkte, men vanligvis vil *X* ha høy sannsynlighet for å ta verdi i intervallet $(\mu - 2\sigma, \mu + 2\sigma)$. Det kan vises at den er minst 0,75, og ofte 0,90 til 0,95.

Egenskaper til *E(X)* og *Var(X)* (bevis henvises til oppgave A9)

- *(E1) E(a) = a, a* er en konstant
- *(E2) E(a + bX) = a + bE(X), a* og *b* er konstanter
- *(V1) Var(a) = 0 , a* er en konstant
- *(V2) Var(a + bX) = b^2Var(X), a* og *b* er konstanter
- *(V3) Var(X) = $E(X^2)$ - μ^2* .

Vi skal nå se på forventning og varians for summer av tilfeldige variable, og starter med summen av *to* variable *X* og *Y, X + Y.* Alle utledninger gjøres kun for 0/1- variable. Dvs. vi skal konsentrere oss om variable av typen :

$$X = \begin{cases} 1 & \text{hvis begivenhet A inntreffer} \\ 0 & \text{ellers} \end{cases}$$

$$Y = \begin{cases} 1 & \text{hvis begivenhet B inntreffer} \\ 0 & \text{ellers} \end{cases}$$

Et eksempel :

X : yrkesaktiv/ikke yrkesaktiv og *Y :* mann/kvinne

Slike 0/1-variable (også kalt indikatorvariable) er sentrale for utvalgsteorien. Av spesiell interesse for utvalgsteorien er indikatorvariablene for enhetene i utvalget, dvs. av typen: $X = 1$ hvis enhet *i* trekkes til utvalget, og 0 ellers.

Alle resultatene som gjelder for 0/1- variable gjelder også generelt for tilfeldige variable. La nå

$$p_A = P(A) = P(X = 1)$$

$$p_B = P(B) = P(Y = 1)$$
$$p_{AB} = P(A \cap B) = P(X = 1 \cap Y = 1) .$$

Vi ser at

$$\mu_X = E(X) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = P(X = 1) = p_A$$
$$\mu_Y = E(Y) = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) = P(Y = 1) = p_B .$$

Videre fås:

$$\underline{E(X+Y)} = 0 \cdot P(X + Y = 0) + 1 \cdot P(X + Y = 1) + 2 \cdot P(X + Y = 2)$$
$$= P(X = 1 \cap Y = 0) + P(X = 0 \cap Y = 1) + 2\, P(X = 1 \cap Y = 1)$$
$$= P(X = 1 \cap Y = 0) + P(X = 1 \cap Y = 1) + P(X = 0 \cap Y = 1) + P(X = 1 \cap Y = 1)$$
$$= P(X = 1) + P(Y = 1) = \underline{E(X) + E(Y).}$$

$$\underline{Var(X+Y)} = E\{(X+Y) - (\mu_X + \mu_Y)\}^2$$
$$= E\{(X - \mu_X) + (Y - \mu_Y)\}^2$$
$$= E\,(X - \mu_X)^2 + E\,(Y - \mu_Y)^2 + 2E\{(X - \mu_X)(Y - \mu_Y)\}$$
$$= Var(X) + Var(Y) + 2E\{(X - \mu_X)(Y - \mu_Y)\}.$$

Uttrykket $E\{(X - \mu_X)(Y - \mu_Y)\}$ kalles kovariansen mellom X og Y, betegnet med $Cov(X,Y)$, og er et mål på avhengigheten mellom X og Y. Generelt har vi at

$$Cov(X,Y) = E(XY) - \mu_X \mu_Y .$$

Dette ses på følgende vis, generelt:

$$E\{(X - \mu_X)(Y - \mu_Y)\} = E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y)$$
$$= E(XY) - E(\mu_X Y) - E(\mu_Y X) + E(\mu_X \mu_Y)$$
$$= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X \mu_Y$$
$$= E(XY) - \mu_X \mu_Y .$$

For indikatorvariablene :

$$E(XY) = P(XY) = 1 = P(X = 1 \cap Y = 1) = P(A \cap B)$$
$$\mu_X \mu_Y = P(A)P(B)$$

og $$Cov(X,Y) = P(A \cap B) - P(A)P(B).$$

DEFINISJON. X og Y sies å være uavhengige hvis A,B er uavhengige begivenheter.

Hvis X,Y er uavhengige så er $P(A \cap B) = P(A)P(B)$, dvs. $E(XY) = E(X)E(Y)$, og dermed $Cov(X,Y) = 0$.

Vi kan nå summere opp noen egenskaper til forventning og varians :

- *(E3)* $E(X + Y) = E(X) + E(Y)$
- *(E4)* $E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$
- *(V4)* $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
- *(V5)* $Var(X + Y) = Var(X) + Var(Y)$, hvis X og Y er uavhengige
- *(V6)* $Var(X_1 + X_2 + \ldots + X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n)$, hvis $X_1, X_2, \ldots, X_n$ er uavhengige.

I tillegg har vi, fra (E2) og (V2) at $E(aX) = aE(X)$ og $Var(aX) = a^2Var(X)$, slik at vi kan bruke (E4) og (V6) til å finne forventning (generelt) og varians (uavhengige variable) for lineære kombinasjoner $a_1X_1 + a_2X_2 + \ldots + a_nX_n$. Resultatene gjelder for alle typer av tilfeldige variable. Generell definisjon av uavhengighet sier at X og Y er uavhengige hvis, for alle mulige verdier x,y av X,Y, utfallene $\{X = x\}$ og $\{Y= y\}$ er uavhengige begivenheter.

Forventning og varians i binomisk og hypergeometrisk fordeling

Binomisk fordeling

La oss først betrakte binomisk fordeling generelt. Den forekommmer i binomiske forsøk:

(1) Består av n uavhengige enkeltforsøk med to mulige utfall, begivenhet A eller A^c .

(2) p = P(A) er den samme for alle forsøkene.

A kalles «suksess» og A^c «fiasko». Den binomiske variabelen X er antall suksesser i de n forsøkene.

Vi kan utrrykke X som summen av indikatorvariablene til A for de n enkeltutfallene. La

$$X_i = \begin{cases} 1 \text{ hvis A inntreffer i forsøk } i \\ 0 \text{ ellers} \end{cases}$$

Siden forsøkene er uavhengige, så er $X_1, \ldots, X_n$ uavhengige variable. $X = \sum_{i=1}^{n} X_i$. Dermed :

$$E(X) = \sum_{i=1}^{n} E(X_i) \quad \text{og } Var(X) = \sum_{i=1}^{n} Var(X_i) .$$

$E(X_i) = P(X_i = 1) = p$, og $Var(X_i) = E(X_i^2) - p^2 = E(X_i) - p^2 = p - p^2 = p(1-p)$.
Det følger at :

$$E(X) = np$$

$$Var(X) = np(1-p).$$

Hypergeometrisk fordeling

La oss nå betrakte den hypergeometriske fordelingen som forekommer i forbindelse med enkelt tilfeldig utvalg.

Populasjonen består av N enheter, og n er størrelsen på utvalget. Vi trekker et enkelt tilfeldig utvalg, dvs. alle utvalg av størrelse n har samme sannsynlighet for å bli valgt. Vi er interessert i variabelen X

= antall med kjennetegn A i utvalget. Vi lar p være andelen med kjennetegn A i populasjonen, slik at $p = M/N$, hvor M er antall med kjennetegn A i populasjonen. Igjen kan vi uttrykke X som summen av indikatorvariable, nå definert ved

$$X_i = \begin{cases} 1 \text{ hvis enhet fra i' te trekking i utvalget har kjennetegn A} \\ 0 \text{ ellers} \end{cases}$$

I denne situasjonen er X_i 'ene *ikke* uavhengige. Vi skal først se på situasjonen med $n = 2$. Argumentasjonen fra dette tilfellet kan deretter generaliseres. Vi har

$$E(X) = E(X_1) + E(X_2) \text{ og } Var(X) = Var(X_1) + Var(X_2) + 2Cov(X_1, X_2).$$

$$E(X_1) = P(X_1 = 1) = M/N = p$$

$$E(X_2) = P(X_2 = 1) = P(X_1 = 1 \cap X_2 = 1) + P(X_1 = 0 \cap X_2 = 1)$$

$$= P(X_1 = 1)P(X_2 = 1 \mid X_1 = 1) + P(X_1 = 0)P(X_2 = 1 \mid X_1 = 0)$$

$$= \frac{M}{N} \cdot \frac{M-1}{N-1} + \frac{N-M}{N} \cdot \frac{M}{N-1} = \frac{M(M-1) + M(N-M)}{N(N-1)}$$

$$= \frac{M(M-1+N-M)}{N(N-1)} = \frac{M(N-1)}{N(N-1)} = p$$

Dette gir at $\underline{E(X) = 2p = np}$, *gjelder for generell n*

I tillegg: $Var(X_1) = Var(X_2) = p(1-p)$.

