EN INTRODUKTION TIL STATISTIK

Repræsentative undersøgelser

BIND 3A

Poul Thyregod

LYNGBY 1998

IMM

Trykt af IMM - DTU
Bogbinder Hans Meyer

Forord

Nærværende notesæt er udarbejdet som grundbog for undervisningen i faget Statistik 3 ved IMM.

Det har været hensigten at fortsætte den introduktion til statistisk teori, som er givet i bind 1 og bind 2 på en række punkter.

Således introduceres i kapitel 1 de selvstændige teorier for repræsentative undersøgelser, som hører med til statistikerens arbejdsværktøjer, og som ikke er behandlet i de tidligere bind. Formålet med at medtage disse teorier er dels at præentere de specifikke teknikker og værktøjer, som er knyttet til disse teorier, men i lige så høj grad at vise eksempler på brug af statistisk tankegang i forbindelse med planlægning af dataindsamling.

Endvidere suppleres den gennemgang af metoder til analyse af lineære modeller for middelværdien af normalt fordelte variable (den generelle lineære model), der er præsenteret i bind 2, til at omfatte nyere metoder til analyse af (lineære og ikke-lineære) modeller for middelværdistrukturen i de almindelige fordelinger, der blev introduceret i bind 1. Analysen af denne samling af modeller, der sædvanligvis kaldes generaliserede lineære modeller, er tæt knyttet til likelihood-teorien og teorien for eksponentielle familier af fordelinger, hvorfor der også er medtaget en kort introduktion til disse teorier.

Afsnittet om hierarkiske modeller for de almindelige endimensionale fordelinger sigter mod at illustrere en flertrinsvariation, der ofte opleves i praktisk statistisk arbejde og at give nogle simple, gennemskuelige værktøjer, der kan bruges i sådanne sammenhænge. Samtidig finder jeg, at disse modeller danner et godt udgangspunkt for at forstå begrebsapparatet omkring Bayesianske metoder.

Afsnittene om nytteteori og statistisk beslutningsteori er overført nogenlunde uændret fra en tidligere note (P.Thyregod: Statistisk Beslutningsteori, IMSOR 1978). Disse afsnit er medtaget for at give en introduktion til de såkaldte Bayes-metoder og en indsigt i virkemåden for disse metoder, herunder også empirisk Bayes-metoder. Hensigten med at medtage disse afsnit er ydermere at give indsigt i hvorledes den statistiske usikkerhed kombineres tilgodese den statistiske usikkerhed i forbindelse med, eksempelvis økonomiske overvejelser.

Dette bind har været mange år undervejs, og jeg er megen tak skyldig til de deltagere i kurset Statistik 3 ved IMM, som har været udsat for

mine forskellige forsøg på at strukturere det omfattende stof i en strøm af foreløbige noter.

Det er mit håb, at deres lidelser ikke har været forgæves, men at de har bidraget til en stadig forbedring af notesættet sådan at det kan bruges som lærebog og reference i en tid fremover.

Selv med dette lange tilløb er notesættet langt fra fuldkomment. Jeg modtager derfor gerne kommentarer og rettelsesforslag.

Lyngby januar 1998
Poul Thyregod

Indhold

Afsnit 1

Stikprøver fra endelige populationer, Repræsentative undersøgelser

fil: repr1.tex 1998-01-23

1.1 Grundlæggende begreber

1.1.0 Indledning

Sigtet med stikprøveundersøgelse af en forelagt population er at udnytte stikprøveresultatet til at beskrive den resterende ikke-undersøgte del af populationen. Ofte vil man opfatte en stikprøve som et miniaturebillede af populationen, og generalisere de fundne stikprøvekarakteristika til en beskrivelse af hele den undersøgte population. I sædvanligt sprogbrug vil man sige, at en stikprøve er repræsentativ for den population, den er udtaget fra, hvis den netop er et sådant miniaturebillede.

Det turde imidlertid være klart, at en vilkårligt udtaget del af en population ikke nødvendigvis er repræsentativ for populationen. Således vil forskellen i eksponering over for vejrligets påvirkninger medføre, at vandindholdet i det øverste lag korn i et læs med korn sædvanligvis adskiller sig fra vandindholdet i det indre af læsset. En stikprøve, der alene består af korn fra det øverste lag, vil således ikke være repræsentativ for læsset.

For at sikre repræsentativiten er det derfor nødvendigt at formalisere stikprøveudtagningen. Vi vil her kun betragte tilfældig stikprøveudtagning, dvs stikprøver, der udtages på en sådan måde, at ethvert populationselement har en veldefineret sandsynlighed for at indgå i stikprøven. Udtages stikprøven på denne måde kan man benytte sædvanlig statistisk argumentation til at give en kvantitativ vurdering af stikprøvens repræsentativitet i form af konfidensintervaller etc. for de ukendte populationsparametre. Den teori for repræsentative undersøgelser, som vi her skal betragte, beskriver tilfældig stikprøveudtagning, der er struktureret under hensyntagen til populationens struktur. Endvidere omfatter teorien beskrivelse af estimationsprocedurer og vurdering af estimationsusikkerhed.

I de statistiske teorier, vi sædvanligvis betragter, optræder den tilfældige variation oftest i form af processtøj eller målestøj. Ved statistisk modellering af eksperimentelle data forestiller vi os, at forsøgsresultatet er en enkelt realisation af en stokastisk variabel, hvis fordeling beskriver hyppighederne af mulige forsøgsresultater ved gentagelser af eksperimentet under identiske omstændigheder. Tilsvarende, når vi benytter statistiske modeller til beskrivelse af målefejl eller procesdata, forestiller vi os at gentagelser under identiske omstændigheder ville give anledning til data, der ville fordele sig i overensstemmelse med en bestemt fordeling. Den statistiske analyse sigter da sædvanligvis imod at karakterisere denne tænkte hyppighedsfordeling af resultaterne af uendelig mange gentagelser, f.eks. ved estimation af parametre, der fastlægger fordelingen.

Populært sagt, forestiller man sig den tilfældige mekanisme som en integreret del af den datagenerende proces, og observationen opfattes som en stikprøve fra denne uendelige population af mulige realisationer. Sædvanligvis modelleres den datagenererende proces ved en "naturlig" fordeling, som f.eks. beskrevet i afsnit 1 i Introduktion til Statistik, Bind 1.

Den klassiske teori for repræsentative undersøgelser adskiller sig fra denne angrebsvinkel. I teorien for repræsentative undersøgelser betragtes en fastholdt population af værdier, og den tilfældige mekanisme, der optræder, hidrører alene fra stikprøveudtagningen.

Formålet med den statistiske analyse er, som før, at beskrive den undersøgte population. I modsætning til den sædvanlige analyse af eksperimentelle data, hvor populationen er beskrevet ved en tænkt sandsynlighedsfordeling, er populationen her en konkret samling af værdier (nemlig de værdier, hvorfra stikprøven er udtaget), og analysen adskiller sig derfor på to principielle punkter fra sædvanlige statistiske undersøgelser:

1 En stikprøve fra en endelig population vil udtynde populationen, dvs jo større stikprøve, desto større andel af populationen kendes med sikkerhed.

2 En endelig population er karakteriseret ved et konkret, endeligt sæt af værdier. De hyppigheder, hvormed de enkelte værdier optræder i populationen, kan derfor kun med tilnærmelse beskrives ved en sædvanlig sandsynlighedsfordeling.

De klassiske teorier for repræsentative undersøgelser som eksempelvis finder anvendelse ved befolkningsundersøgelser, er derfor i et vidt omfang *ikke-parametriske* (eller rettere: modeluafhængige) i den forstand, at man ikke ønsker at basere analysen på antagelser om at hyppigheden af værdierne i populationen kan beskrives ved en specifik familie af sandsynlighedsfordelinger. I nogle tilfælde kan man opleve, at hyppighedsfordelingen af populationsværdier er multimodal (tætheden har flere toppunkter). I sådanne situationer ønsker man derfor, at de udsagn om populationen, der udledes af stikprøveresultatet, ikke afhænger af andre sandsynlighedsmodeller, end den, der beskriver stikprøveudtagningen.

I de senere år synes der dog - også indenfor befolkningstatistikken - at være en vis opblødning i denne meget rigoristiske holdning. Således diskuteres i stigende omfang de såkaldte superpopulationsmodeller, der bekriver den foreliggende population som en realisation af en parametrisk model.

I industrielle sammenhænge oplever man ofte at populationerne er store; desuden vil de populationer (f.eks. af råvarer, produkter m.v.) , der betragtes i industriel praksis, ofte kunne opfattes som sådanne realisationer af en underliggende stokastisk mekanisme, at den formodede hyppighedsfordeling af populationsværdier er så glat, at det kan være rimeligt at slække på kravet om modeluafhængighed, og modellere populationen ved en parametrisk fordeling.

Den foreliggende fremstilling sigter mod at give en indføring i de begreber og metoder, der benyttes ved repræsentative undersøgelser af en population. Fremstillingen forsøger at tilgodese den modeluafhængige holdning, der er karakteristisk for den klassiske teori.

Hovedvægten i fremstillingen vil dog ligge på udnyttelse af kendskab til populationens struktur ved udvælgelse af stikprøven. Som illustration af konsekvenserne ved brug af sådanne metoder vil vi ofte betragte populationer, der kan beskrives ved parametriske sandsynlighedsfordelinger.

1.1.1 Oversigt

I afsnit 1.2 introduceres de deskriptive statistiske populationsmål, som benyttes i teorien for repræsentative undersøgelser. Afsnit 1.3 diskuterer det ret komplicerede formelapparatur, der er nødvendigt, hvis man ønsker en traditionel beskrivelse ved stokastiske variable, og endvidere introduceres begrebet "simpel tilfældig udvælgelse", hhv. med og uden tilbagelægning, til beskrivelse af simple stikprøveudtagningssituationer.

Afsnit 1.4 og 1.5 beskriver egenskaberne for simple estimater for populationstotal og populationsvarians.

I afsnit 1.6 diskuteres situationer med flere variable pr analyseenhed, og i afsnit 1.7 introduceres det såkaldte kvotientskøn til estimation af en gennemsnitlig rate for populationens interessevariable.

Afsnit 1.8 diskuterer effekten af en udvælgelse, hvor sandsynligheden for at en enhed inddrages i stikprøven, varierer hen over populationsenhederne. Specielt diskuteres situationen, hvor enhederne udvælges stort set proportionalt med deres størrelse.

Endelig diskuteres i afsnit 1.9 de fordele, der kan opnås ved opdeling af populationen i forskelligartede dele, såkaldte strata, og i afsnit 1.10 diskuteres tilsvarende de fordele, der kan opnås ved opdeling af populationen i ensartede grupper, klynger.

1.2 Endelige populationer og tilfældige stikprøver

I abstrakt statistisk formulering er en population afgrænset ved en indeksmængde, $\mathcal{I}$, der specificerer de enkelte elementer i populationen. Som regel vil vi betragte populationer, hvor $\mathcal{I} = 1, 2, \ldots, N$, dvs. populationer med diskrete enheder. Det afgørende er imidlertid at indeksmængden er målelig med et endeligt mål (således at det er muligt at tilknytte sandsynligheder til delmængder af indeksmængden). Såfremt $\mathcal{I} = 1, 2, \ldots, N$, er det naturlige mål, μ, således tællemålet.

Undertiden er der flere muligheder for opdeling af en population i diskrete enheder. Således kan man betragte beboerne i en kommune som enkeltpersoner, fx hvis man er interesseret i rejsemønstret, eller som husstande, hvis man er interesseret i affaldsmængden. For at præcisere denne opdeling bruger man ofte ordet analyseenhed til at betegne den enhed, der bruges ved undersøgelsen.

Populationens værdier er givet ved en afbildning $x : \mathcal{I} \to \mathcal{X}$, hvor $\mathcal{X}$ angiver mængden af mulige værdier for de enkelte analyseenheder. Populationsværdierne er således

$$x(1), x(2), \ldots, x(N).$$

Ofte vil vi blot betegne populationens værdier ved

$$x_1, x_2, \ldots, x_N \quad \text{med} \quad x_i \stackrel{\text{DEF}}{=} x(i)$$

1.2.1 Populationsparametre

Formålet med en repræsentativ undersøgelse er at vurdere nogle simple deskriptive mål for populationens værdier. Almindeligvis benyttes de empiriske momenter til en sådan beskrivelse.

For at kunne skelne mellem de empiriske momenter i populationen og de tilsvarende momenter i stikprøven benytter vi græske bogstaver for populationsparametre.

For en population med værdier x_i indfører vi

Definition 1.2.1 *Populationstotal og gennemsnitlig værdi pr analyseenhed*

Da populationen har en endelig størrelse, har det mening at tale om populationstotalen af interessevariablen,

$$\zeta_x \stackrel{\text{DEF}}{=} \sum_1^N x_i \tag{1.2.1}$$

hvor fodtegnet x angiver, at populationstotalen, ζ_x, er beregnet for interessevariablen x.

Tilsvarende defineres populationgennemsnittet af interessevariablen x

$$\xi_x \stackrel{\text{DEF}}{=} \sum_1^N x_i/N \tag{1.2.2}$$

Det er klart, at kender man gennemsnittet, kan totalen beregnes ved

$$\zeta_x = N\xi_x$$

□

Til beskrivelse af variationen i populationens værdier indfører vi

Definition 1.2.2 ***Populationsvariansen og den korrigerede populationsvarians***

Populationsvariansen for interessevariablen x er

$$\sigma_x^2 = \sum_1^N (x_i - \xi_x)^2/N \tag{1.2.3}$$

□

Bemærkning 1 ***Den korrigerede populationsvarians***

I en række sammenhænge er det mere bekvemt at betragte den korrigerede populationsvarians $\sigma_x'^2$, der defineres som:

$$\sigma_x'^2 = \sum_1^N (x_i - \xi_x)^2/(N-1) \tag{1.2.4}$$

Det er åbenbart, at der gælder

$$\sigma_x'^2 = \frac{N}{N-1}\,\sigma_x^2$$

således at den ene fremkommer af den anden ved multiplikation med en kendt faktor. Som regel vil $N/(N-1)$ være så tæt ved 1, at forskellen på de to størrelser er uden betydning for ethvert praktisk formål. Når vi alligevel har valgt at operere med begge former skyldes det et hensyn til resultater som f.eks. i sætningerne 1.4.1 og 1.5.1, hvor hver af de to former er relevant for hver sin situation. □

Definition 1.2.3 ***Variationskoefficienten og den korrigerede variationskoefficient***

For $\xi_x \neq 0$ defineres variationskoefficienten γ ved:

$$\gamma_x = \sigma_x/\xi_x \tag{1.2.5}$$

og tilsvarende den korrigerede variationskoefficient γ'

$$\gamma_x' = \sqrt{\frac{N}{N-1}}\;\gamma_x = \sigma_x'/\xi_x \tag{1.2.6}$$

hvor ξ_x er givet ved (1.2.2).

Variationskoefficienten kaldes også den relative spredning. Man ser ofte den relative spredning angivet i procent.

Kvadratet på variationskoefficienten benævnes den relative varians. □

Eksempel 1.2.1 ***Alternativ variation af analyseenheder***

Såfremt analyseenhederne har alternativ variation, dvs. såfremt der kun registreres tilstedeværelse eller fravær af en kvalitativ egenskab ved de enkelte analyseenheder, kan vi kode disse værdier ved

$$x_i = \begin{cases} 0 \\ 1 \end{cases}$$

Populationens gennemsnitlige værdi pr. analyseenhed bliver da

$$\xi_x = \Sigma x_i/N = \quad \text{andelen af 1'taller i populationen,}$$

og populationens varians svarer til variansen i en Bernoullifordeling:

$$\sigma_x^2 = \xi_x(1 - \xi_x)$$

□

Definition 1.2.4 ***Populationskovariansen og -korrelationskoefficienten***

Såfremt der er tilknyttet to værdier $x(i)$ og $y(i)$ til hver enkelt analyseenhed definerer vi populationskovariansen, $\sigma_{x,y}$ mellem x og y ved

$$\sigma_{x,y} = \sum_1^N (x_i - \xi_x)(y_i - \xi_y)/N \tag{1.2.7}$$

og den korrigerede populationskovarians

$$\sigma'_{x,y} = \frac{N}{N-1}\ \sigma_{x,y} = \sum_1^N (x_i - \xi_x)(y_i - \xi_y)/(N-1) \tag{1.2.8}$$

hvor $\xi_x = \Sigma x_i/N$ og $\xi_y = \Sigma y_i/N$

Endelig defineres populationens korrelationskoefficient, $\rho_{x,y}$ mellem x og y ved

$$\rho_{x,y} = \sigma_{x,y}/(\sigma_x\sigma_y) = \sigma'_{x,y}/(\sigma'_x\sigma'_y) \tag{1.2.9}$$

hvor $\sigma_x = \sqrt{\sigma_x^2}$ og $\sigma_y = \sqrt{\sigma_y^2}$, ($\sigma'_x$ og σ'_y) angiver populationsvariansen (den korrigerede populationsvarians) for henholdsvis x og y.

I lighed med den relative varians, γ_x^2, indfører vi den relative populationskovarians, $\gamma_{x,y}$, for x og y, samt den tilsvarende korrigerede relative populationskovarians, $\gamma'_{x,y}$, ved

$$\gamma_{x,y} = \sigma_{x,y}/(\xi_x\xi_y) \quad \text{og} \quad \gamma'_{x,y} = \frac{N}{N-1}\ \gamma_{x,y} = \sigma'_{x,y}/(\xi_x\xi_y) \tag{1.2.10}$$

□

Bemærkning 1 ***Relativ populationskovarians udtrykt ved relative spredninger og korrelationskoefficient***

Den relative populationskovarians kan udtrykkes ved de relative spredninger og korrelationskoefficienten som:

$$\gamma_{x,y} = \rho\gamma_x\gamma_y \quad \text{og} \quad \gamma'_{x,y} = \rho\gamma'_x\gamma'_y \tag{1.2.11}$$

□

Der gælder tilsvarende regler for bestemmelse af populationsvarianser for produkter og kvotienter af populationsværdier som for produkter og kvotienter af stokastiske variable:

Sætning 1.2.1 ***Relativ varians af produkt og af kvotient***

Lad populationsværdierne v og z være dannet fra værdierne x og y ved

$$v(i) = x(i) \times y(i) \quad \text{og} \quad z(i) = y(i)/x(i)$$

Da gælder:

$$\gamma_v^2 \simeq \gamma_x^2 + \gamma_y^2 + 2\gamma_{x,y} \quad \text{og} \quad \gamma'^2_v \simeq \gamma'^2_x + \gamma'^2_y + 2\gamma'_{x,y} \tag{1.2.12}$$

$$\gamma_z^2 \simeq \gamma_x^2 + \gamma_y^2 - 2\gamma_{x,y} \quad \text{og} \quad \gamma'^2_z \simeq \gamma'^2_x + \gamma'^2_y - 2\gamma'_{x,y} \tag{1.2.13}$$

Bevis:
Følger ved linearisering af v og z i lighed med beviset for fejlophobningsloven. □

Eksempel 1.2.2 ***Produkt af LN-fordelte variable***

Såfremt populationen er uendelig stor, og x og y er fordelt over populationen som uafhængige $\mathrm{LN}(\alpha_x, \beta_x^2)$ og $\mathrm{LN}(\alpha_y, \beta_y^2)$- fordelte variable, har vi

$$\gamma_x^2 = \exp(\beta_x^2) - 1; \quad \text{og} \quad \gamma_y^2 = \exp(\beta_y^2) - 1$$

Endvidere er $v = x \times y$ fordelt som en $\mathrm{LN}(\alpha_x + \alpha_y, \beta_x^2 + \beta_y^2)$-fordelt variabel med

$$\gamma_v^2 = (1 + \gamma_x^2)(1 + \gamma_y^2) = \exp(\beta_x^2 + \beta_y^2) - 1$$

Såfremt β_x og β_y begge er små, finder man ved rækkeudvikling af eksponentialfunktionen, at

$$\gamma_x^2 \simeq \beta_x^2; \quad \gamma_y^2 \simeq \beta_y^2; \quad \text{og} \quad \gamma_v^2 \simeq \beta_x^2 + \beta_y^2$$

Jo mindre værdier af de relative varianser, desto bedre er altså approksimationen ved ovenstående sætning. □

Ofte vil den variable $x(\cdot)$ betegne et størrelsesmål for analyseenheden, og den variable $y(\cdot)$ betegne værdien af en interessevariabel for analyseenheden.

I sådanne situationer vil interessen ofte knytte sig til beskrivelse af en normeret værdi af y-variablen.

Definition 1.2.5 ***Relativ værdi af interessevariabel for analyseenhed***

Lad størrelsen af den i'te analyseenhed i en population være $x(i), \quad i = 1, \ldots, N$, og lad $y(i) \quad i = 1, \ldots, N$ betegne værdien af en interessevariabel for den i'te analyseenhed.

Den relative værdi, $b(\cdot)$, af interessevariablen for en analyseenhed er

$$b(i) = y(i)/x(i) \tag{1.2.14}$$

Den gennemsnitlige relative værdi af interessevariablen pr. analyseenhed er

$$\xi_b = \sum_1^N b_i \Big/ N \tag{1.2.15}$$

med populationsvariansen af den relative værdi pr analyseenhed

$$\sigma_b^2 = \sum_1^N (b_i - \xi_b)^2 \Big/ N \tag{1.2.16}$$

□

I en række tilfælde i praksis vil interessen imidlertid snarere være knyttet til en beskrivelse af den relative værdi af $y(\cdot)$ i forhold til størrelsen $x(\cdot)$ i populationen taget under ét.

Vi indfører derfor

Definition 1.2.6 *Relative værdi af interessevariabel for populationen*

Lad de variable $x(\cdot)$ og $y(\cdot)$ være som ovenfor.
Den relative værdi, β , af y i forhold til x for populationen er givet ved

$$\beta = \sum_1^N y_i \Big/ \sum_1^N x_i = \xi_y/\xi_x \tag{1.2.17}$$

□

Bemærkning 1 *Den relative værdi i populationen udtrykt ved de relative værdier pr analyseenhed*

Omformningen

$$\beta = \sum_1^N b_i x_i \Big/ \sum_1^N x_i$$

viser, at β netop er det vejede gennemsnit af de relative værdier pr analyseenhed med analyseenhedens størrelse, x_i, som vægte. □

Relationen mellem de to relative mål for interessevariablen y, ξ_b og β fremgår af

Sætning 1.2.2 *Relation mellem gennemsnitlig relativ værdi pr analyseenhed og relativ værdi i populationen*

Lad ξ_b og β være bestemt ved (1.2.15) og (1.2.17). Da er

$$\beta = \xi_b + \sigma_{b,x}/\xi_x \tag{1.2.18}$$

Bevis:
Sætningen følger ved at bemærke, at populationskovariansen mellem b og x er

$$\begin{aligned} \sigma_{b,x} &= \Sigma(b_i - \xi_b)(x_i - \xi_x)/N = \Sigma b_i x_i/N - \xi_b \xi_x \\ &= \Sigma y_i/N - \xi_b \xi_x = \xi_x(\beta - \xi_b) \end{aligned} \tag{1.2.19}$$

□

Bemærkning 1 ***Geometrisk repræsentation af gennemsnitlig relativ værdi og relativ værdi for populationen***

Hvis man for hver analyseenhed afbilder parret (x_i, y_i) i et koordinatsystem med x_i ud ad abscisseaksen og y_i op ad ordinataksen, vil den relative værdi af interessevariablen for den i'te analyseenhed repræsenteres ved hældningen b_i af linien, der forbinder origo $(0,0)$ med punktet (x_i, y_i). Den gennemsnitlige relative værdi pr analyseenhed er det aritmetiske gennemsnit af disse hældninger.

Den relative værdi β af interessevariablen for populationen repræsenteres ved hældningen af linien, der forbinder origo med tyngdepunktet (ξ_x, ξ_y) af punktskaren.