$E(X_1X_2) = P(X_1 = 1 \cap X_2 = 1) = \frac{M}{N} \cdot \frac{M-1}{N-1}$ hvilket gir at

$$Cov(X_1, X_2) = \frac{M}{N} \cdot \frac{M-1}{N-1} - \frac{M}{N} \cdot \frac{M}{N} = p\frac{(M-1)N - M(N-1)}{(N-1)N}$$

$$= p\frac{M-N}{N(N-1)} = -p\frac{N-M}{N} \cdot \frac{1}{N-1} = -p(1-p)/(N-1).$$

Herav: $Var(X) = 2p(1-p) - 2p(1-p)\frac{1}{N-1} = 2p(1-p)\frac{N-2}{N-1}$.

$$Var(X) = np(1-p)\frac{N-n}{N-1}, \text{ for generell } n.$$

Vi kan nå sammenfatte forventning og varians i binomisk (trekking med tilbakelegging) og hypergeometrisk fordeling (trekking uten tilbakelegging):

Binomisk fordeling: $E(X) = np$
$Var(X) = np(1-p)$

Hypergeometrisk fordeling : $E(X) = np$

$$Var(X) = np(1\text{-}p)\frac{N-n}{N-1} \approx np(1-p)\left(1-\frac{n}{N}\right)$$

Legg merke til at når *n/N* er liten så er $np(1-p)\left(1-\frac{n}{N}\right) \approx np(1\text{-}p)$, den binomiske variansen. For eksempel, la $n = 1000$, og N= 1000 000, så er $np(1-p)\left(1-\frac{n}{N}\right) = 999p(1\text{-}p)$ og $np(1\text{-}p) = 1000p(1\text{-}p)$. Med $p = 0{,}05$, f.eks., fås 47,45 og 47,50 henholdsvis. Det betyr at når n/N er liten så er trekking med og uten tilbakelegging omtrent det samme, sannsynlighetsteoretisk .

Eksempel 2.6. Vi har en populasjon på 100 personer, hvorav 10 er arbeidsledige. Et enkelt tilfeldig utvalg på $n = 30$ personer trekkes, og X = antall arbeidsledige i utvalget. Siden X er hypergeometrisk fordelt med $N = 100$, $n = 30$ og $p = 10/100 = 0{,}1$ fås at :

$$E(X) = np = 3$$

og

$$Var(X) = np(1\text{-}p)\frac{N-n}{N-1} = 30 \cdot (0{,}1) \cdot (0{,}9) \cdot \tfrac{70}{99} = 1{,}91.$$

$$sd(X) = \sqrt{1{,}91} = 1{,}38.$$

Ved gjentatte enkle tilfeldige utvalg på 30 personer så vil gjennomsnittet av antall ledige i utvalget bli omtrent lik 3, og verdiene av X vil stort sett variere i intervallet $(\mu - 2\sigma, \mu + 2\sigma)$, dvs., mellom 0 og 6.

Oppgaver til Appendiks A

A.1 Viser til eksempel 1.6. Utfall k : k riktige svar. Vis at

P(0 riktige svar) = 1/16
P(1 riktig svar) = 4/16
P(2 riktige svar) = 6/16
P(3 riktige svar) = 4/16
P(4 riktige svar) = 1/16

A.2 Vi trekker et enkelt tilfeldig utvalg på 3 personer fra en populasjon på 5 personer, nummerert fra 1 til 5. Det betyr at alle utvalg på 3 personer er like sannsynlige. Betrakt følgende begivenheter :

A : Person 1 er med i utvalget
B : Person 2 er med i utvalget.

(a) Finn P(A) og P(B).
(b) Beskriv i ord begivenhetene : $A \cup B$, $A \cap B$, A^c, B^c .
(c) Uttrykk begivenhetene i punkt (b) som mengder av utfall.
(d) Beregn $P(A \cup B)$, $P(A \cap B)$, $P(A^c)$, $P(B^c)$ ved å bruke regnereglene 1 -3.
(e) Sjekk resultatene i punkt (d) ved å bruke punkt (c).

A.3 Betrakt forsøket i oppgave A.2. Anta at personene 1,3 og 4 er sysselsatte, de andre ikke. La X betegne antall sysselsatte i utvalget. Finn følgende sannsynligheter :

(a) P(X = 0)
(b) P(X = 1)
(c) P(X = 3)
(d) P(X = 1 eller 3).

A.4 Stokastisk forsøk: Kaster to terninger, en hvit og en rød.

(a) Hvor mange mulige utfall har forsøket?
(b) Finn sannsynligheten for at summen av terningverdiene er lik 7.
(c) Finn sannsynligheten for at summen er 3 eller mer.

A.5 Betrakt eksempel 1.7. Finn følgende sannsynligheter direkte fra tabellen:

(a) P(valgte person røyker $\mid$ person har høyt blodtrykk).
(b) P(valgte person røyker $\mid$ person har normalt blodtrykk).
(c) P(valgte person ikke røyker).

A.6 Betrakt alle familier i Norge med 2 barn og minst en gutt.
Hvor stor andel av disse har 2 gutter ?

A.7 Vi har en gruppe på 7 personer, 4 EU-motstandere og 3 EU-tilhengere. Et enkelt tilfeldig utvalg på 2 personer trekkes. X er antall EU- tilhengere i utvalget.

(a) Finn P(X=1)

(b) Beregn $E(X)$ og $Var(X)$.

A.8 Anta vi trekker 5 kort tilfeldig fra en kortstokk på 52 kort. Finn
(a) P(to ess)
(b) P(ingen ess)
Det oppgis at $\binom{52}{5} = 2.598.960$.

A.9 Bevis egenskapene (E1), (E2) til $E(X)$ og (V1)-(V3) til $Var(X)$ i kapittel 2.2.

Appendiks B

Variansestimering

Horvitz- Thompson estimatoren er gitt ved

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

$\hat{t}_{HT}$ er forventningsrett med

$$Var(\hat{t}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + 2\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j$$

$$= \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$ hvis størrelsen på *s* er bestemt på forhånd.

Her er π_{ij} = P(enhetene *i* og *j* er trukket ut til utvalget) = P(I_i = 1 og I_j = 1) . Anta utvalgsplanen har fast størrelse på *s.* Vi skal vise at

$$\hat{Var}(\hat{t}_{HT}) = \frac{1}{2} \sum_{i \in s} \sum_{\substack{j \in s \\ j \neq i}} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 .$$

er forventningsrett. Vi kan uttrykke denne variansestimatoren på følgende form:

$$\hat{Var}(\hat{t}_{HT}) = \frac{1}{2} \sum_{i} \sum_{j \neq i} I_i I_j \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

Dermed:

$$E(\hat{Var}(\hat{t}_{HT})) = \frac{1}{2} \sum_{i} \sum_{j \neq i} E(I_i I_j) \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

$$= \frac{1}{2} \sum_{i} \sum_{j \neq i} P(I_i = 1 \cap I_j = 1) \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

$$= \frac{1}{2} \sum_{i} \sum_{j \neq i} \pi_{ij} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

$$= \frac{1}{2} \sum_{i} \sum_{j \neq i} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = Var(\hat{t}_{HT})$$

Vi ser at forventningsretthet for variansestimatoren krever at $\pi_{ij} > 0$ for alle $i \neq j$.

Anta nå enkelt tilfeldig utvalg . Da er Horvitz-Thompson estimatoren lik ekspansjonsestimatoren

$$\hat{t}_{HT} = \hat{t}_e = N\bar{y}_s \ ,$$

hvor $\bar{y}_s = \sum_{i \in s} y_i / n$ er gjennomsnittet i utvalget *s*.

Fra teorien for Horvitz-Thompson estimatoren vet vi at $\hat{t}_e$ er forventningsrett, $E(\hat{t}_e) = t$. Vi kan finne $Var(\hat{t}_e)$ ved å bruke den generelle formelen ,

$$Var(\hat{t}_{HT}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 .$$

Alle π_{ij} er like og π_{ij} = P(Enhetene *i* og *j* trekkes ut) = P($I_i I_j = 1$) = $E(I_i I_j)$.

Dessuten:

$$\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} I_i I_j = \sum_{i=1}^{N} I_i \sum_{\substack{j=1 \\ j\neq i}}^{N} I_j = \sum_{i=1}^{N} I_i (n - I_i) = n\sum_{i=1}^{N} I_i - \sum_{i=1}^{N} I_i$$

$$= n \cdot n - n = n(n - 1).$$

Dette medfører at

$$E\left(\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} I_i I_j\right) = \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} E(I_i I_j) = n(n-1) .$$

Dermed :

$$\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} \pi_{ij} = n(n-1) .$$

Det er ialt $N(N - 1)$ ledd i summen , slik at hver enkelt π_{ij} er lik :

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} .$$

Populasjonens varians er gitt ved

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 \ \ , \text{ med } \mu = t/N.$$

Vi kan nå utlede $Var(\hat{t}_e)$:

$$Var(\hat{t}_e) = \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}\left(\frac{n^2}{N^2} - \frac{n(n-1)}{N(N-1)}\right)\left(\frac{Ny_i}{n} - \frac{Ny_j}{n}\right)^2$$

$$= \left(1 - \frac{N(n-1)}{(N-1)n}\right)\frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N}(y_i - y_j)^2$$

$$= \frac{N-n}{(N-1)n}\cdot\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - \mu - (y_j - \mu))^2$$

$$= \frac{N-n}{(N-1)n}\cdot\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\{(y_i - \mu)^2 + (y_j - \mu)^2 - 2(y_i - \mu)(y_j - \mu)\}$$

$$= \frac{N-n}{(N-1)n}\cdot\frac{1}{2}\{\sum_{i=1}^{N}(y_i - \mu)^2 N + N\sum_{j=1}^{N}(y_j - \mu)^2 - 2\sum_{i=1}^{N}(y_i - \mu)\sum_{j=1}^{N}(y_j - \mu)$$

Det siste leddet blir 0, og dermed :

$$Var(\hat{t}_e) = \frac{N-n}{(N-1)n}\cdot\frac{1}{2}\cdot N\cdot 2N\sigma^2$$

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

$Var(\hat{t}_e)$ er estimert ved

$$\hat{Var}(\hat{t}_e) = \frac{1}{2}\sum_{i\in s}\sum_{\substack{j\in s \\ j\neq i}}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

$$= \frac{N(N-1)}{n(n-1)}\cdot\frac{N-n}{(N-1)n}\cdot\frac{1}{2}\sum_{i\in s}\sum_{\substack{j\in s \\ j\neq i}}(y_i - y_j)^2$$

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

helt tilsvarende som i $Var(\hat{t}_e)$. La

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

Den forventningsrette estimatoren kan uttrykkes som :

$$\hat{Var}(\hat{t}_e) = N^2\frac{\hat{\sigma}^2}{n}\cdot\frac{N-n}{N-1}.$$

Appendiks C

Sammenligning : Rate-estimatoren og ekspansjonsestimatoren

Vi har kjent tilleggsinformasjon for hele populasjonen , $\boldsymbol{x} = (x_1 , x_2 , \ldots , x_N)$. Alle x_i er positive. Rate-estimatoren ble definert i kapittel 2.2, med $X_\circ$ lik totalsummen av alle x-verdiene :

$$\hat{t}_R = X_\circ \frac{\sum_{i\in s} y_i}{\sum_{i\in s} x_i} = \frac{X_\circ}{N\overline{x}_s}\hat{t}_e \quad .$$

Vi skal nå sammenligne rate-estimatoren og ekspansjonsestimatoren. La $f = n/N$, $R = t / X_\circ$, $\sigma_y^2 = \sigma^2$. Populasjonsvariansen for $\boldsymbol{x}$ betegnes med σ_x^2 , dvs. $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$. Populasjonens kovarians mellom x_i - og y_i - verdiene er definert ved:

$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$. Her er $\mu_x = X_\circ / N$ og $\mu_y = t / N$.

Følgende resultater holder for rate-estimatoren :

$$E(\hat{t}_R) \approx t + X_\circ \frac{1-f}{\bar{x}^2}\cdot\frac{R\sigma_x{}^2 - \sigma_{xy}}{n} \approx t \quad \text{for store } n$$

$$Var(\hat{t}_R) \approx N^2 \cdot \frac{1-f}{n}(\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{xy}).$$

Fra tidligere ,

$$\mathrm{Var}(\hat{t}_e) = N^2 \frac{\sigma_y{}^2}{n}\frac{N-n}{(N-1)} \approx N^2 \frac{\sigma_y{}^2}{n}(1-f).$$

Populasjonens korrelasjonskoeffisient mellom x og y er gitt ved

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad .$$

Korrelasjonskoeffisienten måler graden av lineær avhengighet mellom x- og y-verdiene i populasjonen og tar verdier i intervallet [-1,1] . Vi ser at :

$$\frac{Var(\hat{t}_R)}{Var(\hat{t}_e)} = \frac{\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{xy}}{\sigma_y^2}$$

og

$$Var(\hat{t}_R) < Var(\hat{t}_e) \Leftrightarrow \sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{xy} < \sigma_y^2$$

$$\Leftrightarrow 2R\sigma_x^2 < \sigma_{xy} \Leftrightarrow \rho_{xy} > \frac{1}{2}\cdot R\frac{\sigma_x}{\sigma_y} = \frac{1}{2}\cdot\frac{\sigma_x / \mu_x}{\sigma_y / \mu_y}$$

Et alternativt uttrykk for $Var(\hat{t}_R)$ (se oppgave C.1) er :

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

Dermed har vi også alternativt at

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

$$\Leftrightarrow \sum_{i=1}^{N} (y_i - \mu_y \frac{x_i}{\mu_x})^2 < \sum_{i=1}^{N} (y_i - \mu_y)^2 .$$

$Var(\hat{t}_R)$ kan estimeres ved :

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

Variansestimatoren til $Var(\hat{t}_e)$ er , med $\hat{\sigma}^2 = \frac{N-1}{N} \cdot \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y}_s)^2$, lik :

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

I en gitt situasjon kan vi derfor avgjøre om vi skal benytte $\hat{t}_R$ eller $\hat{t}_e$ ved å sammenligne

$$\left(\frac{\mu_x}{\bar{x}_s} \right)^2 \sum_{i \in s} (y_i - \hat{R}x_i)^2 \quad \text{og} \quad \sum_{i \in s} (y_i - \bar{y}_s)^2 .$$

Et 95% konfidensintervall for t basert på rate-estimatoren er gitt ved:

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

Oppgave til Appendiks C

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

Løsninger til oppgaver

Kapittel 1

1.1 Enkelt tilfeldig utvalg på 133 personer, med X = antall yrkesaktive. Med n uavhengige observasjoner av X, $X_1,\ldots X_n$, så er $\overline{X} = \sum_{i=1}^{n} X_i / n$ en forventningsrett estimator for $E(X)$ og $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ forventningsrett for $Var(X)$.

Vi har 5 utvalg med X- verdier lik 71, 81, 77, 78, 83. $E(X)$ estimeres da ved $\bar{x}$ = 390/5 = 78. $Var(X)$ estimeres ved $s^2 = 84/4$, og estimatet av standardavviket blir $s = \sqrt{21} = 4{,}58$.

(a) X er hypergeometrisk fordelt ; N = 1330, n = 133, M = 797, og $p = M/N$ = 797/1330 = 0,599.
Teoretiske verdier : $E(X) = np = 79{,}7$; $Var(X) = np(1-p)\frac{N-n}{N-1} = 28{,}77$; $Sd(X) = \sqrt{28{,}77} = 5{,}36$

(b) $E(X)$ = 79,7 mot estimat 78,0. $Sd(X)$ = 5,36 mot estimat 4,58. God overensstemmelse, spesielt tatt i betraktning at vi har kun 5 observasjoner av X.

1.2

	Hypergeometrisk fordeling	Binomisk fordeling
Forventning	79,7	79,7
Varians	28,77	31,94
Standardavvik	5,36	5,65

Forventning er like, mens binomisk varians er større. De praktiske følger er at vi kan tilnærme til binomisk med ca. 5% større standardavvik.

1.3* $P(X = y_i) = 1/N$ for i = 1, …, N.

(a) $E(X) = \sum_i y_i P(X = y_i) = \sum_i y_i \frac{1}{N} = \frac{1}{N}\sum_i y_i = \mu$

$Var(X) = E(X-\mu)^2 = \sum_i (y_i - \mu)^2 \frac{1}{N} = \sigma^2$

(b) $\sigma^2 = \frac{1}{N}\sum_i (y_i - p)^2 = \frac{1}{N}\left\{ \sum_{\{i:y_i=1\}} (1-p)^2 + \sum_{\{i:y_i=0\}} (0-p)^2 \right\}$

$= \frac{1}{N}\left\{ Np(1-p)^2 + N(1-p)p^2 \right\} = p(1-p)^2 + (1-p)p^2 = p(1-p)\{1-p+p\} = p(1-p)$

1.4 95% konfidensintervall for andelen p i befolkningen : $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

n = *400*, $\hat{p} = \frac{58}{400} = 0{,}145$ gir konfidensintervallet : $0{,}145 \pm 2 \cdot 0{,}0176 = 0{,}11$ til 0,18.