Regressionskoefficienten svarende til den sædvanlige lineære regression af y på x er

$$\beta_R = \frac{\sum(x_i - \xi_x)(y_i - \xi_y)}{\sum(x_i - \xi_x)^2} \tag{1.2.20}$$

Udtrykt ved kovarianser og varianser finder man

$$\beta_R = \frac{\sigma_{x,y}}{\sigma_x^2} = \rho_{x,y}\sigma_y/\sigma_x = \beta\rho_{x,y}\gamma_y/\gamma_x \tag{1.2.21}$$

□

Bemærkning 2 ***Gennemsnitlig relativ værdi og relativ værdi for populationen minimererer vægtede kvadratafvigelsessummer***

Betragt den vægtede kvadratafvigelsessum

$$S_1(c) = \sum_{i=0}^{N} \frac{1}{x_i} (y_i - cx_i)^2$$

for givne værdier af (x_i, y_i). Det gælder da, at den værdi af c, der minimerer $S_1(c)$, netop er $c = \beta$.

Betragtes i stedet

$$S_2(c) = \sum_{i=0}^{N} \frac{1}{x_i^2} (y_i - cx_i)^2$$

finder man, at $S_2(c)$ minimeres netop for $c = \xi_b$. □

Bemærkning 3 *Maksimal forskel mellem gennemsnitlig relativ værdi og relativ værdi for populationen*

Idet $(\beta - \xi_b) = \sigma_{b,x}/\xi_x$ finder man - ikke overraskende - , at såfremt x'erne er ens, da vil $\xi_b = \beta$. Tilsvarende gælder, at såfremt b_i' erne er ens, dvs. for $y_i = bx_i$, da vil de to gennemsnit ξ_b og β ligeledes være sammenfaldende.
Jo større numerisk værdi af kovariansen mellem b og x, desto større er forskellen på β og ξ_b.

Hvis kovariansen er negativ, vil $\beta < \xi_b$. Hvis kovariansen er positiv, har man $\xi_b < \beta$.

Ved benyttelse af Schwarz' ulighed på relationen (1.2.19) finder man, at

$$\sigma_{b,x} \leq \sigma_b \sigma_x$$

hvorfor der gælder

$$|\beta - \xi_b| \leq \sigma_b \gamma_x \tag{1.2.22}$$

□

Bemærkning 4 *Afvigelse mellem interessevariabel og fremskreven værdi svarende til den relative værdi for populationen*

Betragt nu den endimensionale størrelse d, bestemt som residualet svarende til den i'te enhed ved prædiktion af y_i med $\beta \times x_i$ (linien gennem origo og tyngdepunktet),

$$y_i = \beta\, x_i + d_i$$

Populationsmiddelværdien for d er

$$\xi_d = \frac{1}{N} \sum_{i=1}^{N} d_i = \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta x_i) = \xi_y - \beta \xi_x = 0$$

og populationsvariansen for d er

$$\sigma_d^2 = \frac{1}{N} \sum_{i=1}^{N} d_i^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta x_i)^2$$

Idet

$$y_i - \beta x_i = (y_i - \xi_y) - \beta(x_i - \xi_x)$$

ser man, at der gælder

$$\sigma_d^2 = \sigma_y^2 - 2\beta\sigma_{x,y} + \beta^2\sigma_x^2 \tag{1.2.23}$$

□

Eksempel 1.2.3 ***Effekt af reklamekampagne***

En kæde af detailforretninger gennemførte en større reklamekampagne for en bestemt mærkevare. Varen sælges i 500-g pakninger.

Lad x_i angive antallet af solgte pakninger af denne vare i den i'te butik i de to uger, der gik forud for kampagnen, og lad y_i angive antallet af solgte pakninger i de to uger, der fulgte kampagneugen.

Den gennemsnitlige værdi af omsætningsændringen for butikkerne fås ved at beregne omsætningsændringen $b_i = y_i/x_i$ for hver butik og derefter bestemme den gennemsnitlige værdi ξ_b (hen over butikker) af denne størrelse.

Den samlede omsætningsændring, β, for butikskæden fås ved at dividere det totale antal solgte pakninger efter kampagneugen $\sum y_i$ med det totale antal solgte pakninger før kampagneugen $\sum x_i$. □

Eksempel 1.2.4 ***Gennemsnitlig relativ værdi pr analyseenhed og relativ værdi af y i populationen for todimensional*** LN***-fordeling***

For at illustrere relationen mellem den gennemsnitlige relative værdi pr analyseenhed og den relative værdi i populationen vil vi betragte en parametrisk fordeling af x'er og y'er. Vi vælger igen at betragte logaritmisk normaltfordelte størrelser, da denne familie er afsluttet overfor kvotientdannelse.

Den todimensionale logaritmiske normalfordeling

I lighed med generaliseringen af den endimensionale normalfordeling til en todimensional fordeling kan vi generalisere den endimensionale logaritmiske normalfordeling til at omfatte todimensionale variable.

Vi siger at den todimensionale stokastiske variable (X, Y) følger en todimensional logaritmisk normalfordeling med parametre $(\alpha_x, \alpha_y, \beta_x^2, \beta_y^2, \beta_{x,y})$, såfremt

$$(\ln(X), \ln(Y)) \in \mathrm{N}_2(\alpha, \beta)$$

med

$$\boldsymbol{\alpha} = \begin{Bmatrix} \alpha_x \\ \alpha_y \end{Bmatrix} \quad \text{og} \quad \boldsymbol{\beta} = \begin{Bmatrix} \beta_x^2 & \beta_{x,y} \\ \beta_{x,y} & \beta_y^2 \end{Bmatrix}$$

Der gælder, at de marginale fordelinger af X og Y er henholdsvis $\mathrm{LN}(\alpha_x, \beta_x^2)$ og $\mathrm{LN}(\alpha_y, \beta_y^2)$ fordelinger, hvorfor

$$\mathsf{E}[X] = \exp(\alpha_x + \frac{1}{2}\beta_x^2) \quad \mathsf{V}[X] = \{\mathsf{E}[X]\}^2 \{\exp(\beta_x^2) - 1\}$$

og

$$\mathsf{E}[Y] = \exp(\alpha_y + \frac{1}{2}\beta_y^2) \quad \mathsf{V}[Y] = \{\mathsf{E}[Y]\}^2 \{\exp(\beta_y^2) - 1\}$$

Endvidere har man, at kovariansen mellem X og Y er

$$\begin{aligned} \mathsf{COV}[X,Y] &= \exp(\alpha_x + \alpha_y + \frac{1}{2}\beta_x^2 + \frac{1}{2}\beta_y^2) \{\exp(2\beta_{x,y}) - 1\} \\ &= \mathsf{E}[X]\,\mathsf{E}[Y] \{\exp(2\beta_{x,y}) - 1\} \end{aligned}$$

Den simultane fordeling af X og $V = XY$ er igen en todimensional logaritmisk normalfordeling med parametre:

$$\boldsymbol{\alpha}_1 = \begin{Bmatrix} \alpha_x \\ \alpha_x + \alpha_y \end{Bmatrix} \quad \text{og} \quad \boldsymbol{\beta}_1 = \begin{Bmatrix} \beta_x^2 & \beta_x^2 + \beta_{x,y} \\ \beta_x^2 + \beta_{x,y} & \beta_x^2 + \beta_y^2 + 2\beta_{x,y} \end{Bmatrix}$$

og tilsvarende får man, at den simultane fordeling af X og $Z = Y/X$ er en todimensional LN-fordeling med parametrene

$$\boldsymbol{\alpha}_2 = \begin{Bmatrix} \alpha_x \\ \alpha_y - \alpha_x \end{Bmatrix} \quad \text{og} \quad \boldsymbol{\beta}_2 = \begin{Bmatrix} \beta_x^2 & \beta_{x,y} - \beta_x^2 \\ \beta_{x,y} - \beta_x^2 & \beta_x^2 + \beta_y^2 - 2\beta_{x,y} \end{Bmatrix}$$

Der gælder således specielt for de marginale fordelinger af henholdsvis $V = XY$ og $Z = Y/X$

$$V \in \mathrm{LN}(\alpha_x + \alpha_y, \beta_x^2 + \beta_y^2 + 2\beta_{x,y})$$

og

$$Z \in \mathrm{LN}(\alpha_y - \alpha_x, \beta_x^2 + \beta_y^2 - 2\beta_{x,y})$$

Endvidere gælder

$$\mathsf{COV}[XY, X] = \mathsf{E}[X]^2\,\mathsf{E}[Y]\,\exp(2\beta_{x,y})\,[\exp\{2(\beta_x^2 + \beta_{x,y})\} - 1]$$

og

$$\mathsf{COV}[Y/X, X] = \mathsf{E}[Y]\,\exp(-\beta_x^2)\,[1 - \exp\{2(\beta_x^2 - \beta_{x,y})\}]$$

For en bivariat logaritmisk normaltfordelt population med parametrene $(\alpha_x, \alpha_y, \beta_x^2, \beta_y^2, \beta_{x,y})$ har man således den relative værdi "β" (1.2.17) af y i forhold til x for populationen

$$\mathsf{E}[Y]/\mathsf{E}[X] = \exp\{\alpha_y - \alpha_x + \frac{1}{2}(\beta_y^2 - \beta_x^2)\},$$

mens den gennemsnitlige relative værdi ξ_b pr analyseenhed (1.2.15) er

$$\begin{aligned}\mathsf{E}[Y/X] &= \exp\{\alpha_y - \alpha_x + \frac{1}{2}(\beta_x^2 + \beta_y^2 - 2\beta_{x,y})\} \\ &= \{\mathsf{E}[Y]/\mathsf{E}[X]\}\ \exp(\beta_x^2 - \beta_{x,y})\end{aligned}$$

Jo større numerisk værdi af $\beta_x^2 - \beta_{x,y}$, desto større er altså forskellen mellem de to værdier "β" (1.2.17) og ξ_b (1.2.15).

Vi bemærker iøvrigt, at regressionen af $\ln(Y)$ på $\ln(x)$ er lineær i $\ln(x)$, idet den betingede fordeling af $Y|X = x$ er en $\mathrm{LN}(\alpha_{y|x}, \beta_{y|x}^2)$ - fordeling med

$$\alpha_{y|x} = \alpha_y + \psi_{xy}\ (\beta_y/\beta_x)\ [\ln(x) - \alpha_x]$$

og

$$\beta_{y|x}^2 = \beta_y^2(1 - \psi_{xy}^2)$$

hvor

$$\psi_{xy} = \beta_{x,y}/(\beta_x\beta_y)$$

angiver korrelationskoefficienten i den simultane fordeling af $\ln(X)$ og $\ln(Y)$.

Betragtes i stedet regressionen af de ikke-logaritmerede værdier Y på X finder man,

$$\begin{aligned}\mathsf{E}[Y|X = x] &= \exp(\alpha_{y|x} + \frac{1}{2}\beta_{y|x}^2) \\ &= \exp(\alpha_y)\ [x/\exp(\alpha_x)]^k \qquad (1.2.24)\end{aligned}$$

med

$$k = \psi_{xy}(\beta_y/\beta_x) = \beta_{x,y}/\beta_x^2$$

og

$$\mathsf{V}[Y|X = x] = [\exp(\alpha_y)]^2\ [x/\exp(\alpha_x)]^{2k}\ \{\exp[(1 - \psi_{xy}^2)\ \beta_y^2] - 1\} \qquad (1.2.25)$$

Vi finder altså , at jo mere parameteren k afviger fra 1 (dvs jo mere regressionen af Y på X afviger fra en ret linie), desto større forskel er der på den gennemsnitlige relative værdi pr analyseenhed, ξ_b (1.2.15) og den relative værdi "β"af y i forhold til x for populationen (1.2.17). □

1.3 Stikprøver fra endelige populationer

fil: repr2.tex 1998-01-24

1.3.1 Målgruppe, stikprøveramme, stikprøve og tilfældig stikprøve

I praksis er det ikke altid muligt at udtage en stikprøve fra målgruppen (eng.: target population), som er hele den population, man ønsker at beskrive, men man må nøjes med et betragte en mere veldefineret population, stikprøverammen (eng.: sampling frame), som muliggør stikprøveudtagning. Stikprøverammen består således af de analyseenheder, hvorfra stikprøven udtages.

Vi skal ikke her komme yderligere ind på relationer mellem målgruppe og stikprøveramme. I nærværende fremstilling vil vi kun betragte den veldefinerede situation, hvor den målgruppe (population), der betragtes, er sådan beskaffen, at den direkte kan danne basis for stikprøveudtagning.

Vi indleder med at definere en stikprøve fra en population $\{x(i)\}_{i\in\mathcal{I}}$

Definition 1.3.1 ***Stikprøve fra en endelig population***

En stikprøve af størrelsen n fra en population $\{x(i)\}_{i\in\mathcal{I}}$ er et sæt, $s = (i_1, i_2, \ldots, i_n)$, af værdier fra populationens indeksmængde $\mathcal{I}$. De tilhørende populationsværdier er

$$x(i_1), x(i_2), \ldots, x(i_n).$$

□

Vi vil her kun beskæftige os med tilfældige stikprøver, da kun sådanne stikprøver muliggør en objektiv vurdering af repræsentativitet (skævhed) og usikkerhed.

Definition 1.3.2 ***Tilfældig stikprøve***

En tilfældig stikprøve af størrelsen n fra en population $\{x(i)\}_{i\in\mathcal{I}}$ er bestemt ved et sæt $S = (I_1, I_2, \ldots, I_n)$ af stokastiske variable med værdier i populationens indeksmængde $\mathcal{I}$.
De tilhørende populationsværdier er

$$X_1 = x(I_1), \quad X_2 = x(I_2), \ldots, X_n = x(I_n)$$

□

I stedet for den noget omstændelige notation $x(I_1), x(I_2), \ldots, x(I_n)$, vil vi som regel blot benytte betegnelsen $X_1, X_2, \ldots, X_n$ for en tilfældig stikprøve fra populationen $x_1, x_2, \ldots, x_n$.

1.3.2 Stikprøveudtagning ved simpel tilfældig udvælgelse

Udvælgelse uden tilbagelægning:

I praksis vil man som regel altid udføre stikprøveudtagning uden tilbagelægning, d.v.s. på en sådan måde, at en given analyseenhed kun kan optræde een gang i stikprøven.

Ved stikprøvetagning fra indeksmængden $\mathcal{I}$ foretaget uden tilbagelægning er mængden af mulige stikprøver mængden af talsæt $(i_1, i_2, \ldots, i_n)$, hvor $i_\nu \in \mathcal{I}$ og $i_\nu \neq i_\mu$ for $\nu \neq \mu$. Der er således ialt $N \cdot (N-1) \cdots (N-n+1)$ forskellige stikprøver af størrelsen n, idet der tages hensyn til rækkefølgen.

Udvælgelse med tilbagelægning

Da en række af udtrykkene for de statistiske egenskaber bliver simplere, hvis udvælgelsen foretages med tilbagelægning, vil vi også betragte denne situation. Ved stikprøveudtagning fra indeksmængden $\mathcal{I}$, foretaget med tilbagelægning, er mængden af mulige stikprøver mængden af talsæt $(i_1, i_2, \ldots, i_n)$, hvor $i_\nu \in \mathcal{I}$. Der er således ialt N^n forskellige stikprøver af størrelsen n.

Definition 1.3.3 ***Simpel tilfældig udvælgelse***

Udtagning af stikprøven $S = (I_1, I_2, \ldots, I_n)$ fra indeksmængden $\mathcal{I}$ siges at foregå ved simpel tilfældig udvælgelse, såfremt alle de mulige stikprøver fra $\mathcal{I}$ har samme sandsynlighed for at blive udtaget.

Simpel tilfældig udvælgelse kan foretages såvel med tilbagelægning som uden tilbagelægning.

Ved simpel tilfældig udvælgelse uden tilbagelægning forstås udvælgelse således, at de enheder, der skal indgå i stikprøven, udtages én for én blandt de tilbageværende enheder i populationen. Ved udvælgelsen af første stikprøveenhed har alle analyseenheder sandsynligheden $1/N$ for at blive udtrukket. Ved udvælgelse af anden enhed i stikprøven har alle resterende analyseenheder sandsynligheden $1/(N-1)$ for at blive udtrukket, etc. □

Sætning 1.3.1 ***Udvælgelsessandsynligheder for simpel tilfældig udvælgelse uden tilbagelægning***

Lad $I_1, I_2, \ldots, I_n$ angive en stikprøve udtaget ved simpel tilfældig udvælgelse uden tilbagelægning fra en population med indeksmængde $\mathcal{I} = 1, 2, \ldots, N$, da gælder

$$\mathsf{P}[I_\nu = i] = \frac{1}{N} \quad \text{for} \quad \nu = 1, 2, \ldots, n; \quad i \in \mathcal{I} \tag{1.3.1}$$

$$\mathsf{P}[I_\nu = i, I_\mu = j] = \begin{cases} \dfrac{1}{N(N-1)} & \nu \neq \mu, \quad i \neq j \\ 0 & \nu \neq \mu, \quad i = j \\ 0 & \nu = \mu, \quad i \neq j \\ \dfrac{1}{N} & \nu = \mu, \quad i = j \end{cases} \tag{1.3.2}$$

Bevis:
Beviset følger ved opskrivning af sandsynlighederne for udvalg fra indeksmængden □

De enkelte enheder, der indgår i stikprøven er således ikke stokastisk uafhængige. Fordelingen af den anden enhed, der udtages, afhænger af hvilken enhed, der blev udtaget som den første stikprøveenhed. Såfremt udvalgsbrøken , n/N, er lille, vil man dog som regel i praksis se bort fra denne afhængighed.

Sætning 1.3.2 *Udvælgelsessandsynligheder for simpel tilfældig udvælgelse med tilbagelægning*

Lad $I_1, I_2, \ldots, I_n$ angive en stikprøve udtaget ved simpel tilfældig udvælgelse med tilbagelægning fra en population med indeksmængde $\mathcal{I} = 1, 2, \ldots, N$, da gælder

$$\mathsf{P}[I_\nu = i] = \frac{1}{N} \quad \text{for} \quad \nu = 1, 2, \ldots, n; \quad i \in \mathcal{I} \tag{1.3.3}$$

$$\mathsf{P}[I_\nu = i, I_\mu = j] = \begin{cases} \dfrac{1}{N^2} & \nu \neq \mu, \quad (i,j) \in \mathcal{I}^2 \\ \dfrac{1}{N^2} & \nu = \mu, \quad i = j \\ 0 & \nu = \mu, \quad i \neq j \end{cases} \tag{1.3.4}$$

Bevis:
Beviset følger ved at bemærke, at der ved hver udtrækning er sandsynligheden $1/N$ for at en bestemt populationsenhed indgår i stikprøven, og at de enkelte trækninger er uafhængige. □

For at få en mere direkte beskrivelse af hvilke analyseenheder, der indgår i stikprøven indfører vi

Definition 1.3.4 *Udvælgelsesvektor*

Ved udvælgelsesvektoren for en stikprøve udtaget ved simpel tilfældig udvælgelse forstår vi den N-dimensionale søjlevektor $\mathbf{U}$, hvis i'te element er givet ved:

$$U_i = I_{[I_1=i]} + I_{[I_2=i]} + \cdots + I_{[I_2=i]} \quad \text{for} \quad i = 1, 2, \ldots, N$$

□

Udvælgelsesvektoren er en stokastisk vektor, hvis i'te element angiver, hvorvidt den i'te populationsenhed indgår i stikprøven. Fordelingen af udvælgelsesvektoren beskriver stikprøvevariationen, der som tidligere anført principielt set er det eneste stokastiske element i stikprøveundersøgelsen.

De første momenter af udvælgelsesvektoren er givet i følgende

Sætning 1.3.3 *Momenter for udvælgelsesvektoren*

For simpel tilfældig stikprøveudtagning gælder:

$$\mathsf{E}\,[\mathbf{U}] = \frac{n}{N}\ \mathbf{1}_N \tag{1.3.5}$$

Såfremt stikprøven er udtaget uden tilbagelægning, er dispersionsmatricen for udvælgelsesvektoren:

$$\mathbf{D}\,[\mathbf{U}] = \frac{n}{N-1}\left(1-\ \frac{n}{N}\right)\left(\mathbf{I}_N - \ \frac{1}{N}\ \mathbf{J}_N\right) \tag{1.3.6}$$

For stikprøveudtagning med tilbagelægning er dispersionsmatricen

$$\mathbf{D}\,[\mathbf{U}] = \frac{n}{N}\ \left(\mathbf{I}_N - \ \frac{1}{N}\ \mathbf{J}_N\right), \tag{1.3.7}$$

hvor $\mathbf{I}_N$ som sædvanligt betegner enhedsmatricen af dimension $N \times N$ og $\mathbf{J}_N$ angiver den $N \times N$-dimensionale matrix med lutter ettaller, $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}_N^T$.

Bevis:
For udtagelse uden tilbagelægning har vi

$$\mathsf{E}\,[U_i U_j] = \begin{cases} \dfrac{n}{N} & \text{for} \quad i = j \\ \dfrac{n(n-1)}{N(N-1)} & \text{for} \quad i \neq j \end{cases}$$

hvorfor

$$\mathsf{COV}[U_i U_j] = \begin{cases} \dfrac{n}{N}\left(1-\ \dfrac{n}{N}\right) & \text{for} \quad i = j \\ -\ \dfrac{n}{N}\left(1-\ \dfrac{n}{N}\right)\dfrac{1}{N-1} & \text{for} \quad i \neq j \end{cases}$$

For stikprøveudtagning med tilbagelægning gælder:

$$\mathsf{E}\,[U_iU_j] = \begin{cases} \frac{n}{N} + \frac{n(n-1)}{N^2} & \text{for} \quad i = j \\ \frac{n(n-1)}{N^2} & \text{for} \quad i \neq j \end{cases}$$

hvorfor

$$\mathsf{COV}[U_iU_j] = \begin{cases} \frac{n}{N}\left(1 - \frac{n}{N}\right) & \text{for} \quad i = j \\ -\frac{n}{N^2} & \text{for} \quad i \neq j \end{cases}$$

□

Bemærkning 1 ***Dispersionsmatricen for udvælgelsesvektoren udtrykt som equikorrelatonsmatrix***

Vi bemærker, at dispersionsmatricen, $\mathbf{D}\,[\mathbf{U}]$, kan udtrykkes på den sædvanlige form for equikorrelationsmatricer:

For udtagning uden tilbagelægning finder man

$$\mathbf{D}\,[\mathbf{U}] = \frac{n}{N}\left(1 - \frac{n}{N}\right)\{(1-\rho)\mathbf{I}_N + \rho\mathbf{J}_N\}$$

med $\rho = -1/(N-1)$. Dispersionsmatricen er netop dispersionsmatricen for den N-dimensionale hypergeometriske fordeling.

For udtagning med tilbagelægning gælder

$$\mathbf{D}\,[\mathbf{U}] = \frac{n}{N}\left(1 - \frac{1}{N}\right)\{(1-\rho)\mathbf{I}_N + \rho\mathbf{J}_N\}$$

med $\rho = -1/(N-1)$. Dispersionsmatricen er netop dispersionsmatricen svarende til den N-dimensionale Multinomialfordeling, $\mathrm{Mult}_N(n; 1/N, 1/N, \ldots, 1/N)$. □

1.4 Estimation af populationstotalen eller populationsgennemsnit

Vi betragter en population af størrelsen N med populationsværdierne af de N analyseenheder $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}^T$.

Den gennemsnitlige værdi pr analyseenhed er

$$\xi = \frac{1}{N}\,\mathbf{1}_N\mathbf{x} = \Sigma x_i/N$$

og populationsvariansen er

$$\sigma^2 = \frac{1}{N}\,\mathbf{x}^T\Big(\mathbf{I}_N - \frac{1}{N}\,\mathbf{J}_N\Big)\mathbf{x} = \Sigma(x_i-\xi)^2/N$$

med den korrigerede populationsvarians

$$\sigma'^2 = \frac{N}{N-1}\,\sigma^2$$

Benytter vi betegnelsen $X_\nu = x(I_\nu)$ for den ν-te stikprøveenhed har vi for en stikprøve på n enheder de n stikprøveværdier $X_1, X_2, \ldots, X_n$ og stikprøvegennemsnittet

$$\overline{X}_\cdot = \frac{1}{n}\,\mathbf{x}^T\mathbf{U} = \sum_1^n X_\nu \Big/ n \tag{1.4.1}$$

Sætning 1.4.1 ***Forventningsværdi og varians for stikprøvegennemsnit***
Betragt en stikprøve på n elementer, $X_1, X_2, \ldots, X_n$, og lad stikprøvegennemsnittet $\overline{X}_\cdot$ være givet ved (1.4.1).

Såfremt stikprøven er udtaget ved simpel tilfældig udvælgelse gælder:

$$\mathsf{E}\,[\overline{X}_\cdot] = \xi \tag{1.4.2}$$

For udvælgelse uden tilbagelægning gælder

$$\mathsf{V}\,[\overline{X}_\cdot] = \frac{\sigma'^2}{n}\,\Big(1-\frac{n}{N}\Big) \tag{1.4.3}$$

og for udvælgelse med tilbagelægning er

$$\mathsf{V}\,[\overline{X}.] = \frac{\sigma^2}{n} \tag{1.4.4}$$

Bevis:
Sætningen vises ved at bemærke, at

$$\mathsf{E}\,[\overline{X}.] = \frac{1}{n}\,\mathbf{x}^T\,\mathsf{E}\,[\mathbf{U}]$$

og

$$\mathsf{V}\,[\overline{X}.] = \frac{1}{n^2}\,\mathbf{x}^T\;\mathbf{D}\,[\mathbf{U}]\,\mathbf{x}$$

□

Bemærkning 1 ***Momenter for enkelte stikprøveenheder***

Ved betragtning af de n separate udvælgelsesvektorer svarende til de n stikprøveenheder finder man momenterne for de enkelte stikprøveenheder

$$\mathsf{E}\,[X_\nu] = \xi; \qquad \mathsf{V}\,[X_\nu] = \sigma^2$$

For udvælgelse uden tilbagelægning er for $\nu \neq \mu$

$$\mathsf{COV}[X_\nu, X_\mu] = -\;\frac{1}{N-1}\,\sigma^2 = -\;\frac{1}{N}\,\sigma'^2$$

og for udvælgelse med tilbagelægning er for $\nu \neq \mu$

$$\mathsf{COV}[X_\nu, X_\mu] = 0$$

Faktoren $(1-n/N)$ i udtrykket for variansen på $\overline{X}.$ ved udtagning uden tilbagelægning skyldes den indbyrdes korrelation mellem elementerne i stikprøven på grund af udtyndingen af populationen. Faktoren kaldes også korrektionsfaktoren for den endelige population, (engelsk: *finite population correction*, ofte forkortet til *fpc*). Faktoren er praktisk taget 1 for udvalgsbrøker $f = n/N < 0.01$.

Når bortses fra korrektionen for den endelige population ved stikprøvetagning uden tilbagelægning, er variansen på stikprøvegennemsnittet proportional med den reciprokke stikprøvestørrelse. Usikkerheden på stikprøvegennemsnittet afhænger således i det væsentlige af stikprøvens absolutte størrelse, n. Når blot udvalgsbrøken $f = n/N$ er mindre end 0.1, er korrektionen sædvanligvis af underordnet betydning. □

Eksempel 1.4.1 *Alternativt varierende analyseenheder*

Hvis analyseenhederne har alternativ variation, er fordelingen af stikprøvetotalen

$$X_+ = \sum_1^n X_\nu$$

en hypergeometrisk fordeling, $X_? \in \mathrm{H}(n, N\xi, N)$ med

$$\mathsf{E}[X_+] = n\,\xi; \qquad \mathsf{V}[X_+] = n\,\xi(1-\xi)\Big(1 - \frac{n-1}{N-1}\Big)$$

□

1.5 Estimation af populationsvarians

1.5.1 Momenter for stikprøvevariansen

En naturlig estimator for populationsvariansen er stikprøvevariansen

$$S^2 = \sum_1^n (X_\nu - \overline{X}_\cdot)^2/(n-1) \tag{1.5.1}$$

For stikprøvevariansen (1.5.1) gælder

Sætning 1.5.1 *Forventningsværdi og varians for stikprøvevarians*

Betragt en stikprøve på n elementer, $X_1, X_2, \ldots, X_n$, og lad stikprøvevariansen S^2 være givet ved (1.5.1).

Såfremt stikprøven udvælges uden tilbagelægning er

$$\mathsf{E}[S^2] = \sigma'^2 \tag{1.5.2}$$

$$\mathsf{V}[S^2] \simeq \frac{\sigma^4}{n}\Big(\mu_4/\sigma^4 - 1 + \frac{2}{n-1}\Big)\Big(1 - \frac{n}{N}\Big) \tag{1.5.3}$$

hvor σ'^2 angiver den korrigerede populationsvarians (1.2.4), σ^2 angiver populationsvariansen (1.2.3), og hvor μ_4 betegner populationens fjerde centrale moment

$$\mu_4 = \sum_1^N (x_i - \xi)^4 / N \tag{1.5.4}$$

Såfremt udvælgelsen er med tilbagelægning gælder

$$\mathsf{E}[S^2] = \sigma^2 \tag{1.5.5}$$

$$\mathsf{V}[S^2] = \frac{2\sigma^4}{n-1} + \frac{\mu_4 - 3\sigma^4}{n} \tag{1.5.6}$$

Bevis:
Udtrykket for $\mathsf{E}[S^2]$ fås ved benyttelse af relationerne

$$\mathsf{V}[X_\nu] = \sigma^2; \quad \text{og} \quad \mathsf{COV}[X_\nu, X_\mu] = -\frac{1}{N}\sigma'^2$$

for udtagning uden tilbagelægning, samt

$$\mathsf{V}[X_\nu] = \sigma^2; \quad \text{og} \quad \mathsf{COV}[X_\nu, X_\mu] = 0$$

for udvælgelse med tilbagelægning.

Opskrives udtrykket for $\mathsf{V}[S^2]$ svarende til udtagning uden tilbagelægning får man - efter nogen reduktion - det eksakte udtryk :

$$\begin{aligned}
&\mathsf{V}[S^2] = \\
&\frac{\mu_4}{n}\left(\frac{n}{n-1}\right)^2 \left\{1 - \frac{n-1}{N-1} \times \frac{2}{n}\left[1 - 3\frac{n-1}{N-1} + 2\frac{(n-1)(n-1)}{(N-1)(N-2)}\right]\right. \\
&\left. + \frac{1}{n^2}\left[1 - 7\frac{n-1}{N-1} + 12\frac{(n-1)(n-2)}{(N-1)(N-2)} - 6\frac{(n-1)(n-2)(n-3)}{(N-1)(N-2)(N-3)}\right]\right\} \\
&- \frac{\sigma^4}{n-1}\ \frac{N}{N-1}\left\{1 + \frac{n}{N-1} - 2\frac{n-2}{N-2}\right. \\
&\left. - \frac{3}{n}\left[1 - 2\frac{n-2}{N-2} + \frac{(n-2)(n-3)}{(N-2)(N-3)}\right]\right\}
\end{aligned} \tag{1.5.7}$$

Erstatter man faktorerne $(n-1)/(N-1)$, $(n-2)/(N-2)$ og $(n-3)/(N-3)$ med dvalgsbrøken $f = n/N$, og ser man iøvrigt bort fra led af størrelsen $1/N, 2/N, 3/N$ og $4/N$, får man (1.5.3). Udtrykket for $\mathsf{V}[S^2]$ svarende til udtagning med tilbagelægning følger ved udnyttelse af at de udvalgte stikprøveenheder er uafhængige og identisk fordelte. □

Bemærkning 1 ***Jo tykkere haler, desto større usikkerhed på variansskønnet***

Sætningen viser, at jo større værdi af populationens fjerde centrale moment, desto større bliver usikkerheden på variansskønnet.

For uafhængige observationer fra normalfordelingen gælder $\mu_4 = 3\sigma^4$, hvorfor (1.5.6) i dette tilfælde reduceres til

$$\mathsf{V}[S^2] = \frac{2\sigma^4}{n-1}$$

Hvis fordelingen af x'erne har tykkere haler, end normalfordelingen, som f.eks. en $\mathrm{G}(\alpha,\beta)$- fordeling, har vi $\sigma^2 = \beta^2\alpha$; $\mu_4 = 3\beta^4\alpha(\alpha+2)$, hvorfor (1.5.6) i tilfældet med uafhængige observationer fra en $\mathrm{G}(\alpha,\beta)$-fordeling bliver

$$\mathsf{V}[S^2] = \frac{2\sigma^4}{n-1} + \frac{6\sigma^4}{n\alpha}$$

med $\sigma^2 = \beta^2\alpha$. □

1.5.2 Konfidensgrænser:

Ved brug af den centrale grænseværdisætning kan man konstruere approksimative konfidensgrænser for populationsgennemsnittet $\xi_x = \overline{x}_.$.

Vi bruger stikprøvegennemsnittet (1.4.1) som estimat for populationsgennemsnittet, dvs.

$$\hat{\xi}_x = \overline{X}_.;$$

Udvælgelse uden tilbagelægning

Det følger af en variant af den centrale grænseværdisætning, at $\overline{X}_.$ approksimativt vil følge en normalfordeling med forventningsværdi (1.4.2) og varians (1.4.3).

Såfremt vi kendte den korrigerede populationsvarians $\sigma_x'^2$, ville et approksimativt konfidensinterval være

$$\overline{X}. \pm u_{1-\alpha/2}\frac{\sigma_x'}{\sqrt{n}}\sqrt{1-f} \tag{1.5.8}$$

I stedet for populationsvariansen bruger vi estimatet

$$\widehat{\sigma}_x'^2 = S^2 = \sum_{\nu=1}^{n}(X_\nu - \overline{X}.)^2/(n-1)$$

Vi har dermed det approksimative $(1-\alpha)$ - konfidensinterval for populationsmiddelværdien ξ_x:

$$\overline{X}. \pm u_{1-\alpha/2}\frac{S}{\sqrt{n}}\sqrt{1-f}\,, \tag{1.5.9}$$

hvor $f = n/N$ angiver udvalgsbrøken, og u_P som sædvanlig betegner P-fraktilen i den normerede normale fordeling.

For populationstotalen, $x_+ = N\xi_x$, finder man tilsvarende det approksimative $(1-\alpha)$ - konfidensinterval ved multiplikation med populationsstørrelsen N:

$$N\overline{X}. \pm u_{1-\alpha/2}\, N\,\frac{S}{\sqrt{n}}\sqrt{1-f}\,. \tag{1.5.10}$$

Bemærkning 1 ***Konfidensinterval med fastlagt længde***

Længden af det approksimative $(1-\alpha)$-konfidensinterval (1.5.8) for populationsgennemsnittet ξ_x er

$$\begin{aligned} l &= 2u_{1-\alpha/2}\frac{\sigma_x'}{\sqrt{n}}\sqrt{1-f} \\ &= 2u_{1-\alpha/2}\sigma_x'\sqrt{\frac{1}{n}-\frac{1}{N}} \end{aligned}$$

Såfremt man ønsker et interval af en fastlagt længde, l_0, skal stikprøvestørrelsen n således tilfredstille relationen

$$n \geq \frac{1}{\left(\frac{l_0/\sigma'_x}{2u_{1-\alpha/2}}\right)^2 + \frac{1}{N}} \tag{1.5.11}$$

□

Udvælgelse med tilbagelægning

Såfremt udtagningen foregår med tilbagelægning får man det approksimative $(1-\alpha)$ - konfidensinterval for populationsgennemsnittet ξ_x:

$$\overline{X}. \pm u_{1-\alpha/2}\frac{S}{\sqrt{n}} \tag{1.5.12}$$

og tilsvarende det approksimative $(1-\alpha)$ - konfidensinterval for populationstotalen $\zeta_x = N\xi_x$:

$$N\overline{X}. \pm u_{1-\alpha/2}\, N\, \frac{S}{\sqrt{n}} \tag{1.5.13}$$

Udtrykkene (1.5.12) og (1.5.13), benyttes også ved udvælgelse uden tilbagelægning, når udvalgsbrøken $f = n/N$ er lille.

Bemærkning 2 *Brug af normalfordelingsfraktiler for situationer med ukendt spredning*

Vi minder om reglen fra Introduktion til Statistik, Bind 1 vedrørende konfidensintervaller for forventningsværdien baseret på en stikprøve af uafhængige normaltfordelte størrelser. I situationen, hvor også spredningen må estimeres fra stikprøven, benyttes en fraktil i t-fordelingen for at tilgodese den ekstra variation, der hidrører fra den estimerede spredning.

Når vi til trods herfor benytter normalfordelingsfraktilen $u_{1-\alpha/2}$ i udtrykkene (1.5.9), (1.5.12) og (1.5.13) skyldes det, at selve det at bruge normalfordelingen

er en approksimation, og det vil give et falsk skær af nøjagtighed, hvis man foretager en sådan korrektion af den fraktil, der indgår i approksimationen. □

Bemærkning 3 ***Konfidensinterval ved populationer med stor skævhed***

Man vil somme tider opleve, at fordelingen af de kendetegn, der undersøges, udviser en klar positiv skævhed (få store værdier og mange små). For sådanne populationer gælder det, at den øvre konfidensgrænse viser sig at være for lav med en hyppighed, der er større end $\alpha/2$, mens nedre konfidensgrænse viser sig at være for høj, sjældnere end angivet ved $\alpha/2$. Sammenfattende vil den påståede konfidensgrad (e.g. 95 %) være for høj. □

Sætning 1.5.2 ***Konfidensgrænser ved alternativt varierende analyseenheder***

Såfremt enhederne i populationen har alternativ variation, d.v.s. såfremt x_i kun antager værdierne 0 eller 1, er $100(1-\alpha)$-konfidensintervallet for populationstotalen $x.$ bestemt ved

$$X_{+L} \leq X_+ \leq X_{+U}$$

Såfremt stikprøven udtages med tilbagelægning er X_{+L} og X_{+U} bestemt ved

$$X_{+L} = Np_L \quad \text{og} \quad X_{+U} = Np_U \tag{1.5.14}$$

hvor p_L og p_U er de tilsvarende konfidensgrænser i *binomialfordelingen*, bestemt ved hjælp af fraktiler i Beta-fordelingen eller F-fordelingen (jvf. Introduktion til Statistik, Bind 1):

$$\begin{aligned} p_L &= \mathrm{Be}(X_+, n - X_+ + 1)_{\alpha/2} \\ &= \frac{X_+}{X_+ + (n - X_+ + 1)\mathrm{F}(2n - 2X_+ + 2, 2X_+)_{1-\alpha/2}} \\ p_U &= \mathrm{Be}(X_+ + 1, n - X_+)_{1-\alpha/2} \\ &= \frac{X_+ + 1}{X_+ + 1 + (n - X_+)\mathrm{F}(2n - 2X_+, 2X_+ + 2)_{\alpha/2}} \end{aligned} \tag{1.5.15}$$

Såfremt stikprøven udtages uden tilbagelægning, bestemmes X_{+L} og X_{+U} ved

$$\begin{aligned} \mathrm{H}(X.; n, N, X_{+U}) &\approx \alpha/2 \\ \mathrm{H}(X. - 1; n, N, X_{+L}) &\approx 1 - \alpha/2 \end{aligned} \tag{1.5.16}$$

hvor $\mathrm{H}(c; n, N, M)$ angiver den *kumulerede hypergeometriske fordeling*, $\mathsf{P}\,[\mathrm{H}(n, N, M) \leq c]$.

Konfidensgrænserne X_{+L} og X_{+U} må i dette tilfælde bestemmes ved iteration. Som udgangspunkt for iterationen kan man benytte approksimationen

$$X_{+U} \approx N\,\frac{X_+}{n} + N\,\sqrt{(N-n)/(N-1)}\;(p_U - \frac{X_+}{n})$$

og

$$X_{+L} \approx N\,\frac{X_+}{n} - N\,\sqrt{(N-n)/(N-1)}\;(\frac{X_+}{n} - p_L)$$

hvor p_L og p_U er bestemt ved (1.5.15)

Bevis:
Konfidensgrænserne bestemmes udfra overvejelser i analogi med bestemmelsen af binomialfordelingsintervallerne i Introduktion til Statistik, Bind 1.

Approksimationen består i at indsnævre det konfidensinterval, man ville have fundet, såfremt udtagningen havde været med tilbagelægning. Inds-

nævringen foregår ved at intervallængden multipliceres med den sædvanlige korrektionsfaktor for en endelig population.

Da parameterrummet (de mulige værdier af x_+) er diskret, kan man ikke altid sikre, at konfidensgraden netop er $100(1-\alpha)$. Ønsker man at sikre, at konfidensgraden er mindst $100(1-\alpha)$, kan man bestemme X_{+L} og X_{+U} ved

$$\mathrm{H}(X_+; n, N, X_U) \leq \alpha/2 \quad \text{og} \quad 1 - \mathrm{H}(X_+ - 1; n, N, X_{+L}) \leq \alpha/2$$

□

Eksempel 1.5.1 ***Konfidensinterval ved alternativt varierende analyseenheder***

En husgavl er belagt med 263 beskyttelsesfliser. På grund af vejrets påvirkning og kvalitetsvariationer ved opsætningen er nogle af pladerne imidlertid blevet løse eller forvitrede og er tjenlige til udskiftning.

For at vurdere udskiftningens omfang udvalgtes en tilfældig stikprøve på $n = 56$ fliser, der blev lydprøvet for at vurdere vedhæftningen. Man fandt ialt $X_+ = 6$ "hule"fliser i denne stikprøve.

For at bestemme et 95 % konfidensinterval for det *totale antal hule fliser* bestemmes først konfidensintervallet for *andelen* p af hule fliser under antagelse af at udtagningen var foregået med tilbagelægning.

Man finder $p_L = \mathrm{Be}(6, 51)_{0.025} = 0.0403$ og $p_U = \mathrm{Be}(7, 50)_{0.975} = 0.2188$ således at første approksimation til X_L og X_U

$$X_L \approx 263 \times \{6/56 - \sqrt{0.79}\,(6/56 - 0.0403)\} = 12.55$$

og

$$X_U \approx 263 \times \{6/56 + \sqrt{0.79}\,(0.2188 - 6/56)\} = 54.23$$

Nedenstående tabeller viser værdierne af $1 - \mathrm{H}(5; 56, 263, X_+)$ for $X_+ = 10, 11, \ldots, 14$, samt af $\mathrm{H}(6; 56, 263, X_+)$ for $X_+ = 53, 54, \ldots, 58$:

X_+	10	11	12	13	14
$1 - \mathrm{H}(5; 56, 263, X_+)$	0.0076	0.0139	0.0232	0.0360	0.0526

X_+	53	54	55	56	57	58
$\mathrm{H}(6; 56, 263, X_+)$	0.0314	0.0266	0.0224	0.0189	0.0158	0.0132

Man finder derfor med en sikkerhed på lidt over 95 %, at antallet af fejlbehæftede plader er indeholdt i intervallet bestemt ved $X_L = 12$ og $X_U = 55$.

□

1.6 Stikprøver fra populationer med flere værdier pr analyseenhed

fil: repr3.tex 1998-01-24

1.6.1 Stikprøvekovarians

Såfremt hver analyseenhed er tilknyttet to værdier $x(i)$ og $y(i)$, vil man ofte benytte en enkelt stikprøve til at estimere populationsmiddelværdierne ξ_x og ξ_y af de to variable.

Lad $X_\nu = x(I_\nu)$ og $Y_\nu = y(I_\nu)$; $\nu = 1, 2, \ldots, n$ angive værdierne af de to variable for de n udvalgte enheder $I_1, I_2, \ldots, I_n$, og lad

$$\overline{X}. = \sum_1^n X_\nu/n \qquad \text{og} \quad \overline{Y}. = \sum_1^n Y_\nu/n \tag{1.6.1}$$

angive de tilsvarende stikprøvegennemsnit.

Forventningsværdi og varians for $\overline{X}.$ og $\overline{Y}.$ er givet i sætning 1.4.1. Da værdien af X_ν og Y_ν er knyttet til den samme analyseenhed, vil $\overline{X}.$ og $\overline{Y}.$ imidlertid ikke nødvendigvis variere uafhængigt af hverandre. Samvariationen mellem de to estimater er givet i

Sætning 1.6.1 ***Kovarians mellem stikprøvegennemsnit***

Lad $\overline{X}.$ og $\overline{Y}.$ være bestemt ved (1.6.1) .

Såfremt stikprøven er udtaget uden tilbagelægning er:

$$\mathsf{COV}[\overline{X}., \overline{Y}.] = \frac{1}{n}\,(1-f)\,\sigma'_{xy} \tag{1.6.2}$$

hvor f angiver udvalgsbrøken, $f = n/N$.

For stikprøveudtagning med tilbagelægning er

$$\mathsf{COV}[\overline{X}_\cdot, \overline{Y}_\cdot] = \frac{1}{n}\,\sigma_{xy} \tag{1.6.3}$$

hvor populationskovariansen σ_{xy} og den korrigerede populationskovarians σ'_{xy} er givet ved (1.2.7) og (1.2.8)

Bevis:
Følger ved at udnytte, at

$$\mathsf{COV}[X_\nu, Y_\mu] = \begin{cases} \sigma_{xy} & \text{for} \quad \nu = \mu \\ -\dfrac{1}{N-1}\,\sigma_{xy} & \text{for} \quad \nu \neq \mu \end{cases},$$

såfremt udvælgelsen foregår uden tilbagelægning, og

$$\mathsf{COV}[X_\nu, Y_\mu] = \begin{cases} \sigma_{xy} & \text{for} \quad \nu = \mu \\ 0 & \text{for} \quad \nu \neq \mu \end{cases},$$

såfremt udvælgelsen foregår med tilbagelægning. □

Bemærkning 1 ***For udtagning med tilbagelægning er korrelationen mellem stikprøvegennemsnittene er den samme som mellem populationsværdierne***

Det fremgår af (1.6.3), at korrelationen mellem $\overline{X}$ og $\overline{Y}$ er den samme som korrelationen mellem X og Y. □

Som grundlag for estimation af populationskovariansen σ_{xy}, eller den korrigerede populationskovarians σ'_{xy}, kan man benytte stikprøvekovariansen

$$\widehat{\sigma}'_{xy} = \frac{1}{n-1}\sum_{\nu=1}^{n}(X_\nu - \overline{X}_\cdot)(Y_\nu - \overline{Y}_\cdot) \tag{1.6.4}$$

Der gælder

Sætning 1.6.2 ***Forventningsværdi af stikprøvekovarians***

Lad stikprøvekovariansen $\widehat{\sigma}'_{xy}$ være bestemt ved (1.6.4).
Såfremt stikprøven er udvalgt uden tilbagelægning er:

$$\mathsf{E}\,[\,\widehat{\sigma}'_{xy}] = \sigma'_{xy} \tag{1.6.5}$$

For stikprøveudvælgelse med tilbagelægning er

$$\mathsf{E}\,[\,\widehat{\sigma}'_{xy}] = \sigma_{xy} \tag{1.6.6}$$

Bevis:
Følger i analogi med beviset for sætning 1.6.1.

□

Bemærkning 1 *Centralt estimat for populationskovarians*

Såfremt man ønsker et centralt estimat for den ukorrigerede populationskovarians σ_{xy}, kan man ved stikprøveudtagning uden tilbagelægning benytte det korrigerede estimat

$$\widehat{\sigma}_{xy} = \frac{N-1}{N}\,\widehat{\sigma}'_{xy}$$

Eksempel 1.6.1 *Samvariation mellem affaldsmængde for en husstand og antallet af personer i husstanden*

I en kommune med 2000 husstande er man interesseret i at vurdere mængden af husholdningsaffald.