Dvs. med 95 % sikkerhet anslås andelen til å være mellom 11% og 18 %.

Kapittel 2

2.1* $E(\hat{t}_{HT}) = \sum_{k=1}^{N} \frac{y_k}{\pi_k} E(I_k) = \sum_{k=1}^{N} \frac{y_k}{\pi_k} \pi_k = \sum_{k=1}^{N} y_k = t$

2.2 (a) Utvalgsplan : $p(\{1,4\}) = 0,5$, $p(\{2,4\}) = 0,3$, $p(\{3,4\}) = 0,2$, $p(s) = 0$ ellers. Trekksannsynlighetene: $\pi_1 = 0,5$ $\pi_2 = 0,3$ $\pi_3 = 0,2$ $\pi_4 = 1,0$.

(b) Mulige verdier av $\hat{t}_{HT}$ med tilhørende sannsynligheter :

$s = \{1,4\}$: $\hat{t}_{HT}$ = 1200 med sannsynlighet 0,5

$s = \{2,4\}$: $\hat{t}_{HT}$ = 1666,67 med sannsynlighet 0,3

$s = \{3,4\}$: $\hat{t}_{HT}$ = 2500 med sannsynlighet 0,2

Dette gir : $E(\hat{t}_{HT}) = 1200 \cdot 0,5 + 1666,67 \cdot 0,3 + 2500 \cdot 0,2 = 1600 = t$

og $Var(\hat{t}_{HT}) = (1200 - 1600)^2 \cdot 0,5 + (1666,67 - 1600)^2 \cdot 0,3 + (2500 - 1600)^2 \cdot 0,2 = 243.333,33$

(c) $SE(\hat{t}_{HT}) = \sqrt{243.333,33} = 493,3$, mens i opprinnelig utvalgsplan $SE(\hat{t}_{HT}) = 57,74$. Velger opprinnelig utvalgsplan.

2.3 (a) Mulige verdier av $\hat{t}_e$, hver med sannsynlighet 1/6:

s	{1,2}	{1,3}	{1,4}	{2,3}	{2,4}	{3,4}
$\hat{t}_e$	600	800	2200	1000	2400	2600

$$E(\hat{t}_e) = \frac{1}{6}(600 + 800 + 2200 + 1000 + 2400 + 2600) = \frac{9600}{6} = 1600 = t$$

$$Var(\hat{t}_e) = E(\hat{t}_e - t)^2$$

$$= \frac{1}{6}\left\{(600 - 1600)^2 + (800 - 1600)^2 + (2200 - 1600)^2 + (1000 - 1600)^2 + (2400 - 1600)^2 + (2600 - 1600)^2\right\}$$

$= 666.666,67$ og $SE(\hat{t}_e) = \sqrt{Var(\hat{t}_e)} = 816,50.$

(b) $X_\circ = 300$, og $\hat{t}_R = X_\circ \frac{\sum_s y_i}{\sum_s x_i}$. Mulige verdier av $\hat{t}_R$, hver med sannsynlighet 1/6:

s	{1,2}	{1,3}	{1,4}	{2,3}	{2,4}	{3,4}
$\hat{t}_R$	1800	1714,29	1500	1875	1565,22	1560

$$E(\hat{t}_R) = \frac{1}{6}(1800 + 1714,29 + 1500 + 1875 + 1565,22 + 1560) = \frac{10.014,51}{6} = 1669,1$$

$$Var(\hat{t}_R) = \frac{1}{6}\left\{(1800 - 1669,1)^2 + (1714,29 - 1669,1)^2 + (1500 - 1669,1)^2 + (1875 - 1669,1)^2\right\}$$

$$+ \frac{1}{6}\left\{(1565,22 - 1669,1)^2 + (1560 - 1669,1)^2\right\} = 18.810,07$$

og $SE(\hat{t}_R) = \sqrt{18.810,07} = 137,15.$

2.4 (a) Mulige verdier av $\hat{t}_e$ er 3,6 og 12 , hver med sannsynlighet 1/3. Dermed:

$$E(\hat{t}_e) = \frac{1}{3}(3+6+12) = 7; \text{ dvs. } \hat{t}_e \text{ er forventningsrett}$$

$$Var(\hat{t}_e) = \frac{1}{3}\left\{(3-7)^2 + (6-7)^2 + (12-7)^2\right\} = 14 \text{ og } SE(\hat{t}_e) = \sqrt{14} = 3{,}74.$$

(b) Den opprinnelige utvalgsplanen medførte at $E(\hat{t}_e) = 6{,}15$ og $SE(\hat{t}_e) = 1{,}49$. Totalt sett så er den opprinnelige planen best.

2.5 (a) π_1 = P(trekke husholdning 1 denførste gangen) + P(trekke husholdning 1 den andre gangen)

$= \frac{3}{5} + \frac{2\cdot 3}{5\cdot 4} = \frac{12+6}{20} = \frac{18}{20} = \frac{9}{10}$.

$\pi_2 = \pi_3$, og $\pi_1 + \pi_2 + \pi_3 = 2$. Det betyr at $\pi_2 = \pi_3 = \frac{11}{20} = 0{,}55$.

(b) $\hat{t}_{HT} = \dfrac{400.000}{0{,}9} + \dfrac{200.000}{0{,}55} = 808.081$

$$\hat{t}_R = X_\circ \frac{\sum_s y_i}{\sum_s x_i} = 5\frac{600.000}{4} = 750.000 = t\,!$$

$$\hat{t}_e = 3 \cdot \frac{600.000}{2} = 900.000$$

(c) s = {1,2}: $\hat{t}_{HT} = 717.172$, $\hat{t}_R = 687.150$, $\hat{t}_e = 825.000$.

s = {2,3}: $\hat{t}_{HT} = 636.364$, $\hat{t}_R = 875.000$, $\hat{t}_e = 525.000$.

(d) p({1,2}) = P(trekke husholdning 1 første gang, husholdning 2 andre gang)

+ P(trekke husholdning 2 første gang, husholdning 1 andre gang)

$= \frac{3}{5}\cdot\frac{1}{2} + \frac{1}{5}\cdot\frac{3}{4} = \frac{6+3}{20} = \frac{9}{20}$.

Helt tilsvarende : p({1,3}) = 9/20. Siden summen av alle p(s) er lik 1: p({2,3}) = 2/20.

(e) Mulige verdier og tilhørende sannsynligheter for de tre estimatorene:

s	$\hat{t}_{HT}$	$\hat{t}_R$	$\hat{t}_e$	sannsynlighet
{1,2}	717.172	687.150	825.000	9/20
{1,3}	808.181	750.000	900.000	9/20
{2,3}	636.364	875.000	525.000	2/20

$$E(\hat{t}_{HT}) = 717.172\cdot\tfrac{9}{20} + 808.181\cdot\tfrac{9}{20} + 636.364\cdot\tfrac{2}{20} = 750.000 = t$$

$$E(\hat{t}_R) = 687.150\cdot\tfrac{9}{20} + 750.000\cdot\tfrac{9}{20} + 875.000\cdot\tfrac{2}{20} = 734.375$$

$$E(\hat{t}_e) = 825.000\cdot\tfrac{9}{20} + 900.000\cdot\tfrac{9}{20} + 525.000\cdot\tfrac{2}{20} = 828.750$$

$Var(\hat{t}_{HT}) = E(\hat{t}_{HT} - t)^2 = (717.172 - 750.000)^2 \cdot \frac{9}{20} + (808.181 - 750.000)^2 \cdot \frac{9}{20} + (636.364 - 750.000)^2 \cdot \frac{2}{20}$

$= 3.299.531.905$, $SE(\hat{t}_{HT}) = \sqrt{Var(\hat{t}_{HT})} = 57.441$.

$Var(\hat{t}_R) = 3.340.055.125$ og $SE(\hat{t}_R) = 57.793$.

$Var(\hat{t}_e) = 17.718.750.000$ og $SE(\hat{t}_R) = 133.112$.

Konklusjon: Foretrekker $\hat{t}_{HT}$.

Kapittel 3

3.1 (a) $S\hat{E}(\hat{t}_e) = \sqrt{16 \cdot \frac{0,5}{2} \cdot \{(100 - 200)^2 + (300 - 200)^2\}} = \sqrt{80.000} = 282,84.$

$\hat{R} = 400 / 70 = 40 / 7$. $\hat{V}ar(\hat{t}_R) = \left(\frac{75}{35}\right)^2 \cdot 16 \cdot \frac{0,5}{2}\left\{(100 - \frac{40}{7} \cdot 20)^2 + (300 - \frac{40}{7} \cdot 50)^2\right\} = 7496,88$ og

$S\hat{E}(\hat{t}_R) = \sqrt{7496,88} = 86,58$. Velger rate-estimatoren.

(b) $\hat{t}_e = 800$, $\hat{t}_R = 300 \cdot \frac{400}{70} = 1714,29$. ($t = 1600$)

(c) 95% konfidensintervaller: (i) basert på $\hat{t}_e$: $800 \pm 2 \cdot 282,84 = 800 \pm 566 = (234\ ,\ 1366)$ inkluderer ikke sann *t.* (ii) Basert på $\hat{t}_R$: $1714,29 \pm 2 \cdot 86,58 = 1714 \pm 173 = (1541\ ,\ 1887)$ inkluderer sann *t*.

Kapittel 4

4.1 (a) $E(\bar{y}_s) = \sum verdi \cdot \frac{1}{15} = \frac{2725}{15} = \frac{1090}{6} = 181,7.$

$Var(\bar{y}_s) = \sum (verdi - \frac{1090}{6})^2 \cdot \frac{1}{15} = \frac{10083,33}{15} = 672,22.$

(b) Stratum 1= (1,2,3) , stratum 2 = (4,5,6). De 9 mulige utvalg (alle like sannsynlige) og tilhørende verdier av estimatoren $\bar{y}_s$:

s	{1,4}	{1,5}	{1,6}	{2,4}	{2,5}	{2,6}	{3,4}	{3,5}	{3,6}
$\bar{y}_s$	215	195	195	200	180	180	170	150	150

$E(\bar{y}_s) = \frac{1635}{9} = 181,7$. $Var(\bar{y}_s) = 3950 / 9 = 438,89$.

I begge utvalgsplaner er estimatoren forventningsrett. Standardfeilen er, imidlertid, mindre ved stratifisert utvalgsplan som dermed foretrekkes. $SE_{ETU}(\bar{y}_s) = \sqrt{672,22} = 25,93$, og $SE_{STRAT}(\bar{y}_s) = \sqrt{438,89} = 20,95.$

4.2* $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ og $\sigma_h^2 = \frac{1}{N_h}\sum_{i \in U_h}(y_i - \bar{Y}_h)^2$.

$\sigma^2 = \frac{1}{N}\left\{\sum_{U_1}(y_i - \bar{Y}_1 + \bar{Y}_1 - \bar{Y})^2 + \sum_{U_2}(y_i - \bar{Y}_2 + \bar{Y}_2 - \bar{Y})^2 + ... + \sum_{U_H}(y_i - \bar{Y}_H + \bar{Y}_H - \bar{Y})^2\right\}$

$$= \frac{1}{N}\left\{\sum_{U_1}(y_i - \overline{Y}_1)^2 + \sum_{U_2}(y_i - \overline{Y}_2)^2 + \ldots + \sum_{U_H}(y_i - \overline{Y}_H)^2\right\}$$
$$+ \frac{1}{N}\left\{N_1(\overline{Y}_1 - \overline{Y})^2 + N_2(\overline{Y}_2 - \overline{Y})^2 + \ldots + N_H(\overline{Y}_H - \overline{Y})^2\right\}$$
$$+ 2 \cdot \frac{1}{N}\sum_{h=1}^{H}\{(\overline{Y}_h - \overline{Y})\sum_{U_h}(y_i - \overline{Y}_h)\}$$

$\sum_{U_h}(y_i - \overline{Y}_h) = N_h\overline{Y}_h - N_h\overline{Y}_h = 0$ for alle *h*. Dermed:

$$\sigma^2 = \frac{1}{N}\left\{N_1\sigma_1^2 + N_2\sigma_2^2 + \ldots + N_H\sigma_H^2\right\} + \sum_h W_h(\overline{Y}_h - \overline{Y})^2 .$$

4.3 Stratum 1 (USA) , stratum 2 (Brasil), og stratum 3 (Mexico) har alle $\sigma_h^2 = 0$. Stratum 4 har populasjonsvarians $\sigma_4^2 = 98{,}188$. Dermed fås : $Var(\hat{t}_{st}) = N_4^2 \cdot \frac{\sigma_4^2}{n_4} \cdot \frac{N_4 - n_4}{N_4 - 1} = 5^2 \cdot 98{,}188 = 2454{,}7$ og $SE(\hat{t}_{st}) = \sqrt{2454{,}7} = 49{,}54$.

4.4 (a) $n/N = 0{,}1 \Rightarrow n_h / N_h = 0{,}1$. Det gir $n_1 = 6, n_2 = 92$ og $n_3 = 35$.

(b) $\hat{t}_{st} \approx N\overline{y}_s = 1330 \cdot \frac{78}{133} = 780$. Eksakt

$\hat{t}_{st} = N_1\overline{y}_{s_1} + N_2\overline{y}_{s_2} + N_3\overline{y}_{s_3} = 59 \cdot \frac{1}{2} + 916 \cdot \frac{70}{92} + 355 \cdot \frac{5}{35} = 777{,}2.$

$E(\hat{t}_{st}) = t = 797,$ og $Var(\hat{t}_{st}) = \sum_{h=1}^{3} N_h^2 \cdot \frac{\sigma_h^2}{n_h} \cdot \frac{N_h - n_h}{N_h - 1}$. Fra oppgave 1.3b: $\sigma_h^2 = p_h(1 - p_h)$, hvor p_h er andel yrkesaktive i stratum h.

$p_1 = 28/59 = 0{,}475$, $p_2 = 729/916 = 0{,}796$, $p_3 = 40/355 = 0{,}113$. Dette gir:

$Var(\hat{t}_{st}) = 1792{,}02$ og $SE(\hat{t}_{st}) = \sqrt{1792{,}02} = 42{,}33.$

(c) La *p* være andel yrkesaktive i hele populasjonen, *p* = 797/1330 = 0,599.

$Var_{ETU}(\hat{t}_e) = N^2 \cdot \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1} = 2876{,}76$ og $SE_{ETU}(\hat{t}_e) = \sqrt{2876{,}76} = 53{,}64.$ Siden $SE_{ETU}(\hat{t}_e) > SE(\hat{t}_{st})$ foretrekkes stratifisert utvalgsplan.

4.5 (a) p_1 = andel yrkesaktive i 16-24 år gruppen (= 138/197 = 0,70). p_2 = andel yrkesaktive over 24 år (= 659/1133) = 0,58). Vi har at $n = 133$ med $n_1 = 20$ og $n_2 = 113$ og enkelt tilfeldig utvalg fra hvert stratum. Estimatorer for p_1, p_2 : $\hat{p}_1$ = andel yrkesaktive i utvalget fra stratum 1, og $\hat{p}_2$ = andel yrkesaktive i utvalget fra stratum 2. $Var(\hat{p}_i = \frac{p_i(1-p_i)}{n_i} \cdot \frac{N_i - n_i}{N_i - 1}$ for *i* = 1,2 som gir at $Var(\hat{p}_1) = 0{,}00947$ og $Var(\hat{p}_2) = 0{,}00194$ slik at $SE(\hat{p}_1) = \sqrt{0{,}00947} = 0{,}0973$ og $SE(\hat{p}_2) = \sqrt{0{,}00194} = 0{,}0440.$

(b) Krav for publisering: $SE(\hat{p}_i) / \hat{p}_i \le 0{,}10$. $\hat{p}_1 = 0{,}65$ gir $SE(\hat{p}_1) / \hat{p}_1 = 0{,}150 = 15\%$ og er ikke pålitelig nok. $\hat{p}_2 = 0{,}60$ gir $SE(\hat{p}_2) / \hat{p}_2 = 0{,}073 = 7{,}3\%$ og er iorden.

(c) $n_1 = n_2 = 66$ medfører at $Var(\hat{p}_1) = 0{,}002127$ og $SE(\hat{p}_1) = 0{,}0461$, $SE(\hat{p}_1) / \hat{p}_1 = 0{,}071 = 7{,}1\%$ og estimatet kan publiseres. $Var(\hat{p}_2) = 0{,}003474$ og $SE(\hat{p}_2) = 0{,}0589$, $SE(\hat{p}_2) / \hat{p}_2 = 0{,}098 =$ 9,8% og det holder akkurat for publisering.