Lad y_i angive den årlige affaldsmængde i den i'te hustand, og lad x_i angive størrelsen af den i'te husstand (antallet af personer).

Ved simpel tilfældig udvælgelse har man udvalgt en stikprøve på 10 husstande og igennem et år vejet mængden af husholdningsaffald i husstanden. Den årlige affaldsmængde for de 10 husstande er anført i nedenstående tabel.

Lbnr	1	2	3	4	5	6	7	8	9	10
Antal pers	1	1	2	2	3	3	2	2	1	5
Affalds-mgd [kg]	632	926	670	521	1016	1174	797	968	330	1450

Den gennemsnitlige husstandsstørrelse i stikprøven er $\overline{X}. = 2.20$ [pers./husstand]. Stikprøvevariansen er

$$S_X^2 = \frac{13.60}{9} = 1.5444 = 1.23^2$$

og den gennemsnitlige affaldsmængde pr husstand i stikprøven er $\overline{Y}. = 848.4$ [kg/husstand] med stikprøvevariansen

$$S_Y^2 = \frac{973\ 580.40}{9} = 108\ 175.6 = (328.9\ [\text{kg}])^2$$

Idet

$$\sum_{\nu=1}^{10}(X_\nu - \overline{X}.)(Y_\nu - \overline{Y}.) = 2\ 955.20$$

får man estimatet for kovariansen mellem husstandsstørrelse og affaldsmængde i populationen

$$\widehat{\sigma}'_{xy} = \frac{2\ 955.20}{9} = 328.36$$

Der er altså en positiv samvariation mellem antallet af personer i husstanden og mængden af husstandens husholdningsaffald. □

1.6.2 Relativ værdi pr analyseenhed

Den relative værdi pr analyseenhed $b_i = y_i/x_i$ er en endimensional størrelse, og estimation af de tilsvarende populationsværdier kan derfor udføres efter retningslinierne i afsnit 1.4 og 1.5.

Populationsmiddelværdien af b_i er den gennemsnitlige relative værdi pr analyseenhed

$$\xi_b = \frac{1}{N}\sum_1^N y_i/x_i \tag{1.6.7}$$

Sætning 1.6.3 *Forventningsværdi og varians for stikprøvegennemsnit af relative værdier*

Betragt skønnet

$$\overline{B}. = \frac{1}{n} \sum_{1}^{n} B_\nu \tag{1.6.8}$$

hvor $B_\nu = Y_\nu / X_\nu$; $\nu = 1, 2, \ldots, n$ angiver stikprøveværdierne af de relative værdier pr analyseenhed.

For simpel tilfældig udvælgelse gælder:

$$\mathsf{E}\,[\,\overline{B}.] = \xi_b \tag{1.6.9}$$

Såfremt udvælgelsen foretages uden tilbagelægning er

$$\mathsf{V}\,[\,\overline{B}.] = \frac{1}{n}\,\sigma'^2_b \left(1 - \frac{n}{N}\right) \tag{1.6.10}$$

og såfremt udvælgelsen foretages med tilbagelægning gælder

$$\mathsf{V}\,[\,\overline{B}.] = \frac{1}{n}\,\sigma^2_b \tag{1.6.11}$$

Bevis:
Følger umiddelbart af sætning 1.4.1 □

Kovariansen mellem $\overline{B}.$ og $\overline{X}.$ fremgår ved benyttelse af sætning 1.6.1.

Sætning 1.6.4 *Central estimator for kovariansen mellem analyseenhedens størrelse og den relative værdi af interessevariablen*

Estimatoren

$$\widehat{\sigma}_{b,x} = \frac{N-1}{N}\frac{1}{n-1}\sum_{1}^{n}(B_\nu - \overline{B}.)(X_\nu - \overline{X}.) \tag{1.6.12}$$

er en central estimator for $\sigma_{b,x}$, såfremt udvælgelsen foregår uden tilbagelægning.

Såfremt udvælgelsen foregår med tilbagelægning, vil

$$\widehat{\sigma}'_{b,x} = \frac{N}{N-1}\widehat{\sigma}_{b,x}$$

være central for $\sigma_{b,x}$.

Bevis:
Resultatet følger af sætning 1.6.2. □

Bemærkning 1 ***Beregning af estimatoren for*** $\sigma_{b,x}$

Estimatoren udregnes lettest ved at benytte relationen:

$$\begin{aligned}\widehat{\sigma}_{b,x} &= \frac{N-1}{N}\frac{1}{n-1}\sum_{1}^{n}(B_\nu - \overline{B}.)(X_\nu - \overline{X}.)\\ &= \frac{N-1}{N}\frac{1}{n-1}\sum_{1}^{n}(Y_\nu - \overline{B}.\, X_\nu)\\ &= \frac{N-1}{N}\frac{n}{n-1}(\overline{Y}. - \overline{B}.\,\overline{X}.)\end{aligned} \tag{1.6.13}$$

der udtrykker, at kovariansestimatet fås som afvigelsen mellem gennemsnittet af interessevariablen og den værdi, der fås ved at "fremskrive"den gennemsnitlige x-værdi med en faktor, der svarer til gennemsnittet af de enkelte y-værdiers "fremskrivningsfaktor". □

1.7 Kvotientskøn

Vi skal i dette afsnit betragte estimater for den relative værdi af y i forhold til x for populationen:

$$\beta = \xi_y/\xi_x \tag{1.7.1}$$

1.7.1 Det simple kvotientskøn

Vi vil indledningsvis belyse egenskaberne ved det simple skøn $\widehat{\beta}$, over forholdet $\beta = \Sigma y_i / \Sigma x_i$, bestemt ved

$$\widehat{\beta} = \overline{Y}_\cdot / \overline{X}_\cdot = \sum_{\nu=1}^{n} Y_\nu / \sum_{\nu=1}^{n} X_\nu \tag{1.7.2}$$

Estimatoren kaldes ofte kvotientskønnet.

Sætning 1.7.1 ***Forventningsværdi af kvotientskønnet***

For simpel tilfældig udvælgelse gælder:

$$\mathsf{E}[\widehat{\beta}] = \beta - \mathsf{COV}[\widehat{\beta}, \overline{X}_\cdot]/\xi_x \tag{1.7.3}$$

Bevis:
Relationen fås ved at bemærke, at

$$\mathsf{COV}[\widehat{\beta}, \overline{X}_\cdot] = \mathsf{E}[\widehat{\beta}\,\overline{X}_\cdot] - \mathsf{E}[\widehat{\beta}]\,\mathsf{E}[\overline{X}_\cdot] = \xi_y - \mathsf{E}[\widehat{\beta}]\,\xi_x$$

□

Bemærkning 1 ***Skævhed af kvotientskønnet***

Vi ser at kvotientskønnet ikke er centralt. Skævheden er $-\mathsf{COV}[\widehat{\beta}, \overline{X}_\cdot]/\xi_x$.

Idet kovariansen mellem $\widehat{\beta}$ og $\overline{X}_\cdot$ kan udtrykkes som produktet af korrelationskoefficienten ρ (mellem $\widehat{\beta}$ og $\overline{X}_\cdot$) og de to spredninger $\sigma_{\widehat{\beta}}$ og $\sigma_{\overline{X}}$ ser man, at man kan udtrykke den relative skævhed af kvotientskønnet som

$$\frac{\mathsf{E}[\widehat{\beta}] - \beta}{\beta} = -\rho\gamma_1\gamma_2$$

hvor γ_1 og γ_2 betegner de relative spredninger for henholdsvis $\widehat{\beta}$ og $\overline{X}_\cdot$, ser man, at hvis stikprøvestørrelsen er passende stor, vil de relative spredninger γ_1 og γ_2 være små. Da korrelationskoefficienten $\rho_{\widehat{\beta},\overline{X}}$ er numerisk begrænset af 1, vil skævheden aftage med voksende stikprøvestørrelse. □

En mere eksplicit vurdering af skævheden er givet i nedenstående

Sætning 1.7.2 *Skævhed og varians for kvotientskøn*

For simpel tilfældig udvælgelse uden tilbagelægning gælder for det simple kvotientskøn $\widehat{\beta}$ bestemt ved (1.7.2)

$$\mathsf{E}\,[\,\widehat{\beta}\,] \simeq \beta\{1+\frac{1}{n}[\gamma_x^2-\gamma_{x,y}](1-f)\} \tag{1.7.4}$$

$$\mathsf{V}\,[\,\widehat{\beta}\,] \simeq \frac{\beta^2}{n}[\gamma_x^2+\gamma_y^2-2\gamma_{x,y}](1-f)\,, \tag{1.7.5}$$

hvor f som vanligt betegner udvalgsbrøken, $f = n/N$, og variationskoefficienterne, γ_x og γ_y er bestemt ved (1.2.5) og den relative populationskovarians $\gamma_{x,y}$ er bestemt ved (1.2.10)

Såfremt udvælgelsen foretages med tilbagelægning bortfalder faktoren $(1-f)$ i ovenstående udtryk.

Bevis:
Resultatet vises ved at udvikle udtrykket

$$\widehat{\beta} = \overline{Y}./\overline{X}. = (\xi_y/\xi_x)\left(1+\frac{\overline{Y}.-\xi_y}{\xi_y}\right)\left(1+\frac{\overline{X}.-\xi_x}{\xi_x}\right)^{-1}$$

i en Taylorrrække, se f.eks. Tin (1965). □

Bemærkning 1 *Skævheden af kvotientskønnet udtrykt ved korrelationen mellem x og y*

Det følger af (1.7.4), at skævheden af kvotientskønnet approksimativt kan udtrykkes som

$$\mathsf{E}\,[\,\widehat{\beta}-\beta] \simeq \beta(\gamma_x^2-\gamma_{x,y})\frac{1-f}{n}$$

Idet $\gamma_{x,y} = \rho_{x,y}\gamma_x\gamma_y$, hvor $\rho_{x,y}$ angiver korrelationen mellem x og y (jvf. (1.2.11)) finder man, at skævheden kan udtrykkes som

$$\mathsf{E}\,[\,\widehat{\beta}-\beta] \simeq \frac{1-f}{n\xi_x}(\beta\sigma_x^2-\rho_{x,y}\sigma_x\sigma_y)\,. \tag{1.7.6}$$

□

Bemærkning 2 *Skævheden af kvotientskønnet udtrykt ved regressionskoefficienten svarende til den sædvanlige regression*

Lad β^* angiver regressionskoefficenten svarende til den sædvanlige lineære regression af y på x

$$\beta^* = \sum(x_i - \xi_x)(y_i - \xi_y) / \sum(x_i - \xi_x)^2 = \sigma_{x,y}/\sigma_x^2$$

Man har da af (1.7.4), (se f.eks. (1.2.21)), at skævheden af kvotientskønnet approksimativt kan udtrykkes som

$$\mathsf{E}\,[\,\widehat{\beta} - \beta] \simeq \gamma_x^2(\beta - \beta^*)\,\frac{1-f}{n}\,. \tag{1.7.7}$$

Skævheden er altså nul, hvis $\beta = \beta^*$, dvs. hvis hældningen for linien gennem origo og tyngdepunktet netop er den samme som hældningen for regressionen af y på x.

Jo mere disse to hældninger afviger fra hinanden, desto større er skævheden af kvotientskønnet. □

Bemærkning 3 Middelkvadratafvigelsen for kvotientskøn

Skævheden $\mathsf{E}\,[\,\widehat{\beta} - \beta]$ af kvotientskønnet er af størrelsesordenen $1/n$, hvorfor den kvadratiske afstand $\mathsf{E}\,[\,(\widehat{\beta} - \beta)^2]$ er af størrelsesordenen $1/n^2$. Da variansen $\mathsf{V}\,[\,\widehat{\beta}]$ er af størrelsesordenen $1/n$, gælder det derfor at middelkvadratafvigelsen

$$\mathsf{E}\,[\,(\widehat{\beta} - \beta)^2] = (\mathsf{E}\,[\,\widehat{\beta}] - \beta)^2 + \mathsf{V}\,[\,\widehat{\beta}]$$

vil være domineret af variansen, $\mathsf{V}\,[\,\widehat{\beta}]$, hvorfor man i almindelighed ser bort fra skævheden af skønnet. □

Bemærkning 4 *Estimation af variansen for kvotientskønnet ved residualvarians for linien gennem origo og tyngdepunktet*

Vi bemærker, at udtrykket (1.7.5) for variansen på $\widehat{\beta}$ kan omskrives til

$$\mathsf{V}\,[\,\widehat{\beta}] \simeq \frac{1}{n\xi_x^2}[\sigma_y^2 + \sigma_x^2\beta^2 - 2\beta\sigma_{x,y}](1-f) = \frac{1}{n\xi_x^2}\,\sigma_d^2(1-f)$$

hvor σ_d^2 angiver populationsvariansen for størrelsen $d_i = y_i - \beta x_i$, jvf (1.2.23) og f betegner udvalgsbrøken, $f = n/N$.

Man kan derfor bestemme et skøn over variansen på $\widehat{\beta}$ ved at benytte residualvariansen omkring linien gennem origo og tyngdepunktet:

$$\widehat{\mathrm{V}}\,[\,\widehat{\beta}] = \frac{1-f}{n\xi_x^2}\, s_d^2 \tag{1.7.8}$$

hvor

$$s_d^2 = \frac{1}{n-1}\sum_{\nu=1}^{n}(Y_\nu - \widehat{\beta}X_\nu)^2 \tag{1.7.9}$$

og f angiver udvalgsbrøken, $f = n/N$.

Såfremt populationsmiddelværdien ξ_x af x ikke er kendt på forhånd, estimeres den ved stikprøvegennemsnittet $\overline{X}.$, og skønnet over variansen på $\widehat{\beta}$ bliver i dette tilfælde

$$\widehat{\mathrm{V}}\,[\,\widehat{\beta}] = \frac{1-f}{n\overline{X}_.^2}\, s_d^2 \tag{1.7.10}$$

hvor s_d^2 er givet ved (1.7.9) □

Eksempel 1.7.1 ***Effekt af reklamekampagne***

I en butikskæde, der omfatter 500 detailbutikker, har man gennemført en større reklamekampagne for en bestemt mærkevare, der sælges i 500-g pakninger.

Man udvalgte 10 butikker ved simpel tilfældig udvælgelse. For hver butik opgjorde man salget x_i (antal pakninger) i de to uger, der gik forud for kampagnen samt salget y_i i de to uger, der fulgte kampagnen.

Salgstallene for de 10 butikker er anført i nedenstående tabel.

Lbnr	1	2	3	4	5	6	7	8	9	10
Før x_i	334	270	309	312	233	379	425	269	298	253
Efter y_i	566	290	470	481	333	444	490	420	446	349

Det gennemsnitlige salg for de 10 butikker før kampagnen var $\bar{x} = 308.2$ pakker, og det gennemsnitlige salg efter kampagnen var $\bar{y} = 428.9$ pakker.

Skønnet over omsætningsændringen bliver da

$$\widehat{\beta} = \frac{428.9}{308.2} = 1.39$$

altså en skønnet øgning af kædens samlede salg på 39 %.

Afvigelsen d_i mellem den observerede y_i og den tilpassede værdi $\widehat{\beta}x_i$ for de 10 butikker er vist i nedenstående tabel:

Lbnr	1	2	3	4	5
d_i	101.20	-85.74	39.99	46.81	8.75
Lbnr	6	7	8	9	10
d_i	-83.43	-101.44	45.65	31.29	-3.08

Man finder da (jvf. (1.7.9))

$$s_d^2 = \frac{\sum D_\nu^2}{9} = \frac{41782.84}{9} = 4642.54$$

Variansen på $\widehat{\beta}$ estimeres ved (1.7.10). Idet $n = 10$, $N = 500$ har man $f = n/N = 10/500 = 0.02$, hvorfor man har

$$\widehat{\mathrm{V}}\,[\,\widehat{\beta}] = \frac{0.98}{10 \times 308.2^2}\, 4642.54 = 0.004790 = (0.069)^2$$

Da $u_{0.975} = 1.96$ får man et approksimativt 95 % konfidensinterval for omsætningsændringen som $\widehat{\beta} \pm 1.96 \times 0.069 = 1.39 \pm 1.96 \times 0.069 = 1.39 \pm 0.14$, dvs. intervallet (1.26 ; 1.53).

Approksimationen skal dog tages med noget forbehold, dels på grund af den ringe stikprøvestørrelse og dels på grund af den relativt store variationskoefficient (ca. 19 %) såvel for x som for y. □

1.7.2 Korrigerede kvotientskøn

Sædvanligvis vælger man at se bort fra kvotientskønnets skævhed. I den aktuelle situation kan man evt forsøge at vurdere skævheden ved at indsætte estimater for de parametre, der indgår i udtrykket (1.7.6) for skævheden.

Den statistiske litteratur indeholder en række forslag til eliminering af skævheden på det simple kvotientskøn $\widehat{\beta}$ (1.7.2)

I nedenstående sætninger beskriver vi to af disse forslag:

Sætning 1.7.3 ***Korrigeret kvotientskøn***

Skønnet

$$\widehat{\beta}_1 = \widehat{\beta} + \widehat{\sigma}_{b,x}/(n\xi_x)\ , \tag{1.7.11}$$

hvor $\widehat{\beta}$ er givet ved (1.7.2), er asymptotisk central for den relative værdi β af y i forhold til x for populationen.

Variansen $\mathsf{V}\,[\,\widehat{\beta}_1]$ er givet ved (1.7.5). Et skøn over variansen fås derfor ved (1.7.8).

Bevis:
Fås ved at benytte relationen (1.7.3).

□

Man kan også bare korrigere skønnet $\overline{B}.$ som anført i nedenstående sætning:

Sætning 1.7.4 ***Hartley-Ross estimatoren***

En approksimativt central estimator for den relative værdi β af y i forhold til x for populationen er Hartley-Ross estimatoren:

$$\widehat{\beta}_2 = \overline{B}. + \frac{N-1}{N}\,\widehat{\sigma}_{b,x}/\xi_x \tag{1.7.12}$$

med

$$\mathsf{V}\,[\,\widehat{\beta}_2] \simeq \frac{1}{n}\,(1-f)\,[\xi_b^2\,\gamma_x^2 + \sigma_y^2/\xi_x^2 - 2\xi_b\sigma_{x,y}/\xi_x] \tag{1.7.13}$$

hvor variationskoefficienterne, γ_x bestemt ved (1.2.5) og $f = n/N$

Bevis:
Relationen (1.2.18) mellem ξ_b og β viser, at

$$\mathsf{E}\,[\,\overline{B}.] = \beta - \sigma_{b,x}/\xi_x \tag{1.7.14}$$

For bestemmelse af variansen se f.eks. Kendall & Stuart (1983)

□

Bemærkning 1 ***Skøn over variansen på Hartley-Ross estimatoren***

Vi bemærker, at variansen $\mathsf{V}\,[\,\widehat{\beta}_2]$ kan udtrykkes som

$$\mathsf{V}\,[\,\widehat{\beta}_2] \simeq \frac{1}{n\xi_x^2}\,\mathsf{V}\,[\,Y - \xi_b X]$$

hvorfor et estimat for variansen fås som

$$\widehat{\mathsf{V}}\,[\,\widehat{\beta}_2] \simeq \frac{1}{n\xi_x^2}\,(1-f)\,\frac{1}{n-1}\,\sum_1^n [(Y_\nu - \overline{Y}.) - \overline{B}.(X_\nu - \overline{X}.)]^2 \tag{1.7.15}$$

□

Bemærkning 2 ***Sammenligning mellem det korrigerede kvotientskøn og Hartley-Ross estimatoren***

Forskellen mellem variansen på de to estimater $\widehat{\beta}_1$ og $\widehat{\beta}_2$ er

$$\mathsf{V}\,[\,\widehat{\beta}_1] - \mathsf{V}\,[\,\widehat{\beta}_2] \simeq \frac{1}{n}\,\gamma_x^2[(\beta - \sigma_{x,y}/\sigma_x^2)^2 - (\xi_b - \sigma_{x,y}/\sigma_x^2)^2]$$

Fortegnet for forskellen afhænger således af om regressionskoefficienten $\sigma_{x,y}/\sigma_x^2$ for populationens y-værdier mod x-værdierne ligger nærmest ved β (da har $\widehat{\beta}_1$ den mindste varians), eller om den ligger nærmere ved den gennemsnitlige rate pr enhed, ξ_b (da har $\widehat{\beta}_2$ den mindste varians). □

1.7.3 Kvotientskøn for populationsgennemsnittet

Selv om formålet med en stikprøveundersøgelse alene er at få kendskab til populationsgennemsnittet (eller totalen) af en enkelt variabel y, kan

det i en række situationer være fordelagtigt at udnytte kendskabet til en hjælpevariabel x, der er knyttet til analyseenheden, og som er beslægtet med interessevariablen y, og som derfor indeholder en vis information om denne.

I dette afsnit vil vi betragte en situation, hvor der foreligger en sådan hjælpevariabel x, og hvor populationsgennemsnittet ξ_x af denne variabel er kendt. Vi skal i de følgende afsnit se andre eksempler på udnyttelse af hjælpevaraiable.

Når populationsgennemsnittet ξ_x af x'erne er kendt, er det nærliggende at søge efter en omsætningsfaktor β^K, der fører populationsgennemsnittet ξ_x for x'erne over i populationsgennemsnittet ξ_y for y'erne, dvs. at estimere kvotienten $\beta^K = \xi_y/\xi_x$ ved hjælp af stikprøven, og derefter udnytte relationen

$$\xi_y = \beta^K \xi_x$$

til at estimere ξ_y.

Egenskaberne for kvotientskønnet over β^K overføres umiddelbart til skønnet

$$\hat{\xi}_y^K = \widehat{\beta}^K \xi_x \tag{1.7.16}$$

eller det tilsvarende estimat

$$\hat{\zeta}_y^K = N\hat{\xi}_y^K \tag{1.7.17}$$

for populationstotalen.

Skønnene (1.7.16) og (1.7.17) kaldes ofte for kvotientskøn. Hvis man vil præcisere at man estimerer ξ_y (eller ζ_y) siger man at man benytter et kvotientskøn for populationsgennemsnittet (eller populationstotalen).

Der gælder således

Sætning 1.7.5 ***Approksimative momenter for kvotientskøn for populationsmiddel og populationstotal***

For simpel tilfældig udvælgelse uden tilbagelægning har kvotientskønnene

$\hat{\xi}_y^K$ og $\hat{\zeta}_y^K$ givet ved (1.7.16) og (1.7.17) de approksimative middelværdier

$$\mathsf{E}\,[\hat{\xi}_y^K] \simeq \xi_y\{1+\frac{1}{n}[\gamma_x^2-\gamma_{x,y}](1-f)\}$$

$$\mathsf{E}\,[\hat{\zeta}_y^K] \simeq N\xi_y\{1+\frac{1}{n}[\gamma_x^2-\gamma_{x,y}](1-f)\}$$

og de approksimative varianser

$$\mathsf{V}\,[\hat{\xi}_y^K] \simeq \frac{\sigma_d^2}{n}(1-f) \tag{1.7.18}$$

$$\mathsf{V}\,[\hat{\zeta}_y^K] \simeq N^2\frac{\sigma_d^2}{n}(1-f) \tag{1.7.19}$$

hvor hvor f angiver udvalgsbrøken, $f = n/N$, og σ_d^2 angiver populationsvariansen for $d_i = y_i - \beta^K x_i$ se (1.2.23).

Residualvariansen σ_d^2 estimeres ved

$$s_d^2 = \frac{1}{n-1}\sum_{\nu=1}^{n}(Y_\nu - \widehat{\beta}X_\nu)^2 \tag{1.7.20}$$

Den relative varians for skønnene $\hat{\xi}_y^K$ og $\hat{\zeta}_y^K$ er

$$\frac{\mathsf{V}\,[\hat{\xi}_y^K]}{\xi_y^2} \simeq \frac{1-f}{n}\,[\gamma_x^2+\gamma_y^2-2\gamma_{x,y}]\,, \tag{1.7.21}$$

Såfremt udvælgelsen foretages med tilbagelægning bortfalder faktoren $(1-f)$ i ovenstående udtryk.

Bevis:
Resultatet følger umiddelbart af sætning 1.7.2 og bemærkning 4 til sætningen. □

Bemærkning 1 *Sammenligning mellem kvotientskøn og direkte estimation ved stikprøvegennemsnittet*

Det er imidlertid ikke umiddelbart givet, at kvotientskønnet (1.7.16) for populationsgennemsnittet ξ_y af y'erne er at foretrække frem for det enkle estimat ved stikprøvegennemsnittet $\overline{Y}_.$.

Idet vi tillader os at se bort fra betydningen af kvotientskønnets skævhed, vil vi sammenligne de to estimater ved at sammenligne varianserne i fordelingen af estimatet.

Udtrykket (1.7.18) for den approksimative varians for kvotientskønnet (1.7.16) kan omformes til

$$\mathsf{V}\,[\,\hat{\xi}_y^K] \simeq \frac{1-f}{n}[\sigma_y^2 + (\beta^K)^2\sigma_x^2 - 2\beta^K\rho_{x,y}\sigma_x\sigma_y]\,,$$

Variansen for stikprøvegennemsnittet $\overline{Y}_.$ fås af (1.4.3)

$$\mathsf{V}\,[\overline{Y}_.] = \frac{1-f}{n}\,\sigma_y^2$$

hvor vi har tilladt os at se bort fra forskellen mellem σ_y og σ_y'

Betingelsen for at variansen for kvotientskønnet er mindre end variansen for det simple gennemsnit er således, at

$$(\beta^K)^2\sigma_x^2 - 2\beta^K\rho_{x,y}\sigma_x\sigma_y < 0$$

dvs. at

$$\rho_{x,y} > \frac{\beta^K}{2}\,\frac{\sigma_x}{\sigma_y} = \frac{1}{2}\,\frac{\gamma_x}{\gamma_y} \qquad (1.7.22)$$

Der skal således være en tilstrækkelig stærk sammenhæng (stor korrelation) mellem værdierne af y og værdierne af hjælpevariablen x for at man med fordel kan udnytte værdierne af x til estimation af ξ_y.

Det ses, at usikkerheden på nævneren, udtrykt ved den relative spredning γ_x, er en væsentlig størrelse. Hvis den relative spredning af x-værdierne

er mere end dobbelt så stor som den relative spredning af y-værdierne, udtrykker kravet (1.7.22) at $\rho > 1$, hvilket ikke kan opfyldes, da korrelationskoefficienten som bekendt højst kan blive 1.

I mange situationer, hvor x og y er beslægtede variable, vil de relative spredninger for x og y ofte være omtrent lige store. I sådanne tilfælde vil kvotientskønnet altså være at foretrække for det direkte gennemsnit $\overline{Y}_.$, hvis korrelationen ρ mellem x og y er større end 0.50. □

Eksempel 1.7.2 ***Estimation af omsætning i en butikskæde***

I en butikskæde med 500 detailbutikker har man opgjort den gennemsnitlige omsætning pr. butik i året 1993 for samtlige butikker til 17.5 [Mio kr].

Ved udgangen af januar kvartal 1994 ønsker man at vurdere den samlede omsætning i dette kvartal.

Der udvælges 10 butikker ved simpel tilfældig udvælgelse og for hver af disse opgør man omsætningen y_i i januar kvartal. Endvidere registrerer man butikkens samlede omsætning x_i i 1993.

Værdierne (i [Mio kr]) er anført i nedenstående tabel.

Lbnr	1	2	3	4	5
1993	17.542	18.699	17.683	17.713	17.073
1 kvt 1994	5.028	6.046	5.607	5.367	4.529
Lbnr	6	7	8	9	10
1993	18.609	17.144	18.383	18.128	18.868
1 kvt 1994	5.658	5.243	5.975	5.187	5.792

Man har $\overline{Y} = 5.4432$ [Mio kr] og $\overline{X} = 17.9842$ [Mio kr], dvs.
$\widehat{\beta}^K = 5.4432/17.9842 = 0.3027$.

Kvotientskønnet over kædens samlede omsætning i januar kvartal er

$$\hat{\zeta}^K = 500 \times 0.3027 \times 17.500 = 2\,648.63 \text{ [Mio kr]}$$

For at vurdere usikkerheden på dette skøn bestemmes

$$s_d^2 = \frac{\sum D_\nu^2}{9} = \frac{0.970203}{9} = 0.10784 = (0.3284[\text{Mio kr}])^2$$

Skønnet over variansen (1.7.19) på $\hat{\zeta}_y^K$ bliver derfor

$$\hat{V}[\hat{\zeta}_y^K] = \frac{500^2 \times 0.98}{10} \times 0.10784 = 2642.080 = (51.4 \text{ [Mio kr]})^2$$

dvs. at den relative spredning på skønnet er

$$\frac{\sqrt{\hat{V}[\hat{\zeta}_y^K]}}{\hat{\zeta}_y^K} = \frac{51.4}{2\ 648.63} = 0.019 = 1.9\ \%$$

Et 95 % konfidensinterval for det samlede salg i første kvartal er

$$2\ 648.63 \pm 1.96 \times 51.4 = (2548 \text{ [Mio kr]}; 2749 \text{ [Mio kr]})$$

En naturlig del af undersøgelsen er at kontrollere, at kvotientskønnet har mening, dvs. at skønnet har mindre varians end det skøn, der blot er baseret på den gennemsnitlige omsætning i januar kvartal for de 10 butikker.

I en situation som denne, hvor data er indsamlet, kan man sammenligne estimatet for variansen for kvotientskønnet med estimatet for variansen på $500 \times \overline{Y}$. Idet

$$\sum_{\nu=1}^{10}(Y_\nu - \overline{Y}.)^2 = 1.9605$$

har man

$$\hat{\sigma}_y^2 = \frac{1.9605}{9} = 0.2178 = (0.467 \text{ [Mio kr]})^2$$

hvorfor

$$\hat{V}[500 \times \overline{Y}] = 500^2 \times \frac{0.2178}{10} = (73.79 \text{ [Mio kr]})^2$$

dvs. en væsentlig større varians end $\hat{V}[\hat{\zeta}_y^K] = (51.4 \text{ [Mio kr]})^2$.

Havde man i stedet vurderet om betingelsen (1.7.22) er opfyldt, ville resultatet naturligvis også her være til kvotientskønnets fordel. Den estimerede korrelation mellem x og y er

$$\hat{\rho} = \frac{\sum(X_\nu - \overline{X})(Y_\nu - \overline{Y})}{\sqrt{\sum(X_\nu - \overline{X})^2}\sqrt{\sum(Y_\nu - \overline{Y})^2}} = 0.81$$

Stikprøveestimatet for den relative spredning på årsomsætningen i 1993 er $\hat{\gamma}_x = \hat{\sigma}_x/\overline{X} = 0.0369 = 3.69$ % og stikprøveestimatet $\hat{\gamma}_y$ for den relative spredning på kvartalsomsætningen i 1994 er $\hat{\gamma}_y = 0.0857 = 8.57$ %.

Man får derfor skønnet $\hat{\gamma}_x/(2\hat{\gamma}_y) = 0.21$. Korrelationen mellem x og y er altså stærkere end kravet i (1.7.22). □

1.7.4 Regressionsskøn for populationsgennemsnittet

Kvotientskønnet er orienteret mod en situation, hvor relationen mellem interessevariablen y og hjælpevariablen x tilnærmelsesvist er en proportionalitet, sådan at man kan benytte stikprøveresultatet til en multiplikativ fremskrivning af ξ_x.

Undertiden er denne tilnærmelse for grov til at man kan opnå en effektiv udnyttelse af informationen i hjælpevariablen x.

Man kan da overveje at benytte et regressionsskøn.

Ideen bag regressionsskønnet bygger på en beskrivelse af sammenhængen mellem populationsværdierne ved parametrene i en regressionsmodel således som den blandt andet kendes fra Introduktion til Statistik, Bind 1.

Det vides således fra Introduktion til Statistik, Bind 1, at kvadratafvigelsessummen

$$S(a,b) = \sum_{i=1}^{N}[y_i - a - b(x_i - \xi_x)]^2$$

minimeres for $a = \xi_y$ og $b = \beta^R$, hvor regressionskoefficienten β^R for populationen er givet ved (1.2.20)

$$\beta^R = \frac{\sum(x_i - \xi_x)(y_i - \xi_y)}{\sum(x_i - \xi_x)^2} = \frac{\sigma_{x,y}}{\sigma_x^2} = \rho\sigma_y/\sigma_x \tag{1.7.23}$$

Regressionslinien for populationen

$$y = \xi_y + \beta^R(x - \xi_x)$$

er således fastlagt ved mindste kvadraters metode, uden nogen antagelse om fordelingen af populationsværdierne.

Betragt nu den endimensionale størrelse e_i, bestemt som residualet svarende til den i'te enhed ved prædiktion af y_i med det tilsvarende liniepunkt,

$$y_i = \xi_y + \beta^R(x_i - \xi_x) + e_i \tag{1.7.24}$$

Populationsmiddelværdien for e_i er

$$\xi_e = \frac{1}{N}\sum_{i=1}^{N} e_i = \frac{1}{N}\sum_{i=1}^{N}[y_i - \xi_y - \beta^R(x_i - \xi_x)] = \xi_y - \xi_y - \beta^R(\xi_x - \xi_x) = 0$$

og populationsvariansen for e_i er

$$\sigma_e^2 = \frac{1}{N}\sum_{i=1}^{N} e_i^2 = \frac{1}{N}\sum_{i=1}^{N}[y_i - \xi_y - \beta^R(x_i - \xi_x)]^2 \tag{1.7.25}$$

Idet

$$\begin{aligned} SAK_e &= \sum_{i=1}^{N}[y_i - \xi_y - \beta^R(x_i - \xi_x)]^2 = \sum_{i=1}^{N}(y_i - \xi_y)^2 - (\beta^R)^2\sum_{i=1}^{N}(x_i - \xi_x)^2 \\ &= (1-\rho^2)\sum_{i=1}^{N}(y_i - \xi_y)^2 \end{aligned}$$

hvor ρ betegner korrelationskoefficienten (1.2.9) mellem x og y i populationen, finder man udtrykket for populationsvariansen af residualerne

$$\sigma_e^2 = \sigma_y^2(1-\rho^2) \tag{1.7.26}$$

Sætning 1.7.6 ***Regressionsskøn med kendt regressionskoefficient*** β^R

Antag at regressionskoefficienten β^R for populationen er kendt.

Lad $(X_1, Y_1), \ldots, (X_n, Y_n)$ angive værdierne af de to variable i en stikprøve på n enheder udtaget ved simpel tilfældig udvælgelse uden tilbagelægning

og lad $\overline{X}.$ og $\overline{Y}.$ angive de tilsvarende gennemsnit.

Skønnet

$$\hat{\xi}_y^R = \overline{Y}. - \beta^R(\overline{X}. - \xi_x) \tag{1.7.27}$$

kaldes regressionsskønnet med kendt regressionskoefficient.

Der gælder

$$\begin{aligned} \mathsf{E}\,[\hat{\xi}_y^R] &= \xi_y \\ \mathsf{V}\,[\hat{\xi}_y^R] &= \frac{\sigma_e^2}{n}(1-f) \end{aligned} \tag{1.7.28}$$

hvor

$$\sigma_e^2 = \sigma_y^2(1-\rho^2) \tag{1.7.29}$$

angiver populationsvariansen (1.7.25) af residualerne svarende til regressionslinien for populationen og ρ angiver korrelationskoefficienten (1.2.9) mellem x og y.

Såfremt ρ er kendt, benyttes relationen (1.7.29) til estimation af residualvariansen variansen σ_e^2 ved

$$\widehat{\sigma}_e^2 = (1-\rho^2)S_y^2 \tag{1.7.30}$$

hvor

$$S_y^2 = \frac{1}{n-1}\sum_{\nu=1}^{n}(Y_\nu - \overline{Y}.)^2$$

angiver skønnet over populationsvariansen σ_y^2 af interessevariablen y.

Såfremt ρ ikke er kendt, estimeres σ_e^2 ved

$$S_e^2 = \frac{1}{n-1}\sum_{\nu=1}^{n}[Y_\nu - \overline{Y}_\cdot - \beta^R(X_\nu - \overline{X}_\cdot)]^2 \qquad (1.7.31)$$

Såvel estimatet (1.7.27) for populationssmiddelværdien, som estimaterne (1.7.30) og (1.7.31) for variansen er centrale.

Bevis:

Følger direkte □

Eksempel 1.7.3 ***Regressionsskøn for omsætning i en butikskæde ved brug af kendt regressionskoefficient***

Vi betragter atter situationen fra eksempel 1.7.2 på side 49, men vi antager nu, at man fra parallelle undersøgelser i lignende kæder mener at vide, at regressionskoefficienten $\beta^R = 0.50$. Endvidere kender man den gennemsnitlige årsomsætning pr butik i 1993 $\xi_x = 17.5$ [Mio kr].

Omsætningen i første kvartal 1994 samt årsomsætningen i 1993 er angivet i tabellen på side 49.

Man fandt stikprøveresultatet $\overline{Y} = 5.4432$ [Mio kr] og $\bar{x} = 17.9842$ [Mio kr].

Regressionslinien, der sammenknytter omsætningen i januar kvartal med butikkens årsomsætning i 1993 har formen

$$y = \xi_y + 0.50(x - 17.5)$$

Indsætter man heri stikprøveresultatet $(\overline{X}, \overline{Y}) = (17.9842; 5.4432)$ finder man udtrykket for $\hat{\xi}_y^R$

$$\hat{\xi}_y^R = 5.4432 - 0.50(17.9842 - 17.5) = 5.201 \text{ [Mio kr]}$$

Da de udvalgte butikker havde en større gennemsnitlig årsomsætning i 1993, end kæden som helhed, skal den gennemsnitlige kvartalsomsætning i stikprøven reduceres for at kunne udtrykke gennemsnitsværdien for kæden.

Usikkerheden på skønnet bestemmes af (1.7.28). Residualerne

$$E_\nu = Y_\nu - \overline{Y}_\cdot - 0.50(X_\nu - \overline{X}_\cdot)$$

svarende til stikprøveværdierne er angivet i nedenstående tabel:

Lbnr	1	2	3	4	5
1993	17.542	18.699	17.683	17.713	17.073
1 kvt 1994	5.028	6.046	5.607	5.367	4.529
E_ν	-0.1941	0.2454	0.3144	0.0594	-0.4586
Lbnr	6	7	8	9	10
1993	18.609	17.144	18.383	18.128	18.868
1 kvt 1994	5.658	5.243	5.975	5.187	5.792
E_ν	-0.0976	0.2199	0.334	-0.3281	-0.0931

Man finder

$$\widehat{\sigma}_e^2 = \frac{1}{9}\sum_{\nu=1}^{10} E_i^2 = \frac{0.695274}{9} = 0.077253 = (0.2779\ [\text{Mio kr}])^2$$

således at

$$\widehat{\mathrm{V}}\,[\hat{\xi}_y^R] = \frac{0.98}{10} \times 0.077253 = 0.007571 = (0.0870\ [\text{Mio kr}])^2$$

Et 95 % konfidensinterval for den gennemsnitlige omsætning i kædens butikker i januar kvartal er således

$$5.201 \pm 1.96 \times 0.0870 = (5.031; 5.372)\ [\text{Mio kr}]$$

Værdierne for den totale omsætning i kædens 500 butikker fås ved multiplikation med 500, dvs. 95 % konfidensintervallet for det samlede salg bliver

$$500 \times (5.031; 5.372) = (2\,515; 2\,686)\ [\text{Mio kr}]$$

□

Når populationens regressionskoefficient β^R ikke kendes, kan man bruge stikprøveværdier til at estimere denne koefficient.

Sætning 1.7.7 *Approksimative momenter for regressionsskøn med estimeret regressionskoefficient* β^R

Lad $(X_1, Y_1), \ldots, (X_n, Y_n)$ angive værdierne af de to variable i en stikprøve på n enheder udtaget ved simpel tilfældig udvælgelse uden tilbagelægning

og lad $\overline{X}.$ og $\overline{Y}.$ angive de tilsvarende gennemsnit.

Skønnet

$$\hat{\xi}_y^R = \overline{Y}. - \hat{\beta}^R(\overline{X}. - \xi_x) \tag{1.7.32}$$

med

$$\hat{\beta}^R = \frac{\sum_\nu (X_\nu - \overline{X}.)(Y_\nu - \overline{Y}.)}{\sum_\nu (X_\nu - \overline{X}.)^2} \tag{1.7.33}$$

kaldes regressionsskønnet med estimeret regressionskoefficient.

Skønnet er ikke centralt. Skævheden for skønnet er af størrelsesordenen $1/\sqrt{n}$.

Variansen på skønnet er approksimativt

$$\mathrm{V}[\hat{\xi}_y^R] \simeq \frac{\sigma_e^2}{n}(1-f) \tag{1.7.34}$$

hvor $\sigma_e^2 = \sigma_y^2(1-\rho^2)$ angiver populationsvariansen (1.7.25) for residualerne svarende til regressionslinien for populationen, og ρ angiver korrelationskoefficienten (1.2.9) mellem x og y.

Som estimat for residualvariansen σ_e^2 benyttes

$$S_e^2 = \frac{1}{n-2}\sum_{\nu=1}^{n}[Y_\nu - \overline{Y}. - \hat{\beta}^R(X_\nu - \overline{X}.)]^2 \tag{1.7.35}$$

Bevis:
Overspringes, se f.eks. Cochran (1963). □

Bemærkning 1 ***Brug af sædvanlige edb-programmer til regressionsanalyse***

Skønnet (1.7.33) over regressionskoefficienten for populationen er det sædvanlige mindste kvadraters estimat. Skønnet kan derfor bestemmes ved brug af standardprogrammer til regressionsanalyse.

Skønnet (1.7.35) over residualvariansen er ligeledes det sædvanlige skøn svarende til variationen omkring regressionslinien. Det bemærkes, at skønnet er baseret på $n-2$ frihedsgrader (i modsætning til skønnet (1.7.31), hvor koefficienten β^R var kendt og antallet af frihedsgrader derfor er $n-1$).

I de sædvanlige regressionsanalysemodeller antages de uafhængige variable x_i at være faste. I disse modeller er skønnene over regressionskoefficient og residualvarians derfor centrale.

I skønnet (1.7.33) over regressionskoefficienten opfattes de observerede værdier af hjælpevariablen som stokastiske variable (da de afhænger af hvilke analyseenheder, der indgår i stikprøven). Variationen af skønnet (1.7.33) er således også påvirket af variationen i nævneren, og skønnet er derfor sædvanligvis ikke centralt. □

Eksempel 1.7.4 ***Regressionsskøn for omsætning i en butikskæde ved brug af estimeret regressionskoefficient***

Vi betragter atter situationen fra eksempel 1.7.2 og 1.7.3.

Vi antager nu, at man på grund af kædens specielle forhold foretrækker at estimere regressionskoefficienten β^R fra stikprøven.

Den gennemsnitlige årsomsætning pr butik i 1993 var $\xi_x = 17.5$ [Mio kr].

Omsætningen i første kvartal 1994 samt årsomsætningen i 1993 er angivet i tabellen på side 49.

Man finder estimatet (1.7.33) for regressionskoefficienten

$$\widehat{\beta}^R = \frac{2.2048}{3.7582} = 0.5867$$

Regressionslinien, der sammenknytter omsætningen i januar kvartal med butikkens årsomsætning i 1993 har formen

$$y = \xi_y + 0.5867(x - 17.5)$$

Ved indsættelse af stikprøveresultatet $(\bar{x}, \bar{y}) = (17.9842, 5.4432)$ i dette udtryk finder man udtrykket for $\hat{\xi}_y^R$

$$\hat{\xi}_y^R = 5.4432 - 0.5867(17.9842 - 17.5) = 5.159 \text{ [Mio kr]}$$

Da den estimerede regressionskoefficient, $\widehat{\beta}^R$, er større end den værdi, der blev benyttet i eksempel 1.7.3, er korrektionen også større.

Til bestemmelse af usikkerheden på skønnet bestemmer man residualerne

$$e_\nu = y_\nu - \bar{y} - 0.5867(x_\nu - \bar{x})$$

Lbnr	1	2	3	4	5
1993	17.542	18.699	17.683	17.713	17.073
1 kvt 1994	5.028	6.046	5.607	5.367	4.529
e_ν	-0.1558	0.1834	0.3405	0.0829	-0.3796

Lbnr	6	7	8	9	10
1993	18.609	17.144	18.383	18.128	18.868
1 kvt 1994	5.658	5.243	5.975	5.187	5.792
e_ν	-0.1518	0.2927	0.2978	-0.3406	-0.1697

Man finder

$$\hat{\sigma}_e^2 = \frac{1}{8}\sum_{\nu=1}^{10} e_i^2 = \frac{0.6671}{8} = 0.083382 = (0.2888 \text{ [Mio kr]})^2$$

således at

$$\hat{\mathrm{V}}\,[\hat{\xi}_y^R] = \frac{0.98}{10} \times 0.083382 = 0.008171 = (0.0903 \text{ [Mio kr]})^2$$

Til trods for at residualkvadratafvigelssummen er lidt mindre end kvadrafvigelsessummen omkring den linie, der blev bestemt ved antagelse af en kendt regressionskoefficient, bliver skønnet over residualvariansen alligevel lidt større, idet vi skal tage højde for den ekstra frihedsgrad, der svarer til estimation af liniens hældning.

Man finder således 95 % konfidensintervallet for den gennemsnitlige omsætning i kædens butikker i januar kvartal

$$5.159 \pm 1.96 \times 0.0903 = (4.982; 5.336) \text{ [Mio kr]}$$

Og endelig fås 95 % konfidensintervallet for den totale omsætning i kædens 500 butikker

$$500 \times (4.982; 5.336) = (2\,491; 2\,668) \text{ [Mio kr]}$$

□

1.7.5 Sammenligning mellem regressionsskøn, kvotientskøn og direkte estimation ved stikprøvegenemsnittet.

Såfremt stikprøven er tilstrækkelig stor, har man udtrykkene for de approksimative varianser for de tre estimater for populationsgennemsnittet ξ_y

$$\begin{aligned} \mathrm{V}[\hat{\xi}_y^R] &\simeq \frac{1-f}{n}\sigma_y^2(1-\rho^2) \\ \mathrm{V}[\hat{\xi}_y^K] &\simeq \frac{1-f}{n}(\sigma_y^2 + (\beta^K)^2\sigma_x^2 - 2\beta^K\rho\sigma_x\sigma_y) \\ \mathrm{V}[\overline{Y}.] &= \frac{1-f}{n}\sigma_y^2 \end{aligned}$$

Det ses, at med mindre korrelationen mellem x og y er nul, vil regressionsskønnet have en mindre varians end det simple gennemsnit. Variansen reduceres med faktoren $1-\rho^2$. Korrelationen skal således have en vis størrelse for at opnå en nævneværdig reduktion.

Vi har tidligere (i bemærkningen til sætning 1.7.5) set, at kvotientskønnet $\hat{\xi}_y^K$ kun er bedre end det simple gennemsnit hvis korrelationen mellem x og y er positiv og tilstrækkelig stærk (større end $\gamma_x/(2\gamma_y)$),

Betingelsen for at variansen på regressionsskønnet $\hat{\xi}_y^R$ vil være mindre end variansen på kvotientskønnet $\hat{\xi}_y^K$ er, at

$$-\rho^2\sigma_y^2 < (\beta^K)^2\sigma_x^2 - 2\beta^K\rho\sigma_x\sigma_y$$

dvs.

$$(\rho\sigma_y - \beta^K\sigma_x)^2 > 0 \quad \text{eller} \quad (\beta^K - \rho)^2 > 0$$

Disse betingelser er trivielt opfyldt med mindre $\beta^K = \rho$, dvs. med mindre regressionslinien for populationen netop er en linie gennem nulpunktet, og i så fald har de to estimater samme varians.