4.6 (a) $\hat{t}_e = N\bar{y}_s = 1330 \cdot \frac{59}{100} = 784{,}7 \approx 785.$

(b) $\hat{t}_{est} = N_1\bar{y}_{s_1} + N_2\bar{y}_{s_2} = 705 \cdot \frac{32}{45} + 625 \cdot \frac{27}{55} = 808{,}2 \approx 808.$

4.7 (a) Uten etterstratifisering , estimator er $\hat{t}_e = N\bar{y}_s$ for EU-tilhengere, EU-motstandere og tvilere. (i) EU- tilhengere : $\hat{t}_e = 1273 \cdot \frac{20}{150} = 170,\ 13{,}3\%$. (ii) EU-motstandere : $\hat{t}_e = 1273 \cdot \frac{105}{150} = 891,\ 70\%.$ (iii) Tvilere : $\hat{t}_e = 1273 \cdot \frac{25}{150} = 212,\ 16{,}7\%.$

Med etterstratifisering, 8 etterstrata, estimator er $\hat{t}_{est} = \sum_{h=1}^{8} N_h\bar{y}_{s_h}$. Vi har følgende verdier av n_i : 24,3,9,4,89,1,12 og N_i : 300,32,62,87,367,80,41,304. (i). EU- tilhengere : $\hat{t}_{est} = 241,\ 18{,}9\ \%.$ (ii) EU-motstandere : $\hat{t}_{est} = 599,\ 47{,}1\%.$ (iii) Tvilere : $\hat{t}_{est} = 433,\ 34\%.$

(b) Stoler mest på de etterstratifiserte estimatene.

(c) De estimerte andelene av EU-motstandere blir for *små* i strata utenfor SP, mens i SP-stratum omtrent det samme. Resultatet blir at vårt estimat av EU-motstandere blir for *lavt*.

(d) Det kan være et problem hvis det er forskjell mellom kvinner og menn. Kan løses ved å etterstratifisere med hensyn på kjønn.

4.8 (a) La t være antall gifte i populasjonen. Etterstratifisert estimat: $\hat{t}_{est} = 800 \cdot \frac{40}{60} + 700 \cdot \frac{10}{90} = 611,$ mellom anslagene 500 og 800.

(b) Etterstratifisert estimat er mest pålitelig: Retter opp skjevhet i utvalget, hvor for mange er trukket blant de registrerte ikke-gifte slik at $\hat{t}_e$ blir for liten.

Kapittel 5

5.1 Krav: $\pi_k = \pi_{Ii}\pi_{k|i} = 0{,}1$. $\pi_{Ii} = \frac{N_i}{N_1+N_2}$ for $i = 1{,}2$, og $\pi_{Ii} = \frac{N_i}{N_3+N_4}$ for $i = 3{,}4.$ $\pi_{k|i} = n_i \ / \ N_i$ slik at, for i = 1,2: $\frac{N_i}{N_1+N_2} \cdot \frac{n_i}{N_i} = \frac{1}{10} \Rightarrow n_i = \frac{N_1+N_2}{10} = 3$. For $i = 3{,}4$: $\frac{N_i}{N_3+N_4} \cdot \frac{n_i}{N_i} = \frac{1}{10} \Rightarrow n_i = \frac{N_3+N_4}{10} = 22.$

Kapittel 6

6.1 (a) Nord-Aurdal: $\pi_{I1} = 6601 / 26021 = 0{,}254.$ Nordre Land/Etnedal: $\pi_{I2} = 8474 / 26021 = 0{,}326.$ Sør-Aurdal: $\pi_{I3} = 3550 / 26021 = 0{,}136.$ Vestre Slidre/Vang: $\pi_{I4} = 4296 / 26021 = 0{,}165.$ Øystre Slidre: $\pi_{I5} = 3100 / 26021 = 0{,}119.$

(b) For PU Nordre Land/Etnedal : $N =$ 4 324 815, $N^{99} =$ 3 511 082, $N_h = 26\ 021,$ $N_h^{99} = 20\ 518,$ $N_{h,i} =$ 8474 og $N_{h,i}^{99} = 6865.$ (6.1) gir $n_i = \frac{n}{N} N_h = 60{,}17 = 60.$ (6.2) gir eksakt $n_i = 60{,}17 \cdot \frac{6865/8474}{3511082/4324815} = 60{,}17 \cdot \frac{0{,}8101}{0{,}8118} = 60{,}04 = 60.$ (6.3) gir $n_i = \frac{n}{N^{99}} N_h^{99} = 58{,}44 = 58.$

(c) (6.1) og (6.2) $\Rightarrow n_i = 60$. Trekksannsynlighet $\pi_k = \frac{N_{h,i}}{N_h} \cdot \frac{60}{N_{h,i}^{99}} = 0{,}002846$. (6.3) gir $n_i = 58$ og trekksannsynlighet $\pi_k = \frac{N_{h,i}}{N_h} \cdot \frac{58}{N_{h,i}^{99}} = 0{,}002751$. Sammenlignet med selvveiende $\pi_k = \frac{10000}{3511082} = 0{,}002848$.

(d) $\pi_k = \frac{N_{h,i}}{N_h} \cdot \frac{60}{N_{h,i}^{99}} = \frac{3100}{26021} \cdot \frac{60}{2444} = 0{,}002925.$

6.2 (a) Tettbygde PU : $N_{h,i}$ øker, π_{Ii} er fast, og $n_i = \frac{n}{N} N_h$ er fast for hele stratum. Dermed $\pi_k = \pi_{Ii} \cdot \frac{n_i}{N_{h,i}}$ avtar og vi har ikke lenger et selvveiende utvalg, $\pi_k < n / N$ for tettbygde PU og $\pi_k > n / N$ for de landlige PU. Estimatoren blir dermed skjev.

(b) Ansettelse av flere intervjuere, eventuelt økte reisekostnader.

Kapittel 7

7.1* (a) Antall i responsstratum $N_R = q_R N$. Antallet i responsdel i stratum *h*: $q_h N_h$. Dermed:

$$\overline{Y}_R = \frac{1}{q_R N} \cdot \sum_{U_R} y_i = \frac{1}{q_R N} \cdot \sum_{h=1}^{H} q_h N_h \overline{Y}_{rh} = \frac{1}{q_R} \cdot \sum_{h=1}^{H} q_h \frac{N_h}{N} \overline{Y}_{rh} = \frac{1}{q_R} \cdot \sum_{h=1}^{H} q_h W_h \overline{Y}_{rh}.$$

Tilsvarende for $\overline{Y}_F$: Antall i U_F = *N*- N_R = $(1 - q_R)N$ og antallet i frafallsdelen i stratum *h* = N_h - $q_h N_h$ = $(1\text{-}q_h)N_h$. Dette gir at

$$\overline{Y}_F = \frac{1}{(1-q_R)N} \cdot \sum_{U_F} y_i = \frac{1}{(1-q_R)N} \cdot \sum_{h=1}^{H} (1-q_h) N_h \overline{Y}_{fh} = \frac{1}{1-q_R} \cdot \sum_{h=1}^{H} (1-q_h) W_h \overline{Y}_{fh}.$$

(b) $(1-q_R)(\overline{Y}_R - \overline{Y}_F) = \frac{1-q_R}{q_R} \sum_h q_h W_h \overline{Y}_{rh} - \sum_h (1-q_h) W_h \overline{Y}_{fh} =$

$$= \frac{1}{q_R} \sum_h (q_h - q_R) W_h \overline{Y}_{rh} + \sum_h W_h \overline{Y}_{rh} - \sum_h q_h W_h \overline{Y}_{rh} - \sum_h (1-q_h) W_h \overline{Y}_{fh}$$

$$= \frac{1}{q_R} \sum_h (q_h - q_R) W_h \overline{Y}_{rh} + \sum_h (1-q_h) W_h (\overline{Y}_{rh} - \overline{Y}_{fh}).$$

7.2 Ekspansjonsestimatene, for H_2: $\frac{1}{2} \cdot 4131874 \cdot \frac{267}{1111} = 496.494$, for H_3: $\frac{1}{3} \cdot 4131874 \cdot \frac{229}{1111} = 283.888$, for H_4 : $\frac{1}{4} \cdot 4131874 \cdot \frac{301}{1111} = 279.859$, for $H_{\geq 5}$: $\frac{1}{5{,}25} \cdot 4131874 \cdot \frac{209}{1111} = 148.054$.

Etterstratifiseringsestimatene:

For H_2: $\frac{1}{2}\left[793869 \cdot \frac{48}{162} + 816880 \cdot \frac{177}{230} + 784581 \cdot \frac{25}{212} + 1066016 \cdot \frac{13}{300} + 670528 \cdot \frac{4}{207}\right] = 507.768$

For H_3: $\frac{1}{3}\left[793869 \cdot \frac{20}{162} + 816880 \cdot \frac{37}{230} + 784581 \cdot \frac{131}{212} + 1066016 \cdot \frac{37}{300} + 670528 \cdot \frac{4}{207}\right] = 286.2221$

For H_4: $\frac{1}{4}\left[793869 \cdot \frac{9}{162} + 816880 \cdot \frac{4}{230} + 784581 \cdot \frac{40}{212} + 1066016 \cdot \frac{231}{300} + 670528 \cdot \frac{17}{207}\right] = 270.561$

For $H_{\geq 5}$: $\frac{1}{5{,}25}\left[793869 \cdot \frac{2}{162} + 816880 \cdot \frac{3}{230} + 784581 \cdot \frac{6}{212} + 1066016 \cdot \frac{17}{300} + 670528 \cdot \frac{181}{207}\right] = 131.310$.