Sammenfatning

Det simple gennemsnit $\overline{Y}.$ har den fordel at det er nemt at regne ud, og nemt at forstå. Brug af skønnet kræver ikke kendskab til hjælpevariable og endelig er skønnet centralt, selv for små stikprøver.

Når der er mulighed for at registrere hjælpevariable, hvis populationsgennemsnit ξ_x er kendt, kan man overveje at korrigere det simple populationsgennemsnit ved at udnytte relationen mellem hjælpevariablen og interessevariablen.

Ideen bag kvotientskønnet er, at man omsætter populationsgennemsnittet ξ_x for hjælpevariablen til skønnet over populationsgennemsnittet for interessevariablen ved at estimere den relative værdi $\beta^K = \xi_y/\xi_x$ af interessevariablen for populationen ud fra stikprøven.

$$\hat{\xi}_y^K = \hat{\beta}^K \xi_x = (\overline{Y}./\overline{X}.)\xi_x$$

Skønnet har den intuitive fordel, at det udnytter kendskabet til hjælpevariablen. Omskrivningen

$$\hat{\xi}_y^K = \overline{Y}. \times (\xi_x/\overline{X}.)$$

viser, at skønnet kan opfattes som en multiplikativ korrektion af stikprøvegennemsnittet for y'erne med en faktor, der afhænger af forholdet mellem stikprøvegennemsnit og populationsgennemsnit for x'erne.

Skønnet er kun godt, hvis der er en tilstrækkelig stor positiv korrelation mellem y og x.

Skønnet er skævt, skævheden aftager med voksende stikprøvestørrelse.

Ideen bag regresssionsskønnet er, at man korrigerer populationsgennemsnittet $\overline{Y}$ for interessevariablen med en additiv korrektion, der afhænger af differensen mellem stikprøvegennemsnit og populationsgennemsnit for $x$erne.

$$\hat{\xi}_y^R = \overline{Y}. - \beta^R(\overline{X}. - \xi_x)$$

Skønnet er godt, hvis der er korrelation mellem y og x. (Korrelationen skal dog have en vis størrelse for at opnå en nævneværdig forbedring i forhold til det simple gennemsnit).

Skønnet er skævt, skævheden aftager med voksende stikprøvestørrelse.

Metode	estimat	varians	variansestimat	Ref.
Simpelt gennemsnit	$\hat{\xi}_y = \overline{Y}.$	$\frac{\sigma_y^2}{n}(1-f)$	$\widehat{\sigma}_y^2 = \frac{1}{n-1}\sum_{\nu=1}^{n}(Y_\nu - \overline{Y})^2$	Sætn. 1.4.1
Kvotient-skøn	$\hat{\xi}_y^K = \widehat{\beta}^K \xi_x$ $\widehat{\beta}^K = \overline{Y}./\overline{X}.$	$\frac{\sigma_d^2}{n}(1-f)$	$\widehat{\sigma}_d^2 = \frac{1}{n-1}\sum_{\nu=1}^{n} D_\nu^2$	Sætn. 1.7.5
Regr. skøn	$\hat{\xi}_y^R = \overline{Y}. - \widehat{\beta}^R(\overline{X}. - \xi_x)$ $\widehat{\beta}^R = \frac{\sum(Y_\nu - \overline{Y}.)(X_\nu - \overline{X}.)}{\sum(X_\nu - \overline{X}.)^2}$	$\frac{\sigma_e^2}{n}(1-f)$ $= \frac{\sigma_y^2}{n}(1-\rho^2)(1-f)$	$\widehat{\sigma}_e^2 = \frac{1}{n-1}\sum_{\nu=1}^{n} E_\nu^2$	Sætn. 1.7.7

Tabel 1.1. Oversigt over estimater for populationsmiddelværdi ξ_y

Populationsstørrelse N, stikprøvestørrelse n, $\quad f = n/N \quad .$

$$
\begin{aligned}
\beta^K &= \xi_y/\xi_x\,; \qquad \beta^R = \sigma_{x,y}/\sigma_x^2 \\
\overline{Y} &= \frac{1}{n}\sum_{\nu=1}^{n} Y_\nu\,; \qquad \overline{X} = \frac{1}{n}\sum_{\nu=1}^{n} X_\nu\,, \\
E_\nu &= Y_\nu - \overline{Y}. - \widehat{\beta}^R(X_\nu - \overline{X}.) \\
D_\nu &= Y_\nu - \widehat{\beta}^K X_\nu
\end{aligned}
$$

1.8 Udvælgelse med varierende sandsynligheder

fil: repr4.tex 1998-01-24

1.8.1 Indledning

Hidtil har vi betragtet simpel tilfældig udvælgelse, dvs situationer, hvor alle analyseenheder har samme chance for at indgå i stikprøven.

I dette afsnit vil vi betragte den generelle situation, hvor udvælgelsessandsynlighederne for de enkelte analyseenheder kan være forskellige. Sådanne situationer kan undertiden forekomme, når en af de variable, der er tilknyttet analyseenheden, udtrykker en form for størrelse længde, areal, antal personer el. lign) for analyseenheden.

For at undgå de massive formuleringsproblemer, der optræder ved udvælgelse uden tilbagelægning, vil vi i dette afsnit antage at udvælgelse foretages med tilbagelægning, med mindre andet udtrykkeligt angives (sætning 1.8.5).

I sætning 1.8.1 illustreres skævheden af det sædvanlige gennemsnit når sandsynligheden for at inddrage en analyseenhed afhænger af interessevariablen.

Sætning 1.8.2 anviser hvorledes man kan korrigere observationerne med udvalgssandsynlighederne og opnå et centralt estimat for populationsmiddelværdien, og endelig viser sætning 1.8.3 hvorledes man ved et bevidst valg af udvalgssandsynlighederne kan opnå en effektiv estimation af den relative værdi β^R for populationen. Sætning 1.8.5 supplerer dette resultat for den mere realistiske situation, hvor udvælgelsen foregår uden tilbagelægning.

1.8.2 Fordelingsforhold ved udvælgelse med varierende sandsynligheder

Vi antager at populationen består af enhederne $\{x_1, x_2, \ldots, x_N\}$, samt at stikprøveudtagningen foretages således at sandsynligheden for at enhed i udvælges i den ν 'te trækning er $\mathsf{P}\,[I_\nu = i] = p_i; \quad i = 1, 2, \ldots, N.$

Såfremt stikprøven $X_1, X_2, \ldots, X_n$ vælges således at $\mathsf{P}\,[I_\nu = i] = p_i$, da gælder

$$\mathsf{E}\,[\,\overline{X}.] = \sum_{i=1}^{N} p_i x_i \tag{1.8.1}$$

Stikprøvegennemsnittet $\overline{X}.$ er altså i almindelighed ikke centralt for populationsmiddelværdien $\xi_x = \sum_1^N x_i/N$. Man kan imidlertid angive simple udtryk for forventningsværdien af stikprøvegennemsnittet i en række specielle situationer.

Vi betragter først den enkle situation, hvor udvælgelsessandsynligheden for en enhed er proportional med populationsværdien x_i. Det er åbenbart, at stikprøven vil være domineret af analyseenheder med store værdier af x, og der vil derfor være en tendens til at stikprøvegennemsnittet er større end populationsgennemsnittet. Der gælder

Sætning 1.8.1 ***Forventet værdi af stikprøvegennemsnit ved udvælgelse proportional med interesssevariablen***

Betragt en population med positive værdier $x(\cdot)$, og antag, at sandsynligheden for at udtage en analyseenhed er proportional med værdien x_i af interessevariablen for den pågældende analyseenhed, d.v.s. at

$$p_i = \mathsf{P}\,[\,I_\nu = i] = x_i \Big/ \sum_{j=1}^{N} x_j \tag{1.8.2}$$

Såfremt stikprøven $X_1, X_2, \ldots, X_n$ udvælges med tilbagelægning med udvalgssandsynlighederne (1.8.2), da gælder for stikprøvegennemsnittet $\overline{X}. = (X_1 + \cdots + X_n)/n$

$$\mathsf{E}\,[\,\overline{X}.] = \xi_{x^2}/\xi_x \tag{1.8.3}$$

med

$$\xi_{x^2} = \sum_1^N x_i^2 \Big/ N \quad \text{og} \quad \xi_x = \sum_1^N x_i \Big/ N$$

Bevis:

Resultatet følger ved indsættelse af (1.8.2) i (1.8.1). □

Bemærkning 1 ***Forventningsværdien af stikprøvegennemsnit ved udvælgelse proportional med interessevariablen er forventningsværdien i den første momentfordeling***

Det ses, at (1.8.3) netop udtrykker forventningsværdien i den første momentfordeling for $x(\cdot)$ se Introduktion til Statistik, Bind 1og Oversigt over fordelinger med anvendelser i Statistik, IMM 1998.

Resultatet er ikke overraskende, idet udtagningen jo netop foregår i overensstemmelse med den første momentfordeling for $x(\cdot)$. □

Eksempel 1.8.1 ***Udvælgelse af buskunder***

Ved en større husstandsundersøgelse spurgtes blandt andet om hyppigheden af benyttelsen af områdets offentlige transportmidler. Blandt de personer, som overhovedet angav at benytte offentlige transportmidler, fandt man følgende fordeling af brugshyppighed:

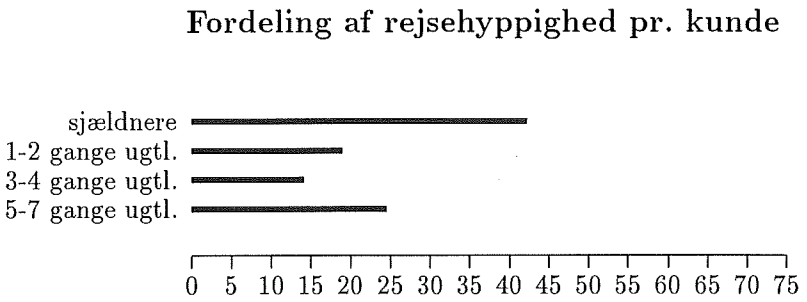
Tabel 1.2. Fordeling af rejsehyppighed for buskunder

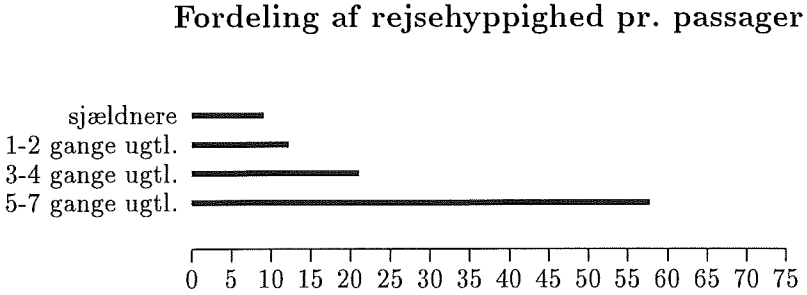
Brugshyp.	nom. værdi	Andel pers.	Rejsedage	Andel Rejsedage
$<$ 1 gang ugtl.	(0.6)	42.3 %	0.2538	10.6 %
1-2 gange ugtl.	(1.5)	19.0 %	0.2850	12.0 %
3-4 gange ugtl.	(3.5)	14.1 %	0.4935	20.7 %
5-7 gange ugtl.	(5.5)	24.6 %	1.3530	56.7 %
Ialt		100.0 %	2.3853	100.0 %

Benytter man den værdi, der er angivet i parentes, til bestemmelse af momentfordelingen, finder man, at de forventede hyppigheder af de fire grupper ved en stikprøveundersøgelse af tilfældigt udvalgte passagerer er som anført i tabellens sidste søjle.

Det fremgår - næppe overraskende - at ca 55% af kundeturene er kunder, der rejser ofte (5-7 gange ugentligt), mens kun 25% af kunderne rejser 5-7 gange ugentligt.

De to fordelinger er anskuet grafisk nedenfor:





□

Eksempel 1.8.2 ***Udvælgelse af fibre***

En uldtråd er spundet af et antal uldfibre af varierende længde. Der udtages nu en stikprøve af trådens fibre ved at der udvælges et tilfældigt tværsnit af tråden, hvorefter tråden klemmes sammen på dette sted, alle de berørte fibre trækkes forsigtigt ud, og længderne $X_1, X_2, \ldots, X_n$ af de udtagne fibre bestemmes.

Vi vil først vurdere udvælgelsessandsynlighederne: Lad trådtykkelsen være n fibre. Lad $x_1, x_2, \ldots, x_N$ betegne længderne af samtlige trådens fibre, og

antag at fibrene i tråden ligger n ved siden af hinanden, blandet uafhængigt af deres længder. Vælger man nu et tilfældigt sted på tråden, da er sandsynligheden for at dette snit vil ramme en fiber af længden x_i netop proportional med fiberlængden, dvs at udvælgelsessandsynligheden netop er givet ved (1.8.2)

Da stikprøven er udtaget proportionalt med fiberlængden er forventningsværdien af det fundne gennemsnit, $\overline{X}.$ netop givet ved (1.8.3), altså som forventningsværdien i den første momentfordeling.

Havde man i stedet udtaget stikprøven ved at definere et "tværsnit" i tråden af længden f.eks. 1 mm, og udvalgt de fibre, hvis venstre ende faldt indenfor dette "tværsnit", ville udvælgelsessandsynligheden være uafhængig af fiberlængden, $\mathsf{P}\ [\ I_\nu = i] = 1/N$, dvs simpel tilfældig udvælgelse, og stikprøvegennemsnittet, $\overline{X}.$, ville være en central estimator for den gennemsnitlige fiberlængde, altså

$$\mathsf{E}\ [\ \overline{X}.] = \xi_x$$

I praksis kan en simpel tilfældig udvælgelse eksempelvis udføres ved at afskære en ende af tråden og fjerne alle de overskårne fibre, sådan at der kun ses naturligt afsluttede fibre. Man griber da 1 mm ned i tråden fra trådenden, og udtager alle de fibre, man herved får fat i. □

Når udvalgssandsynlighederne for de enkelte analyseenheder er kendt, kan man blot korrigere observationerne og derved opnå et centralt estimat for populationsmiddelværdien. Der gælder:

Sætning 1.8.2 ***Fordeling af sandsynlighedskorrigeret stikprøvegennemsnit ved udtagning med kendte udvalgssandsynligheder***

Antag, at stikprøven $X_1, X_2, \ldots, X_n$ fra populationen $\{x_1, x_2, \ldots, x_N\}$ udvælges sådan at $\mathsf{P}\ [I_\nu = i] = p_i; \quad i = 1, 2, \ldots, N$ og betragt de transformerede variable $z(\cdot)$ givet ved

$$z_i = x_i/(Np_i); \quad i = 1, 2, \ldots, N \tag{1.8.4}$$

De således transformerede værdier, z_i kaldes undertiden for de sandsynlighedskorrigerede værdier.

For stikprøvegennemsnittet $\overline{Z}. = \sum_1^n Z_\nu/n$ af Z-værdierne gælder:

$$\mathsf{E}\,[\,\overline{Z}.] = \xi_x; \quad \text{og} \quad \mathsf{V}\,[\,\overline{Z}.] = \frac{1}{n}\,\sigma_z^2 \tag{1.8.5}$$

med

$$\sigma_z^2 = \sum_1^N (z_i - \xi_x)^2\, p_i$$

(Bemærk, at σ_z^2 ikke er den rå populationsvarians, men at de enkelte led er vægtet med udvælgelsessandsynligheden, p_i).

For stikprøvevariansen for Z-værdierne,

$$S_z^2 = \sum_1^n (Z_\nu - \overline{Z}.)^2/(n-1) \tag{1.8.6}$$

gælder der

$$\mathsf{E}\,[\,S_z^2] = \sigma_z^2 \tag{1.8.7}$$

Størrelsen

$$\widehat{\mathsf{V}}\,[\,\overline{Z}.] = \frac{1}{n}\,S_z^2$$

er således et centralt estimat for $\mathsf{V}\,[\,\overline{Z}.]$.

Bevis:

Følger direkte □

Bemærkning 1 *Varians ved udvælgelse proportional med værdien af interessevariablen*

Vi bemærker, at såfremt udvalgssandsynlighederne, p_i, vælges proportionale med populationsværdierne x_i, fås $\sigma_z^2 = 0$, idet alle $z_i = \xi_x$

Disse udvalgssandsynligheder er naturligvis uinteressante i praksis, men bemærkningen viser, at såfremt interessevariablen y_i tilnærmelsesvist er proportionale med en størrelse, x_i, der er kendt på undersøgelsestidspunktet, da kan det svare sig at benytte udvælgelsessandsynligheder, der er proportionale med x_i. □

1.8.3 Udvælgelse proportional med størrelse (PPS)-sampling

Sætning 1.8.3 *Estimation af populationsgennemsnit ved udvælgelse proportional med størrelse (under udvælgelse med tilbagelægning)*

Betragt en population med værdierne $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, hvor værdierne af x_i er kendt. Antag at udvælgelsessandsynlighederne er proportionale med x_i, dvs.

$$p_i = x_i \Big/ \sum_{j=1}^{N} x_j \quad \text{og} \quad \xi_x = \sum_{1}^{N} x_i \Big/ N \ .$$

Lad som vanligt $b_i = y_i/x_i$ angive den relative værdi af interessevariablen for den i'te analyseenhed og lad

$$\overline{B}_\cdot = \frac{1}{n} \sum_{\nu=1}^{n} B_\nu = \frac{1}{n} \sum_{\nu=1}^{n} Y_\nu \Big/ X_\nu \tag{1.8.8}$$

Da gælder

$$\mathsf{E}\,[\overline{B}_\cdot] = \beta^K = \xi_y/\xi_x$$

og

$$\mathsf{V}\,[\overline{B}_\cdot] = \frac{1}{n}(\sigma_b^*)^2$$

hvor

$$(\sigma_b^*)^2 = \frac{1}{N} \sum_{i=1}^{N} \frac{(y_i - \beta^K x_i)^2}{\xi_x x_i} \tag{1.8.9}$$

$(\sigma_b^*)^2$ estimeres centralt ved

$$(S_b^*)^2 = \frac{1}{n-1} \sum_{\nu=1}^{n} (B_\nu - \overline{B})^2 \tag{1.8.10}$$

Estimatoren

$$\hat{\xi}_y^K = \xi_x \times \overline{B}_\cdot \tag{1.8.11}$$

er en central estimator for populationsgennemsnittet ξ_y med variansen

$$\mathsf{V}\,[\hat{\xi}_y^K] = \frac{1}{n}\xi_x^2(\sigma_b^*)^2 \tag{1.8.12}$$

Bevis:
Beviset følger af den foregående sætning ved at bemærke, at de sandsynlighedskorrigerede værdier er

$$z_i = y_i/(Np_i) = \xi_x(y_i/x_i) = \xi_x b_i$$

Der gælder (jvf sætningen), at

$$\mathsf{E}\,[\overline{Z}] = \xi_y$$

men da $\overline{Z} = \xi_x \times \overline{B}$ gælder åbenbart $\mathsf{E}\,[\overline{B}] = \beta^K$.

Man finder

$$\begin{aligned}\sigma_z^2 &= \sum_{i=1}^{N}(z_i - \xi_y)^2 p_i = \sum_{i=1}^{N}(\xi_x/x_i)^2(y_i - \beta^K x_i)^2 x_i \Big/ (N\xi_x) \\ &= \frac{\xi_x}{N}\sum_{i=1}^{N}(y_i - \beta^K x_i)^2 \frac{1}{x_i} = \xi_x^2(\sigma_b^*)^2\end{aligned}$$

□

Bemærkning 1 ***Udvælgelse proportional med størrelse kaldes PPS-sampling***

Hvis x_i er et udtryk for størrelsen af den i'te analyseenhed siger man, at udvælgelsen er proportional med populationsværdiernes størrelse. Ofte bruges den engelske betegnelse, PPS-sampling (**P**robability **P**roportional to **S**ize).

PPS-sampling kan være fordelagtig (give større nøjagtighed i situationer, hvor den relative værdi af interessevariablen ikke varierer så meget hen over den undersøgte population, som f.eks. i undersøgelser af høstudbytte, hvor man kan benytte det tilsåede areal som størrelsesvariabel. Metoden er mindre effektiv, hvis der er stor variation af den relative værdi af interessevariablen, som feks. hvis man interesserer sig for det totale antal

får og bruger brugsstørrelsen som interessevariabel. Hvis undersøgelsen tjener flere formål, som feks både bestemmelse af høstudbytte og antal får er PPS-sampling derfor ikke nødvendigvis den fordelagtigste strategi.

PPS-sampling kan have den praktiske fordel, at stikprøveenhederne blot kan udvælges, feks ved at vælge punkter tilfældigt på et kort. □

Bemærkning 2 ***Ved PPS-sampling er stikprøvegennemsnittet af de relative værdier pr analyseenhed en central estimator for den relative værdi af interessevariablen***

Vi bemærker, at når stikprøven udtages proportional med størrelsen, da er stikprøvegennemsnittet $\overline{B}.$ af de relative værdier pr. analyseenhed en central estimator for den relative værdi (1.2.17) af interessevariablen for populationen. (Dette er ikke tilfældet ved simpel tilfældig udvælgelse jvf sætning 1.6.3). □

I analogi med sætning 1.8.2, kan teorien for kvotientskøn i afsnit 1.7.3 tilsvarende udvides til at omfatte situationer med vilkårlige udvalgssandsynligheder. Der gælder således generelt

Sætning 1.8.4 ***Kvotientskøn ved udvælgelse med vilkårlige sandsynligheder***

Betragt en population, hvor der til hver analyseenhed er knyttet to værdier $x(i)$ og $y(i)$, og antag at interessevariablen er y, og at populationsmiddelværdien ξ_x af x er kendt.

Antag, at stikprøven $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ fra udvælges fra populationen sådan at $\mathsf{P}\,[I_\nu = i] = p_i;\ i = 1, 2, \ldots, N$, og betragt de sandsynlighedskorrigerede variable $(v(\cdot), z(\cdot))$ givet ved

$$v_i = x_i/(Np_i); \quad z_i = y_i/(Np_i); \quad i = 1, 2, \ldots, N \tag{1.8.13}$$

Da gælder for stikprøvegennemsnittene

$$\overline{V}. = \frac{1}{n}\sum_{\nu=1}^{n} V_\nu; \quad \text{og} \quad \overline{Z}. = \frac{1}{n}\sum_{\nu=1}^{n} Z_\nu$$

at

$$\mathsf{E}\,[\overline{V}] = \xi_x\,; \quad \mathsf{E}\,[\overline{Z}] = \xi_y \tag{1.8.14}$$

Lad som vanligt β^K betegne den relative værdi af interessevariablen i populationen, $\beta^K = \xi_y/\xi_x$, og betragt kvotientskønnet

$$\widehat{\beta}^K = \frac{\overline{Z}.}{\overline{V}.} \tag{1.8.15}$$

Der gælder de approksimative relationer

$$\mathsf{E}\,[\widehat{\beta}^K] \simeq \beta^K \tag{1.8.16}$$

$$\mathsf{V}\,[\widehat{\beta}^K] \simeq \frac{1}{\xi_x^2}\,\frac{\sigma_d^2}{n} \tag{1.8.17}$$

hvor

$$\sigma_d^2 = \sum_{i=1}^{N}(z_i - \beta^K v_i)^2 p_i = \sum_{i=1}^{N}(y_i - \beta^K x_i)^2 \Big/ (N^2 p_i) \tag{1.8.18}$$

Variansstørrelsen σ_d^2 estimeres ved

$$S_d^2 = \frac{1}{n-1}\sum_{\nu=1}^{n}(Z_\nu - \widehat{\beta}^K V_\nu)^2 = \frac{1}{n-1}\sum_{\nu=1}^{n}(Y_\nu - \widehat{\beta}^K X_\nu)^2 \Big/ (N p_i)^2 \tag{1.8.19}$$

Kvotientskønnet for populationsgennemsnittet af y'erne fås som

$$\hat{\xi}_y^K = \widehat{\beta}^K \xi_x$$

Estimatet har den approksimative varians

$$\mathsf{V}\,[\hat{\xi}_y^K] \simeq \frac{\sigma_d^2}{n}$$

Bevis:
Se fex Cochran (1963). □

Sætning 1.8.5 ***Estimation af populationsgennemsnit ved PPS-udvælgelse uden tilbagelægning***

Betragt samme situation som i sætning 1.8.3, men antag nu at udvælgelsen foretages uden tilbagelægning.