7.3 (a) $y_i^* = \overline{y}_r \Rightarrow \overline{y}_s^* = \frac{1}{n}\left(\sum_{i \in s_r} y_i + \sum_{i \in s - s_r} y_i^*\right) = \frac{1}{n}\left(n_r \overline{y}_r + (n - n_r)\overline{y}_r\right) = \overline{y}_r$, $n_r = n_1 + n_2$.

(b) $y_i^* = \overline{y}_2 \Rightarrow \overline{y}_s^* = \frac{1}{n}\left(\sum_{i \in s_r} y_i + (n - n_r)\overline{y}_2\right) = \frac{1}{n}\left(n_1 \overline{y}_1 + n_2 \overline{y}_2 + (n - n_1 - n_2)\overline{y}_2\right) = p\overline{y}_1 + (1-p)\overline{y}_2$.

7.4* Gitt data i s_r : $E(y_i^*) = \sum_{s_r} verdi \cdot sannsynlighet = \sum_{k \in s_r} y_k \cdot \frac{1}{n_r} = \bar{y}_r$. $Var(y_i^*) = E(y_i^* - \bar{y}_r)^2 =$ $\sum_{k \in s_r} (y_k - \bar{y}_r)^2 P(y_i^* = y_k) = \sum_{k \in s_r} (y_k - \bar{y}_r)^2 \cdot \frac{1}{n_r}$.

7.5 (a) De 1403 personene utgjør et enkelt tilfeldig utvalg. $\hat{p}$ = 1190/1403 = 0,848.

$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1}} = 0{,}00958$. 95% konfidensintervall: $\hat{p} \pm 2SE(\hat{p}) = 0{,}848 \pm 0{,}019 =$ (0,829 , 0,867).

(b) Estimat og konfidensintervall er meget skjeve. Frafallet er ikke tilfeldig. Det er større frafall blant de som ikke stemte.

(c) $\hat{p}_{est} = \hat{t}_{est} / N$, hvor $\hat{t}_{est} = \sum_{h=1}^{3} N_h \bar{y}_h$. De observerte stratumandelene: $\bar{y}_1 = \frac{1060}{1192} = 0{,}8893$, $\bar{y}_2 = \frac{57}{115} = 0{,}4957$, $\bar{y}_3 = \frac{73}{96} = 0{,}7604$. Det gir $\hat{t}_{est} = 2.667.953$ og $\hat{p}_{est} = \frac{2.667.953}{3.259.957} = 0{,}818$. Estimatet er fremdeles skjevt. Det tyder på at etterstratifiseringen ikke er tilstrekkelig, dvs. at svargruppen ikke er representativ for frafallsgruppen innen hvert stratum.

(d) Basisestimatoren er etterstratifisert estimator. De imputerte verdiene i frafallsgruppen i stratum *h* er alle lik den observerte stratumandel $\bar{y}_h$ eller ekvivalent : 88,93% i frafallsgruppen i stratum 1 gis verdi 1, 49,57% i frafallsgruppen i stratum 2 gis verdi 1, og 76,04% i frafallsgruppen i stratum 3 gis verdi 1.

(e) Etterstratifiseringsestimatoren er forventningsrett når svargruppen er representativ for frafallsgruppen, $\bar{Y}_{rh} = \bar{Y}_{fh}$ for alle *h*. Det holder ikke her. En mulig angrepsmåte er å anta at sannsynligheten for at en person i et gitt stratum svarer er avhengig av om personen stemte eller ikke.

7.6 (a) $\bar{y}_s^* = 241, \hat{\sigma}_*^2 = \frac{1}{9}\left\{ \sum_{i \in s_r} (y_i - \bar{y}_s^*)^2 + \sum_{i \in s - s_r} (y_i^* - \bar{y}_s^*)^2 \right\} = 17090 / 9 = 1898{,}89, \hat{\sigma}_* = 43{,}576.$

95% konfidensintervall: $\bar{y}_s^* \pm 1{,}96 \frac{\hat{\sigma}_*}{\sqrt{n}} = 241 \pm 27{,}0 = (214{,}268)$.

(b) $\bar{y}_s^* = 253, \hat{\sigma}_*^2 = 17810 / 9 = 1978{,}89, \hat{\sigma}_* = 44{,}485.$ 95% KI: $253 \pm 27{,}6 = (225{,}4 \ , \ 280{,}6)$.

(c) $\bar{y}_r = 1540 / 6 = 256{,}67, \hat{\sigma}_r^2 = \frac{1}{5} \sum_{i \in s_r} (y_i - \bar{y}_r)^2 = 9933{,}34 / 5 = 1986{,}67, \hat{\sigma}_r = 44{,}572.$ 95% KI : $\bar{y}_r \pm 1{,}96 \frac{\hat{\sigma}_r}{\sqrt{6}} = 256{,}67 \pm 35{,}67 = (221 \ , \ 292{,}3).$

(d) Rubin's kombinasjon: $B_* = (241 - 247)^2 + (253 - 247)^2 = 72$. $\bar{\sigma}_*^2 = \frac{1898{,}89 + 1978{,}89}{2} = 1938{,}89,$ $V_* = \frac{1938{,}89}{10} + \frac{3}{2} \cdot 72 = 301{,}89.$ 95% KI: $247 \pm 1{,}96\sqrt{301{,}89} = 247 \pm 34{,}1 = (212{,}9 \ , \ 281{,}1).$ Modifisert kombinasjon bruker $V_*^{'} = 193{,}889 + (\frac{1}{0{,}6} + \frac{1}{2})72 = 349{,}89.$ Det gir 95% KI :

$247 \pm 1{,}96\sqrt{349{,}89} = 247 \pm 36{,}7 = (210{,}3 \ , \ 283{,}7)$. Vi ser at konfidensintervallene basert på multippel imputering har samme presisjon som intervallet basert på svarutvalget, som vi vet er gyldig siden frafallet er rent tilfeldig. Intervallene i (a) og (b) er for korte.

Appendix A

A.1 P(A) = m/16. {0 riktige svar} = {GGGG} og m = 1, {1 riktig svar} = {RGGG, GRGG, GGRG, GGGR} og m = 4, {2 riktige svar} = {RRGG, RGRG, RGGR, GRRG, GRGR, GGRR} og m = 6, {3 riktige svar} = {RRRG, RRGR, RGRR, GRRR} og m = 4, {4 riktige svar} ={RRRR} og m = 1.

A.2 Utfallsrom $\boldsymbol{S}$ består av 10 mulige utvalg, alle like sannsynlige, $\boldsymbol{S}$ = {(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)}.

(a) P(A) = 6/10 = 0,6. P(B) = 6/10 = 0,6.

(b) $A \cup B$: 1 eller 2 eller begge er med i utvalget. $A \cap B$: 1 og 2 er begge i utvalget. A^c: 1 er ikke med i utvalget. B^c: 2 er ikke med i utvalget.

(c) $A \cup B = \boldsymbol{S}$ - (3,4,5). $A \cap B$ = {(1,2,3), (1,2,4), (1,2,5)}. A^c = {(2,3,4), (2,3,5), (2,4,5), (3,4,5)}. B^c = {(1,3,4), (1,3,5), (1,4,5), (3,4,5)}.

(d) $P(A \cap B) = 3/10 = 0,3$. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0,6 + 0,6 - 0,3 = 0,9$. $P(A^c) = 1 - P(A) = 0,4 = P(B^c)$.

(e) $m(A \cup B) = 9$, $m(A \cap B) = 3$, $m(A^c) = m(B^c) = 4$.

A.3 (a) X = 0 er umulig, P(X = 0) = 0.

(b) (X = 1) = {(1,2,5), (2,3,5), (2,4,5)} og P(X = 1) = 3/10 = 0,3.

(c) (X = 3) = {(1,3,4) og P(X = 3) =1/10 = 0,1.

(d) P(X = 1 eller 3) = P(X = 1) + P(X = 3) = 0,4.

A.4 (a) 36 = 6·6.

(b) {Sum = 7} = {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)} (rød terning først). P(sum = 7) = 6/36 = 1/6.

(c) A = {sum er 3 eller mer}. A^c = {sum = 2} = {(1,1)} $\Rightarrow$ $P(A^c) = 1/36$ og $P(A) = 1 - P(A^c) = 35/36$.