Dette er kun meningsfuldt, hvis $n \max\{x_i\}$ er mindre end populationstotalen $N\xi_x$, og i dette tilfælde er sandsynligheden for at den i'te enhed indgår i stikprøven

$$\Pi_i = np_i = nx_i / \sum_{j=1}^{N} x_j$$

$$\overline{B}_{\cdot}^{u} = \frac{1}{n} \sum_{\nu=1}^{n} B_\nu = \frac{1}{n} \sum_{\nu=1}^{n} Y_\nu / X_\nu \tag{1.8.20}$$

Da gælder

$$\mathsf{E}\,[\overline{B}_{\cdot}^{u}] = \beta^K = \xi_y / \xi_x$$

og

$$\mathsf{V}\,[\overline{B}_{\cdot}^{u}] = \frac{1}{N^2 \xi_x^2} \sum_{i<j}^{N} \left(x_i x_j - \frac{N^2 \xi_x^2}{n^2} \Pi_{ij} \right) (b_i - b_j)^2 \,,$$

hvor Π_{ij} angiver sandsynligheden for at enhed i og j begge inkluderes i stikprøven.

Variansen estimeres centralt ved

$$\widehat{\mathsf{V}}\,[\overline{B}_{\cdot}^{u}] = \frac{1}{N^2 \xi_x^2} \sum_{\nu<\mu}^{n} \left(X_\nu X_\mu - \frac{N^2 \xi_x^2}{n^2} \Pi_{\nu\mu} \right) (B_\nu - B_\mu)^2 \,.$$

Bevis:
Se Horvitz og Thompson (1952) og Yates og Grundy (1953). □

Bemærkning 1 ***Estimatoren er den samme som ved udvælgelse med tilbagelægning***

Vi bemærker at estimatoren $\overline{B}_{\cdot}^{u}$ givet ved (1.8.20) er den samme som estimatoren $\overline{B}_{\cdot}$ givet i (1.8.8), men da deres fordelingsforhold er forskellige, har vi valgt at benytte forskellige betegnelser.

Estimatoren $\overline{B}_{\cdot}^{u}$ kaldes ofte Horvitz-Thompson estimatoren. □

1.9 Udnyttelse af populationens struktur, stratifikation

Hvis populationen er opdelt i k naturlige grupper (strata), kan man gennem stikprøveplanen sikre, at alle grupper repræsenteres i stikprøven ved at udtage et fastlagt antal (n_i) enheder fra hvert af de i strata, $i = 1, 2, \ldots, k$

Vi benytter betegnelserne

N_i = antal elementer i det i'te stratum, $i = 1, 2, \ldots, k$

$x_{i1}, x_{i2}, \ldots, x_{iN}$, karakteristika for elementerne i det i'te stratum, $i = 1, 2, \ldots, k$

$\xi_i = \sum_j x_{ij}/N_i$, gennemsnitsværdien i det i'te stratum, $i = 1, 2, \ldots, k$

$\sigma_i^2 = \sum_j (x_{ij} - \xi_i)^2/(N_i - 1)$: den korrigerede populationsvarians i det i'te stratum, $i = 1, 2, \ldots, k$

$N. = \sum_i N_i$: populationens størrelse

$w_i = N_i/N$: relativ stratumstørrelse for det i'te stratum, $i = 1, 2, \ldots, k$

n_i stikprøvestørrelse i det i'te stratum, $i = 1, 2, \ldots, k$

$f_i = n_i/N_i$: udvalgsbrøk for det i'te stratum, $i = 1, 2, \ldots, k$

$X_{i\nu}$ ν'te observation fra det i'te stratum, $\nu = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$

$\overline{X}_{i\cdot} = \sum_\nu X_{i\nu}/n_i$: stikprøvegennemsnit fra det i'te stratum.

$S_i^2 = \sum_\nu (X_{i\nu} - \overline{X}_{i\cdot})^2/(n_i - 1)$: stikprøvevariansen i det i'te stratum, $i = 1, 2, \ldots, k$

$n. = \sum_i n_i$: stikprøvens samlede størrelse

Bemærk: Antallet af observationer, n_i, fra hvert stratum fastlægges før stikprøven udtages. Ved udtagning fra det enkelte stratum benyttes simpel tilfældig udvælgelse.

Vi vil endelig indføre omkostningsparameteren

c_i omkostningen pr observation ved undersøgelse af det i'te stratum.

1.9.1 Vilkårlig allokering

Estimatoren for populationsmiddelværdien ξ_x fremkommer umiddelbart ved at vægte de observerede stratumstikprøvegennemsnit $\overline{X}_{i\cdot}$ med de tilsvarende relative stratumstørrelser w_i. Der gælder

Sætning 1.9.1 Fordeling af vægtet stratumgennemsnit

Det stratumvægtede gennemsnit af stikprøvegennemsnittene, $\overline{X}_{i\cdot}$ fra de enkelte strata

$$\hat{\xi}^w = \sum_1^k w_i \, \overline{X}_{i\cdot} \tag{1.9.1}$$

er en central estimator for populationsgennemsnittet ξ med varians

$$\mathsf{V}\,[\,\hat{\xi}^w] = \sum_1^k w_i^2 \frac{\sigma_i^2}{n_i}\,(1-f_i) \tag{1.9.2}$$

Udtrykket (1.9.2) gælder såfremt udvælgelsen er foretaget uden tilbagelægning. For udvælgelse med tilbagelægning bortfalder korrektionsfaktoren $(1-f_i)$ i udtrykket for variansen.

Som skøn over variansen benyttes

$$\widehat{\mathsf{V}}\,[\,\hat{\xi}^w] = \sum_1^k w_i^2 \frac{S_i^2}{n_i}\,(1-f_i) \tag{1.9.3}$$

$(1-\alpha)$ - konfidensgrænser for ξ bestemmes ved:

$$\hat{\xi}^w \pm u_{1-\alpha/2}\sqrt{\widehat{\mathsf{V}}\,[\,\hat{\xi}^w]} \tag{1.9.4}$$

1.9.2 Proportional fordeling af stikprøven på strata

Den simpleste form for stratificeret udvælgelse består i en allokering af stikprøveomfanget n_i proportionalt med stratumstørrelsen N_i, dvs. samme udvalgsfraktion $f = n_i/N_i$ i alle strata. Idet $n.$ angiver den totale stikprøvestørrelse, har vi ved proportional allokering

$$n_i = w_i \, n. \tag{1.9.5}$$

Der gælder

Sætning 1.9.2 ***Estimation ved proportional allokering***

Såfremt stikprøven er allokeret i overensstemmelse med (1.9.5) kan den centrale estimator (1.9.1) udtrykkes som

$$\hat{\xi}^{prop} = \frac{1}{n.} \sum_{i=1}^{k} \sum_{\nu=1}^{n_i} X_{i\nu} = \overline{X}.. \tag{1.9.6}$$

I dette tilfælde udregnes det vægtede gennemsnit af gruppegennemsnittene udregnes ganske simpelt, idet det blot er det simple gennemsnit af samtlige observationer. Man siger, at estimatoren er selvvægtende.

Den tilsvarende varians for estimatet over populationsgennemsnittet ved proportional allokering og stikprøveudvælgelse uden tilbagelægning

$$\mathsf{V}\,[\,\hat{\xi}^{prop}] = \left(\frac{1}{n.} - \frac{1}{N.}\right) \sum_{1}^{k} w_i \sigma_i^2 \tag{1.9.7}$$

Såfremt stikprøven udvælges med tilbagelægning, bortfalder leddet $1/N.$

Bemærkning 1 ***Samme varians i alle strata***

Såfremt variansen er den samme, σ^2, i alle strata, reduceres variansen på estimatet for populationsgennemsnittet til det simple udtryk:

$$V[\hat{\xi}^{prop}] = \frac{\sigma^2}{n.}\left(1 - \frac{n.}{N.}\right)$$

□

Eksempel 1.9.1 ***Stratificeret udvælgelse ved proportional allokering***

En virksomhed har 3000 forhandlere, der hver har et selvstændigt lager af en given bulkvare. De 1000 af forhandlerne er specialforhandlere med et rimelig stort lager, mens de øvrige 2000 er småforhandlere med et lille lager. For at vurdere det totale lager vil man udvælge 120 forhandlere og foretage en lageropgørelse for hver af disse. Benyttes opdelingen i specialforhandlere og småforhandlere som stratifikationsvariabel vil en proportional allokering indebære udtagelse af stikprøveandelen $120/3000 = 3$ % fra hver af de to strata, dvs en tilfældig udvælgelse af $n_1 = 40$ specialforhandlere og $n_2 = 80$ småforhandlere.

Antag at den totale lagerbeholdning for de 40 specialforhandlere fandtes at være 4000 tons med gennemsnittet $\overline{X}_{1\cdot} = 100$ [ton] og med den empiriske varians $S_1^2 = (90\ [\text{ton}])^2$, og at den totale lagerbeholdning for de 80 småforhandlere ligeledes var 4000 [ton], dvs. at gennemsnittet var $\overline{X}_{2\cdot} = 50$ [ton], og at man fandt $S_2^2 = (45\ [\text{ton}])^2$.

Skønnet over den gennemsnitlige lagerbeholdning fås nu (jvf. (1.9.6)) som den fundne stikprøvetotal, divideret med stikprøvestørrelsen, dvs

$$\hat{\xi}^{prop} = (4000 + 4000)/120\ [\text{ton/forhandler}] = 66.7\ [\text{ton per forhandler}]$$

Den estimerede varians på dette skøn bliver (jvf (1.9.7))

$$\widehat{V}[\hat{\xi}^{prop}] = \left(\frac{1}{120} - \frac{1}{3000}\right)\left[\frac{1}{3}\ 90^2 + \frac{2}{3}\ 45^2\right] = 4050/120 = (5.81\ [\text{ton}])^2$$

For at sammenligne med usikkerheden ved simpel tilfældig udvælgelse af 120 forhandlere, vil vi skønne populationsvariansen på basis af stikprøveresultatet. Vi får skønnet over populationsvariansen

$$\widehat{\sigma}^2_{pop} = \frac{1}{3}\ 90^2 + \frac{2}{3}\ 45^2 + \left[\frac{1}{3}\ (100 - 66.7)^2 + \frac{2}{3}\ (50 - 66.7)^2\right] = 4605$$

hvorfor vi har at variansen på gennemsnittet af lagerbeholdningen hos 120 forhandlere, tilfældigt udvalgt blandt hele populationen er

$$\widehat{\mathrm{V}}\,[\overline{X}.] = 4605/120 = (6.20\ [\mathrm{ton}])^2\ .$$

Forskellen mellem de to variansudtryk hidrører fra udtrykket i den kantede parentes, der beskriver variationen mellem de to strata, udtrykt gennem de to stratumgennemsnits variation omkring populationsgennemsnittet. □

1.9.3 Optimal fordeling på strata

Da opdelingen af populationen i strata foreligger før stikprøveindsamlingen påbegyndes, er det muligt at fordele det totale stikprøveomfang på de enkelte strata på en sådan måde at man får den mindst mulige varians for et givet stikprøvebudget, eller eventuelt opnår en given varians med en minimal udgift.

Vi antager, at omkostningen pr observation ved undersøgelse af det i'te stratum er c_i.

Den totale stikprøveudgift ved udvælgelse af n_i enheder fra hvert stratum er da

$$C = \sum_{1}^{k} c_i n_i \tag{1.9.8}$$

Der gælder:

Sætning 1.9.3 ***Optimal allokering ved fastlagt total stikprøveudgift***

Antag, at stikprøveomkostningerne under stratificeret stikprøveudvælgelse er givet ved (1.9.8), hvor det totale budget, C, er givet. Da antager variansen (1.9.2) på estimatet for populationsgennemsnittet sit minimum, såfremt stratumstikprøvestørrelserne n_i tilfredsstiller

$$n_i/n. = \left[N_i\sigma_i \Big/ \sqrt{c_i}\right] \Big/ \left[\sum_{j=1}^{k}(N_j\sigma_j/\sqrt{c_j}\right] \tag{1.9.9}$$

med

$$n. = C \left[\sum_1^k (N_i \sigma_i / \sqrt{c_i})\right] \Big/ \left[\sum_1^k (N_i \sigma_i \sqrt{c_i})\right] \tag{1.9.10}$$

Bevis:

Sætningen vises for eksempel ved benyttelse af Lagrange multiplikatormetoden. Man finder herved, at stikprøvestørrelserne skal tilfredsstille

$$n_i = \alpha \ \sigma_i \ w_i / \sqrt{c_i} \tag{1.9.11}$$

hvor α er en proportionalitetskonstant.

Man finder da, at det optimale forhold mellem størrelsen n_i af stikprøven fra det i'te stratum og den totale stikprøvestørrelse $n.$ er bestemt ved (1.9.9). Indsættes (1.9.9) i (1.9.8) og løses ligningen med hensyn til $n.$ fås udtrykket (1.9.10). □

Bemærkning 1 ***Betingelser for udtagelse af stor stikprøve fra et stratum***

Det fremgår af (1.9.11), at stikprøvestørrelserne skal være proportionale med $\sigma_i \ w_i / \sqrt{c_i}$, dvs vi har umiddelbart, at

Der bør udtages en stor stikprøve (større end den gennemsnitlige stikprøve) fra et givet stratum, såfremt:

a) Det pågældende stratum har mange elementer

b) Det pågældende stratum udviser stor intern variation

c) Stikprøveudtagning fra det pågældende stratum er billig

□

Korollar 1.9.1 ***Neyman allokering ved samme omkostning i alle strata*** Såfremt stikprøveomkostningen er den samme i alle strata, dvs.

$c_i = c$, vil en fastlagt værdi, C af den totale stikprøveomkostning svare til en fastsat total stikprøvestørrelse $n. = C/c$.

I dette tilfælde får man den optimale allokering for et fastlagt totalt stikprøveomfang, $n. = \sum_1^n n_i$:

$$n_i = n. \frac{w_i \sigma_i}{\sum_j w_j \sigma_j} \tag{1.9.12}$$

dvs udvalgsfraktionen $f_i = n_i/N_i$ er proportional med statumspredningen σ_i.

Allokeringen (1.9.12) kaldes Neyman-allokeringen efter den polsk-amerikanske statistiker Jerzy Neyman.

Såfremt allokeringen af stikprøven til de k strata er udført i overensstemmelse med (1.9.12) betegnes det tilsvarende stratumvægtede gennemsnit af stikprøvegennemsnittene

$$\hat{\xi}^{Ney} = \sum_1^k w_i \overline{X}_i.$$

Variansen af det stratumvægtede gennemsnit for denne allokering er

$$\mathsf{V}[\hat{\xi}^{Ney}] = \frac{1}{n.} \left(\sum_1^k w_i \sigma_i \right)^2 - \frac{1}{N.} \sum_1^k w_i \sigma_i^2 \tag{1.9.13}$$

Bemærkning 2 ***Optimal strategi for population stratificeret efter størrelse***

Betragt en stratificeret population med værdierne (x_{ij}, y_{ij}), hvor værdierne af x_{ij} er et udtryk for "størrelsen" af den j'te enhed i det i'te stratum. Antag at populationen er stratificeret efter størrelse, dvs. at x_{ij} er nogenlunde

ensartet indenfor hvert stratum, dvs. $x_{ij} \approx x_i^*$. Antag endvidere at spredningen σ_i af interessevariablen y_{ij} indenfor det i'te stratum er proportional med størrelsen x_i^* i det i'te stratum, og at der er samme omkostning i alle strata. Da vil den optimale allokering efter (1.9.12) være

$$n_i = n. \frac{N_i x_i^*}{\sum_h N_h x_h^*} \approx n. \frac{\sum_j x_{ij}}{\sum_h \sum_j x_{hj}}$$

altså proportional med den totale "størrelse", $\sum_j x_{ij}$ af de enkelte strata. □

Eksempel 1.9.2 *Neyman allokering*

En brancheorganisation ønskede at vurdere den gennemsnitlige nettoindtjening per virksomhed i branchen. Der forelå et register over de enkelte virksomheders egenkapital. Den grupperede fordeling af de 7590 virksomheder efter egenkapital er angivet i nedenstående tabel. Tabellen indeholder desuden forhåndsskøn over den gennemsnitlige indtjening i hver af grupperne, samt skøn over spredningen af indtjeningen inden for hver af grupperne.

Tabel 1.3. Fordeling af 7590 virksomheder efter egenkapital. Skønnede stratumgennemsnit og spredning af nettoindtjeningen er anført i tabellen. (Efter Deming : Some Theory of sampling)

Stratumgrænser i 100 000 kr	Antal virksomheder N_i	skønnet Nettoindk. i 100 000 kr	skønnet spredning 100 000 kr
Ukendt	560	1	5
Under 50	2870	1	5
50 - 99	1110	5	8
100 - 249	1300	15	20
250 - 499	750	50	65
500 - 999	510	100	130
1000 - 5000	580	300	390

Vi bemærker, at en væsentlig del af den totale nettoindkomst synes koncentreret på en lille del af virksomhederne.

Der skønnes at være samme stikprøveudgift ved undersøgelse af alle strata, og man vælger derfor en Neyman-allokering. Det totale antal undersøgte virksomheder sættes til $n. = 760$.

Tabel 1.4. Tabel til brug for bestemmelse af Neyman allokering af stikprøve fra populationen i tabel i 1.3.

Stratumgrænser i 100 000 kr	$N_i\sigma_i$	$N_i\sigma_i$ i % af total	n_i
Ukendt	2800	0.71	6
Under 50	14350	3.65	28
50 - 99	8880	2.26	17
100 - 249	26000	6.61	50
250 - 499	48750	12.40	94
500 - 999	66300	16.86	128
1000 - 5000	226200	57.52	437
Total	393280	100.01	760

Tabel 1.4 illustrerer beregningen af den optimale allokering af stikprøven. □

Sætning 1.9.4 ***Optimal allokering ved fastlagt total varians***

Antag at den ønskede varians V for skønnet over populationsgennemsnittet er fastlagt. Da vil en allokering

$$n_i = (w_i\sigma_i/\sqrt{c_i})\left[\sum_{j=1}^{k}(w_j\sigma_j/\sqrt{c_j})\right] \Big/ \left[V + (\sum_{j=1}^{k} w_j\sigma_j^2)/N.\right] \quad (1.9.14)$$

minimere den totale stikprøveomkostning (1.9.8) under hensyn til dette varianskrav.

Bevis:

Beviset følger af sætning 1.9.3 ved indsættelse af den optimale allokering (1.9.9) og bestemmelse af den tilhørende varians. □

Korollar 1.9.2 ***Optimal allokering ved ens stikprøveomkostninger***

Såfremt stikprøveomkostningen, c_i, er den samme i alle strata, finder man, at den stikprøveallokering, der minimerer den totale stikprøveomkostningen

under den betingelse, at variansen på skønnet over populationsgennemsnittet ikke må overstige V, er givet ved:

$$n_i = w_i\sigma_i \left[\sum_{j=1}^{k} w_j\sigma_j\right] \Big/ \left[V + (\sum_{j=1}^{k} w_j\sigma_j^2)/N.\right] \qquad (1.9.15)$$

1.9.4 Sammenligning mellem simpel tilfældig og stratificeret udvælgelse

Betragt en population på ialt $N.$ elementer, opdelt på k strata med N_i elementer i det i'te stratum og lad den sande stratummiddelværdi og spredning i det i'te stratum være henholdsvis ξ_i og σ_i, og lad tilsvarende $\xi.$ og $\sigma.$ betegne populationsmiddelværdi og spredning.

Antag nu at der udvælges en stikprøve på ialt $n.$ elementer. Såfremt stikprøven udtages ved simpel tilfældig udvælgelse i hele populationen finder vi variansen på stikprøvegennemsnittet:

$$\mathsf{V}[\overline{X}.] = \left(\frac{1}{n.} - \frac{1}{N.}\right)\sigma.^2 \qquad (1.9.16)$$

Såfremt stikprøven vælges ved stratificeret udvælgelse med proportional allokering (dvs stratumstikprøvestørrelse proportional med stratumstørrelse) finder vi variansen (1.9.7) på estimatet for populationsmiddelværdien

$$\mathsf{V}[\hat{\xi}^{prop}] = \left(\frac{1}{n.} - \frac{1}{N.}\right)\sum_{1}^{k} w_i\sigma_i^2 \qquad (1.9.17)$$

Havde vi endelig valgt at udtage stikprøven optimalt ved en Neyman allokering (idet udvælgelsesomkostningerne c_i antages at være de samme i alle strata), havde vi fået variansen

$$\mathsf{V}[\hat{\xi}^{Ney}] = \frac{1}{n.}\left(\sum_{1}^{k} w_i\sigma_i\right)^2 - \frac{1}{N.}\sum_{1}^{k} w_i\sigma_i^2 \qquad (1.9.18)$$

Det er åbenbart, at der må gælde

$$\mathsf{V}\,[\,\hat{\xi}^{Ney}] \leq \mathsf{V}\,[\,\hat{\xi}^{prop}] \leq \mathsf{V}\,[\,\overline{X}.]$$

Betragter vi nu forskellen mellem simpel tilfældig udvælgelse og stratificeret udvælgelse proportional med stratumstørrelsen, finder vi

$$\mathsf{V}\,[\,\overline{X}.] - \mathsf{V}\,[\,\hat{\xi}^{prop}] = \left(\frac{1}{n.} - \frac{1}{N.}\right) \sum_{1}^{k} w_i(\xi_i - \xi.)^2$$

der viser, at stratifikationsgevinsten opnås såfremt der er forskel på stratummiddelværdierne. Jo større indbyrdes forskelle, desto større er gevinsten.

Stratifikation er således fordelagtig i sig selv, når opdelingen i strata er foretaget således at der er stor variation mellem strata, men ringe variation indenfor strata.

Tilsvarende finder vi forskellen mellem stratificeret udvælgelse proportional med stratumstørrelsen og den optimale Neyman-allokering:

$$\mathsf{V}\,[\,\hat{\xi}^{prop}] - \mathsf{V}\,[\,\hat{\xi}^{Ney}] = \frac{1}{n.} \sum_{1}^{k} \sum_{j=1}^{k} w_i w_j (\sigma_i - \sigma_j)^2$$

Gevinsten ved yderligere at foretage en optimal allokering opnås altså såfremt der er forskel på stratumvarianserne. Jo større indbyrdes forskelle, desto større er gevinsten.