A.5 (a) 0,2/0,3 = 2/3.

(b) 0,2/0,7 = 2/7.

(c) 0,6.

A.6 Trekker tilfeldig en familie fra populasjonen med 2 barn og minst en gutt. Utfallsrommet $\boldsymbol{S}$ = {GG, GJ, JG}, yngste barn først, alle tilnærmet like sannsynlige. P(to gutter) = P({GG}) = 1/3. Dvs. andelen er 1/3.

Kan også betraktes som betinget sannsynlighet, P(to gutter|minst en gutt), ved å trekke tilfeldig en familie med 2 barn, med utfallsrom $\boldsymbol{S}$ = {GG, GJ, JG, JJ}. P(to gutter|minst en gutt) = P(to gutter)/P(minst en gutt) = $\frac{1/4}{3/4} = 1/3$.

A.7 X er hypergeometrisk fordelt med $N = 7$, $M = 3$ og $n = 2$. $P(X = x) = \dfrac{\binom{3}{x}\binom{4}{2-x}}{\binom{7}{2}} = \dfrac{\binom{3}{x}\binom{4}{2-x}}{21}$.

(a) $P(X = 1) = \dfrac{\binom{3}{1}\binom{4}{1}}{21} = \frac{12}{21} = 0{,}57.$

(b) $E(X) = np = 2 \cdot \frac{3}{7} = 6/7 = 0{,}857$. $Var(X) = np(1-p)\frac{N-n}{N-1} = \frac{6}{7} \cdot \frac{4}{7} \cdot \frac{5}{6} = \frac{20}{49} = 0{,}408.$

A.8 La X være antall ess blant de 5 uttrukne kortene. X er hypergeometrisk fordelt med $N = 52$, $M = 4$ og $n = 5$. $P(X = x) = \dfrac{\binom{4}{x}\binom{48}{5-x}}{\binom{52}{5}} = \dfrac{\binom{4}{x}\binom{48}{5-x}}{2.598.960}$.

(a) $P(X = 2) = \dfrac{\binom{4}{2}\binom{48}{3}}{2.598.960} = \frac{6 \cdot 17296}{2.598.960} = 0{,}040.$

(b) $P(X = 0) = \dfrac{\binom{4}{0}\binom{48}{5}}{2.598.960} = \frac{1.712.304}{2.598.960} = 0{,}659.$

A.9 (E1) $E(a) = a \cdot P(X = a) = a \cdot 1 = a.$

(E2) $E(a + bX) = \sum_i (a + bx_i) P(X = x_i) = \sum_i aP(X = x_i) + \sum_i bx_i P(X = x_i) = a\sum_i P(X = x_i)$ $+ b\sum_i x_i P(X = x_i) = a + bE(X).$

(V1) $\mu = a, X - \mu = 0 \Rightarrow E(X - \mu)^2 = 0,$ fra $(E1)$.

(V2) $Var(a + bX) = E\{(a + bX) - (a + b\mu)\}^2 = E(bX - b\mu)^2 = b^2 E(X - \mu)^2 = b^2 Var(X).$

(V3) $Var(X) = E(X - \mu)^2 = E\{X^2 - 2\mu X + \mu^2\} = E(X^2) - E(2\mu X) + E(\mu^2) =$
$E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2$.

Appendiks C

C.1 Må vise at $\sum_{i=1}^N (y_i - Rx_i)^2 = N(\sigma_y^2 + R^2\sigma_x^2 - 2R\sigma_{xy})$, $R = \mu_y / \mu_x$.

$\sum_{i=1}^N (y_i - Rx_i)^2 = \sum_{i=1}^N (y_i - \mu_y - (Rx_i - \mu_y))^2$

$= \sum_{i=1}^N (y_i - \mu_y)^2 + \sum_{i=1}^N (Rx_i - \mu_y)^2 - 2\sum_{i=1}^N (y_i - \mu_y)(Rx_i - \mu_y)$

$= N\sigma_y^2 + R^2 \sum_{i=1}^N (x_i - \mu_x)^2 - 2R\sum_{i=1}^N (y_i - \mu_y)(x_i - \mu_x) = N\sigma_y^2 + R^2 N\sigma_x^2 - 2RN\sigma_{xy}.$

Litteratur

Cochran, W. (1977). *Sampling Techniques*. Wiley.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. Wiley.

Rubin, D.B (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley

Särndal, C-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Thomsen, I. (1977). Prinsipper og metoder for SSB's utvalgsundersøkelser. *Samfunnsøkonomiske studier 33*. Statistisk sentralbyrå.

Thomsen, I & Siring, E.(1983). On the causes and effects of nonresponse. Norwegian experiences. *Incomplete Data in Sample Surveys, vol. 3, Session 1*. Academic Press.

De sist utgitte publikasjonene i serien Notater

2000/22 B. Strøm: MSG-6 Utslippsmodellens ligningsstruktur: Teknisk dokumentasjon. 49s.

2000/23 T. Risberg. G. Rogdaberg og R.M. Rosvold: Sykepleiernes tilpasning i arbeidsmarkedet: En kort beskrivelse av teorier og dataregistre. 46s.

2000/24 A.S. Brørs, K. Dybendal, A.H. Foss og T. Jakobsen: Dokumentasjon av BESYS - befolkningsstatistikksystemet: Befolkningsendringer i 1998 og befolkningsbasen (BEBAS) 1. januar 2000. 43s.

2000/25 E. Høydahl: FoB2001: Kommunenes innspill om kommunehefter. 18s.

2000/26 T. Kalve og J. Sørøy: Revisjon av barnevernsdata. 30s.

2000/27 A. Skoglund: Publikasjoner fra forskningsvirksomheten 1991-1999. 72s.

2000/28 H. Hungnes: Omregning av KVARTS-relasjoner til MODAG-relasjoner. 12s.

2000/29 R.N. Johnsen: Undersøking om foreldrebetaling i barnehagar, januar 2000. 36s.

2000/30 O. Rognstad: Plan for landbruksstatistikken etter 1999. 23s.

2000/31 Ø. Kleven: Levekårsundersøkelsen i Longyearbyen 2000: Dokumentasjon og tabellrapport. 188s.

2000/32 E. Rønning: Omnibusundersøkelse - mars 2000: Dokumentasjonsrapport. 34s.

2000/33 J. Johansen og J. Lajord: FD-trygd. Dokumentasjonsrapport. Utdanning. 1992-1997. 119s.

2000/34 A.L. Brathaug, J. Holmøy og H. Tønseth: Årsrapport: Kontaktutvalget for helse- og sosialstatistikk 1999. 24s.

2000/35 N. Barrabés: Norsk standard for utdanningsgruppering: Høringsnotat. 110s.

2000/36 D. Roll-Hansen og C.M. Hansen: En evaluering av datainnsamling gjennom KOSTRA: Rapportering av data fra 1999. 94s.

2000/37 B.R. Joneid og Ø. Sivertstøl: FD - trygd: Dokumentasjonsrapport: Foreløpig uførestønad, 1992-1998. 30s.

2000/38 R.N. Johnsen: Kommunale gebyrer knyttet til bolig. Januar 2000. 27s.

2000/39 J-A.S. Lie: Revisjon av data til Pleie- og omsorgsstatistikken i 1997 og 1998. 83s.

2000/40 Y. Holm, A.H. Tangen og O.M. Tidemann: Forprosjektrapport om etablering av IMFs internasjonale investeringsposisjon (IIP) for Norge. 97s.

2000/41 K.O. Olsen: Forsikring i nasjonalregnskapet. 42s.

2000/42 J. Johansen og J. Lajord: FD - Trygd: Dokumentasjonsrapport: Arbeidssøkere. 1992-1998. 74s.

2000/43 H.V. Sæbø: Til statistikkens pris: Prispolitikk i statistikkbyråene med hovedvekt på elektronisk formidling. 9s.

2000/44 E. Rønning: Omnibusundersøkelse - mai/juni 2000. Dokumentasjonsrapport. 32s.

2000/45 A. Holmøy og M. Høstmark: Undersøkelse om omfanget av utgifter til helse- og sosialtjenester: Dokumentasjon og tabellrapport. 116s.

2000/46 Fagseminar om arealpolitikk og arealstatistikk i opptakten til et nytt årtusen. Seminarrapport 30. mars 2000. 167s.

2000/47 Publikasjoner fra forskningsvirksomheten 1991-1999: Revidert versjon. 82s.

2000/48 A.-K. H. Grorud: Bedrifts- og foretaksregisteret: Regler og rutiner for ajourhold. 121s.

2000/49 T. Hoel, B.R. Joneid og G.E. Wangen: Trekkbas: Brukerdokumentasjon. 35s.