Allokering	Udvalgs-andel f_i	estimat	varians	Ref.
Vilkårlig	-	$\hat{\xi}_x^w = \sum w_i \overline{X}_{i\cdot}$	$\sum w_i^2 \frac{\sigma_i^2}{n_i}(1-f_i)$	Sætn. 1.9.1
Proportional	$n_i/N_i = c$	$\hat{\xi}_x^{prop} = \frac{1}{n_\cdot}\sum_i \sum_\nu X_{i\nu}$	$\left(\frac{1}{n_\cdot} - \frac{1}{N_\cdot}\right)\sum_i w_i \sigma_i^2$	Sætn. 1.9.2
Neyman	$n_i/N_i = c\sigma_i$	$\hat{\xi}_x^{Ney} = \sum_i w_i \overline{X}_{i\cdot}$	$\frac{1}{n_\cdot}\left(\sum_i w_i \sigma_i\right)^2 - \frac{1}{N_\cdot}\sum_i w_i \sigma_i^2$	Korollar 1.9.1

Tabel 1.5. Oversigt over estimater for populationsmiddelværdi $\xi_\cdot$ ved stratificeret udvælgelse

N_i	Stratumstørrelse af i'te stratum
$w_i = N_i/N_\cdot$	Relativ størrelse af i'te stratum
n_i	Stikprøvestørrelse fra i'te stratum
$\overline{X}_{i\cdot}$	Stikprøvegennemsnit i i'te stratum, : $\overline{X}_{i\cdot} = \sum_\nu X_{i\nu}/n_i$
σ_i^2	Varians i i'te stratum
S_i^2	stikprøvevarians i i'te stratum: $S_i^2 = \sum_\nu (X_{i\nu} - \overline{X}_{i\cdot})^2/(n_i - 1)$ (estimat for σ_i^2)

1.10 Udnyttelse af populationens struktur, Klyngeudvælgelse

fil: repr5.tex 1998-01-24

Når populationen er opdelt i en række administrative enheder (klynger), vil det ofte være bekvemt at udtage et antal klynger ved simpel tilfældig udvælgelse blandt alle klynger i populationen, hvorefter de udvalgte klynger gøres til genstand for total tælling (et-trinsudvælgelse), eller der udtages en stikprøve fra hver af de udvalgte klyngerne (to-trinsudvælgelse).

Ved klyngeudvælgelse opfatter man således klyngerne som værende i det væsentlige ens, og hver for sig giver klyngerne et minibillede af den undersøgte population (med lige så stor variation indenfor en klynge, som i hele populationen. Derfor nøjes man med at udvælge et antal klynger, som til gengæld udsættes for en total tælling.

Ved stratificeret udvælgelse opfatter man de enkelte strata som værende væsensforskellige, og variationen i populationen hidrører i et vidt omfang fra variationen mellem strata. Derfor skal alle strata repræsenteres i stikprøven, men til gengæld nøjes man så med en stikprøve fra hvert stratum.

Vi benytter betegnelserne

K = antal klynger i populationen

N_i = antal analyseenheder i den i'te klynge, $i = 1, 2, \ldots, K$

N_{tot} = $\sum_i N_i$: totalt antal analyseenheder i populationen

$x_{i1}, x_{i2}, \ldots, x_{iN}$, værdier af interessevariablen for enhederne i den i'te klynge, $i = 1, 2, \ldots, K$

$\zeta_{i\cdot}$ = $\sum_j x_{ij}$ Klyngetotalen i den i'te klynge, $i = 1, 2, \ldots, K$

$\xi_{i\cdot}$ = $\zeta_{i\cdot}/N_i$, Klyngegennemsnittet i den i'te klynge, $i = 1, 2, \ldots, K$

$\bar{\zeta}_{\cdot\cdot}$ = $\sum_i \zeta_{i\cdot}/K$, den gennemsnitlige klyngetotal

$\zeta_{\cdot\cdot}$ = $\sum_i \zeta_{i\cdot}$, Populationstotalen

σ_ζ^2 = $\sum_i (\zeta_{i\cdot} - \bar{\zeta}_{\cdot\cdot})^2/(K-1)$, Populationsvariansen mellem klyngetotaler

$\bar{\xi}_{..} = \zeta_{..}/(\sum_i N_i)$, Gennemsnit pr analysesenhed i populationen

$\overline{N}_{.} = N_{tot}/K$, Gennemsnitlig klyngestørrelse

$\sigma_x^2 = \sum_i \sum_j (x_{ij} - \xi_{..})^2/(N_{tot} - 1)$, variansen mellem populationsenheder

$\sigma_N^2 = \sum_1^k (N_i - \overline{N}_{.})^2/(K-1)$, variansen mellem klyngestørrelser

$\sigma_{\zeta,N} = \sum_1^k (N_i - \overline{N}_{.})(\zeta_{i.} - \bar{\zeta}_{..})/(K-1)$, Kovariansen mellem klyngetotal og klyngestørrelse

$\gamma_N^2 = \sigma_N^2/\overline{N}_{.}^2$. Den relative varians for klyngestørrelserne

$\gamma_{\zeta,N} = \sigma_{\zeta,N}/(\overline{N}_{.}\bar{\zeta}_{..})$, den relative kovarians mellem klyngetotaler $\zeta_{i.}$ og klyngestørrelser N_i

Sætning 1.10.1 ***Central estimation ved brug af gennemsnitlig klyngetotal***

Antag at der udvælges k klynger ved simpel tilfældig udvælgelse blandt de K klynger.

Lad

$$Z_\kappa = \sum_{\nu=1}^{N_\kappa} X_{\kappa\nu} \tag{1.10.1}$$

betegne klyngetotalen i den κ'te klynge, $\kappa = 1, 2, \ldots, k$ og lad

$$\overline{Z}_{.} = \frac{1}{k} \sum_{\kappa=1}^{k} Z_\kappa \tag{1.10.2}$$

angive stikprøvegennemsnittet af klyngetotalerne.

Da gælder at $\overline{Z}_{.}$ er en central estimator for den gennemsnitlige klyngetotal $\bar{\zeta}_{..}$ i populationen. Variansen er

$$\mathsf{V}[\overline{Z}_{.}] = \frac{1}{k}\ \sigma_\zeta^2 \left(1 - \frac{k}{K}\right) \tag{1.10.3}$$

Skønnet

$$\hat{\zeta}_{..} = K\,\overline{Z}_{.} \tag{1.10.4}$$

er en central estimator for populationstotalen $\zeta_{..}$ med variansen

$$\mathsf{V}\,[\hat{\zeta}_{..}] = \frac{K(K-k)}{k}\ \sigma_{\zeta}^{2} \tag{1.10.5}$$

og tilsvarende er skønnet

$$\hat{\bar{\xi}}_{..} = \frac{1}{\overline{N}_{.}}\,\overline{Z}_{.} \tag{1.10.6}$$

en central estimator for gennemsnittet pr analyseenhed i populationen $\bar{\xi}_{..}$ med variansen

$$\mathsf{V}\,[\hat{\bar{\xi}}_{..}] = \frac{1}{k\overline{N}^{2}}\ \sigma_{\zeta}^{2}\left(1-\frac{k}{K}\right) \tag{1.10.7}$$

Populationsvariansen mellem klyngetotaler σ_{ζ}^{2} estimeres centralt ved stikprøvevariansen mellem klyngetotaler

$$S_{z}^{2} = \frac{1}{k-1}\sum_{\kappa=1}^{k}(Z_{\kappa}-\overline{Z}_{.})^{2} \tag{1.10.8}$$

Bevis:
Resultatet er en direkte konsekvens af at klyngetotalerne udvælges ved simpel tilfældig udvælgelse. □

Bemærkning 1 ***Variansen for klyngeskønnet udtrykt ved intra-klyngekorrelationen***

Såfremt alle klynger er lige store, $N_i = N$, kan variansen (1.10.7) udtrykkes ved intraklyngekorrelationen. Sætter vi

$$\rho = \frac{1}{(N-1)(K \times N - 1)} \sum_{i=1}^{K} \sum_{j=1}^{N} \frac{(x_{ij} - \xi_{..})(x_{ij} - \xi_{..})}{\sigma_x^2}, \tag{1.10.9}$$

har vi

$$\mathrm{V}[\hat{\bar{\xi}}_{..}] = \frac{1}{k \times N} \sigma_x^2 [1 + (N-1)\rho] \left(1 - \frac{k}{K}\right) \tag{1.10.10}$$

Intraklyngekorrelationen, ρ, udtrykker graden af ensartethed mellem analyseenhederne i en klynge sammenlignet med den totale variation mellem samtlige analyseenheder. Jo større del af den totale variation, der er knyttet til forskelle mellem klynger (dvs. jo mere ensartede enhederne er indenfor en klynge), desto mindre effektivt er klyngeskønnet. □

De centrale skøn, der blev betragtet i sætning 1.10.1 tog udgangspunkt i en simpel tilfældig udvælgelse af klyngetotalerne. Variansen på skønnene afhang derfor af variationen σ_ζ^2 mellem klyngetotalerne.

Hvis der er stor forskel på klyngestørrelserne, og hvis der er korrelation mellem klyngestørrelse og klyngetotal (hvad der meget ofte vil være), vil variansen mellem klyngetotalerne være influeret af denne variation mellem klyngetotaler, og det kan derfor være fordelagtigt at forsøge at tilgodese forskellene i klyngestørrelse ved at betragte gennemsnittet $\overline{\overline{X}}_{..}$ pr analyseenhed i stikprøven.

Vi sætter

$$\overline{\overline{X}}_{..} = \sum_{\kappa=1}^{k} \sum_{\nu=1}^{N_\kappa} X_{\kappa\nu} \Big/ \sum_{\kappa=1}^{k} N_\kappa = \sum_{\kappa=1}^{k} Z_\kappa \Big/ \sum_{\kappa=1}^{k} N_\kappa \tag{1.10.11}$$

Skønnet $\overline{\overline{X}}_{..}$ er et kvotientskøn, idet den totale klyngestørrelse for stikprøven $\sum N_\kappa$ varierer afhængigt af hvilke klynger, der bliver udvalgt.

Egenskaberne for skønnet er givet i

Sætning 1.10.2 ***Kvotientskøn for gennemsnitlig værdi pr analyseenhed***

Kvotientskønnet for populationsgennemsnittet

$$\hat{\bar{\xi}}_{..}^{K} = \overline{\overline{X}}_{..} \tag{1.10.12}$$

hvor stikprøvegennemsnittet pr analyseenhed $\overline{\overline{X}}_{..}$ er bestemt ved (1.10.11) er en ikke-central estimator for gennemsnittet pr analyseenhed $\bar{\xi}_{..}$. Skønnet har skævheden

$$\mathsf{E}\left[\hat{\bar{\xi}}_{..}^{K} - \bar{\xi}_{..}\right] \simeq \frac{1}{k}\left(\gamma_N^2 - \gamma_{\zeta,N}\right)\frac{K-k}{K-1}, \tag{1.10.13}$$

hvor γ_N^2 angiver den relative varians på klyngestørrelserne og $\gamma_{\zeta,N}$ angiver den relative kovarians mellem klyngetotaler ζ_i og klyngestørrelser N_i

Variansen på kvotientskønnet er

$$\mathsf{V}\left[\hat{\bar{\xi}}_{..}^{K}\right] \simeq \frac{1}{k\overline{N}^2}\sigma_d^2\left(1 - \frac{k}{K}\right) \tag{1.10.14}$$

hvor σ_d^2 angiver populationsvariansen mellem de størrelseskorrigerede klyngetotaler

$$\sigma_d^2 = \frac{1}{K-1}\sum_{i=1}^{K}(\zeta_i - N_i\bar{\xi}_{..})^2 \tag{1.10.15}$$

Variansen σ_d^2 estimeres ved

$$\hat{\sigma}_d^2 = S_d^2 = \frac{1}{k-1}\sum_{\kappa=1}^{k}(Z_\kappa - N_\kappa\overline{\overline{X}}_{..})^2 = \frac{1}{k-1}\sum_{\kappa=1}^{k}N_\kappa^2(\overline{X}_\kappa - \overline{\overline{X}}_{..})^2 \tag{1.10.16}$$

Bevis:
Resultaterne følger af teorien for kvotientskøn, sætning 1.7.2. Med betegnelserne fra afsnit 1.7 har vi jo en situation, hvor stikprøveenheden er en klynge med de to tilknyttede værdier $Y_i = \zeta_i$ og $X_i = N_i$, og hvor vi ønsker at estimere den relative værdi $\beta = \sum \zeta_i / \sum N_i$ af summen af klyngetotaler i forhold til det summen af analyseenheder for populationen af klynger.

Estimatet (1.10.11) er netop kvotientskønnet $\hat{\beta}^K$ (1.7.2) svarende til denne situation. □

Bemærkning 1 *Alternativ formulering af variansen for kvotientskønnet*

Et alternativt udtryk for variansen på kvotientskønnet for populationsgennemsnittet er

$$\mathsf{V}[\widehat{\bar{\xi}}_{..}^{K}] \simeq \frac{1}{k\overline{N}_{\cdot}^{2}} \left(\sigma_\zeta^2 + \xi_{..}^2 \sigma_N^2 - 2\xi_{..}\sigma_{\zeta,N}\right) \left(1 - \frac{k}{K}\right)$$

hvor σ_N^2 angiver variationen mellem klyngestørrelser, og $\sigma_{\zeta,N}$ angiver kovariansen mellem klyngestørrelse og klyngetotal. □

Bemærkning 2 *Brug af kvotientskønnet uden kendskab til det totale antal analyseenheder*

Undertiden kender man ikke det totale antal analyseenheder (eller, hvad der er det samme: populationsværdien af det gennemsnitlige antal analyseenheder i klyngerne). Det er jo netop klyngeplanens styrke, at man ikke behøver en liste over samtlige enheder i populationen, men kan nøjes med at generere en liste over klyngerne samt lister for de udvalgte klynger.

Da kvotientskønnet (1.10.12) kun udnytter kendskabet til klyngestørrelserne for de udvalgte klynger, kan skønnet også benyttes i disse tilfælde.

I sådanne tilfælde erstattes populationens gennemsnitlige klyngestørrelse $\overline{N}$ i udtrykket (1.10.14) for $\mathsf{V}[\widehat{\bar{\xi}}_{..}^{K}]$ med den gennemsnitlige klyngestørrelse i stikprøven,

$$\widehat{\overline{N}}_{\cdot} = \frac{1}{k}\sum_{\kappa=1}^{k} N_\kappa \tag{1.10.17}$$

□

Eksempel 1.10.1 *Undersøgelse af skoleelevers TV-forbrug*

Som led i en regional undersøgelse af skoleelevers TV-vaner ønsker man for elever i 5 klassse at bestemme antallet af timer pr. elev, der er i en given uge er tilbragt med at se tysk TV.

Undersøgelsens ramme er 35 klasser med ialt 628 elever.

Analyseenheden er en elev, og interessevariablen er antallet af tyske TV-timer i den betragtede uge.

Det er naturligt at opfatte en klasse som en klynge. Der er således ialt $K = 35$ klynger med ialt $N_{tot} = 628$ elever.

Man har udvalgt seks klasser tilfældigt blandt de 35 klasser og udspurgt samtlige elever i hver af de udvalgte klasser. Resultatet er anført i nedenstående tabel (Efter Axel Schultz Nielsen: Indføring i teorien for stikprøveundersøgelser, Samfundslitteratur 1978):

Klasse	Antal elever N_κ	tot. ant. tim. Z_κ	gnsn. ant. tim $\overline{X}_{\kappa\cdot}$
1	20	328.0	16.4
2	22	336.6	15.3
3	17	307.7	18.1
4	15	220.5	14.7
5	18	293.4	16.3
6	20	380.0	19.0
ialt	112	1866.2	99.8
snit	18.67	311.03	16.63
spredning	2.50	53.32	

Vi vil indledningsvist betragte stikprøveresultatet som resultatet af en simpel tilfældig udvælgelse af klyngetotaler (dvs. ζ_i angiver det totale antal timer i den i'te klasse).

Det følger af sætning 1.10.1, at stikprøvegennemsnittet pr klasse er et centralt estimat for populationsmiddelværdien pr. klasse. Vi finder jvf (1.10.2) stikprøvegennemsnittet

$$\overline{Z}_{\cdot} = \frac{1}{6}\sum_{\kappa=1}^{6} Z_\kappa = \frac{1866.2}{6} = 311.03 \text{ [timer/klasse]}$$

Variansen σ_ζ^2 estimeres jvf (1.10.8) ved

$$\widehat{\sigma}_\zeta^2 = S_z^2 = \frac{1}{5}\sum_{\kappa=1}^{6}(Z_\kappa - 311.03)^2 = (53.32\ [\text{timer/klasse}])^2\ ,$$

hvorfor variansen for det gennemsnitlige antal TV-timer pr klasse estimeres ved (1.10.3)

$$\widehat{\mathrm{V}}\,[\,\overline{Z}_\cdot\,] = \frac{1}{6}\ 53.32^2\left(1 - \frac{6}{35}\right) = (19.81\ [\text{timer/klasse}])^2$$

Da vi kender den gennemnsnitlige klassestørrelse i populationen,

$$\overline{N}_\cdot = 628/35 = 17.94\ [\text{elever/klasse}]\ ,$$

følger det af sætning 1.10.1, at skønnet (1.10.6)

$$\widehat{\overline{\xi}}_{\cdot\cdot} = \frac{311.03}{17.94} = 17.33\ [\text{timer/elev}]$$

over det gennemsnitlige timetal pr elev er et centralt skøn. Variansen for dette skøn estimeres jvf (1.10.7) ved

$$\widehat{\mathrm{V}}\,[\,\widehat{\overline{\xi}}_{\cdot\cdot}\,] = \left(\frac{19.81}{17.94}\right)^2 = (1.104\ [\text{timer/elev}])^2\ .$$

Problemet ved denne fremgangsmåde er imidlertid, at skønnet essentielt er baseret på simpel tilfældig udvælgelse af totale TV-timer i en klasse. Usikkerheden er derfor udtrykt ved variationen mellem totale TV-timer i klasserne. Hvis der er stor forskel på klassestørrelserne, kan der også tænkes at være stor variation mellem klassetotalerne

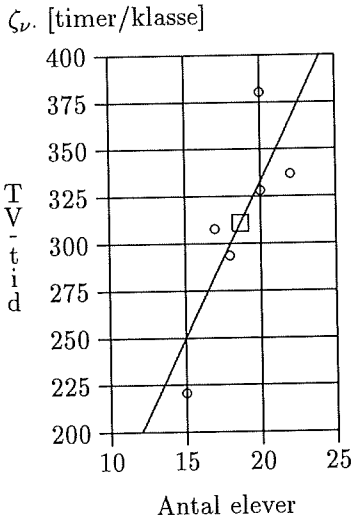
I figur 1.1 er optegnet samhørende værdier af klassestørrelse og total TV-tid pr. klasse. Det ses, at der er en positiv samvariation mellem klassestørrelse og klassetotal, hvorfor et kvotientskøn kan tænkes at være fordelagtigt.

I øvrigt viser figuren, at klassegennemsnit og klassestørrelse varierer uafhængigt af hinanden. Da skævheden på det simple gennemsnit af de klassevise TV-timetal pr elev (1.6.8)

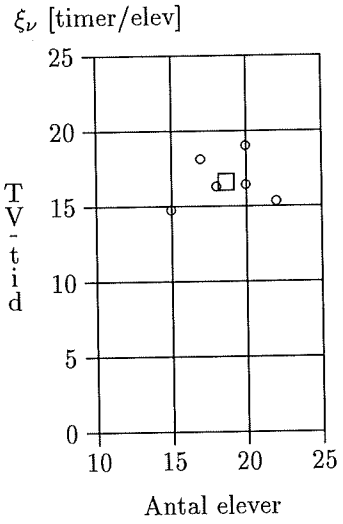
$$\frac{1}{6}\sum_{\nu=1}^{6}\overline{X}_{\nu\cdot} = \frac{99.8}{6} = 16.63$$

Figur 1.1. *Analyse af klyngestikprøveplan*

total TV-tid pr klasse mod klassestørrelse



klassesnit af TV-tid pr elev mod klassestørrelse



□ angiver stikprøvegennemsnittet
Den indtegnede linie angiver linien igennem origo og stikprøvegennemsnittet

essentielt afhænger af denne samvariation (se (1.7.14)), er dette skøn måske ikke særlig skævt. (Vi skal i sætning 1.10.4 se at dette skøn ville være centralt, såfremt udvælgelsen var foretaget med sandsynligheder, der var proportionale med klassestørrelsen).

Vi vil her betragte kvotientskønnet (1.10.12) fra sætning 1.10.2.

Man finder skønnet

$$\hat{\bar{\xi}}_{..}^{K} = \sum_{\kappa=1}^{6} Z_\kappa \Big/ \sum_{\kappa=1}^{6} N_\kappa = \frac{1866.2}{112} = 16.66 \text{ [timer/elev]}$$

De tilsvarende "fittede" værdier $N_\kappa \hat{\bar{\xi}}_{..}^{K}$ af det totale antal TV-timer i klassen er anført i nedenstående tabel

Klasse	1	2	3	4	5	6
Tot ant. tim. Z_κ	328.0	336.6	307.7	220.5	293.4	380.0
Fitted antal $N_\kappa \bar{\bar{X}}_{..}$	333.25	366.57	283.26	2249.94	299.92	333.25
Afvigelse	-5.25	-29.97	24.44	-29.44	-6.52	46.75

Man kan nu bestemme estimatet for variansen mellem de størrelseskorrigerede klyngetotaler

$$S_d^2 = \frac{1}{5} \sum_{\kappa=1}^{6} (Z_\kappa - N_\kappa \bar{\bar{X}}_{..})^2 = \frac{4617.96}{5} = 923.59 = (30.39 \text{ [timer/klasse]})^2$$

sådan at man får estimatet over variansen

$$\hat{V}[\hat{\bar{\xi}}_{..}^{K}] = \frac{1}{6 \times 17.94^2} \, 923.59 \left(1 - \frac{6}{35}\right) = \frac{923.591}{1931.06} \times 0.8286$$

$$= 0.3963 = (0.630 \text{ [timer/elev]})^2$$

Man har således opnået en væsentlig reduktion af usikkerheden i forhold til usikkerheden på det centrale skøn (1.10.6). □

Eksempel 1.10.2 *Stikprøveudvælgelse af æbler i kasser*

Ved stikprøvekontrol af et parti bestående af 1 000 kasser med æbler udvalgtes tilfældigt 8 kasser, og man optalte antallet af æbler i hver kasse, samt antallet af æbler med brune pletter. Man fandt følgende resultat:

Kasse nr	1	2	3	4	5	6	7	8	Ialt
Antal æbler N_κ	50	40	45	55	70	65	35	40	400
Antal plettede Z_κ	4	21	6	30	50	4	20	15	150

Det er klart, at der er tale om klyngeudvælgelse, hvor en kasse svarende til en klynge.

Mens antallet af kasser $K = 1\ 000$ er kendt, kender man ikke det totale antal æbler.

Man bruger kvotientskønnet (1.10.12) til estimation af andelen af plettede æbler

$$\begin{aligned} \hat{p}^K &= \sum_{\kappa=1}^{8} Z_\kappa \Big/ \sum_{\kappa=1}^{8} N_\kappa \\ &= 150/400 = 0.38 \end{aligned}$$

Altså 38 % af æblerne er plettede.

Variansen på dette skøn estimeres ved (1.10.14). Da det gennemsnitlige antal æbler pr kasse for partiet ikke kendes, benytter man estimatet $\widehat{\overline{N}} = 50$ [æbler/ kasse]

De størrelseskorrigerede klyngetotaler fremgår af nedenstående tabel:

Kasse nr	1	2	3	4	5	6	7	8
Antal æbler N_κ	50	40	45	55	70	65	35	40
Antal plettede Z_κ	4	21	6	30	50	4	20	15
$N_\kappa \hat{p}^K$	18.75	15.00	16.88	20.63	26.25	24.38	13.13	15.00

Man har

$$S_d^2 = \frac{1}{7}\sum_{\kappa=1}^{k}(Z_\kappa - N_\kappa 0.38)^2 = \frac{1486.19}{7} = 212.31 = (14.57)^2$$

hvorfor man får

$$\widehat{\mathrm{V}}\,[\hat{p}^K] = \frac{212.31}{8 \times 50^2}\left(1 - \frac{8}{1000}\right) = 0.01053 = (0.103)^2$$

Forestiller man sig, at man kunne udtage 400 æbler ved simpel tilfældig udvælgelse blandt samtlige æbler i partiet, og at man fandt 38 % plettede æbler i denne stikprøve, ville et skøn over spredningen på dette estimat være $\hat{p}(1-\hat{p})/400 = 0.00059$.

Variansen på klyngeskønnet er altså væsentligt større end binomialvariansen. Dette skyldes at pletterne på æblerne har en tendens til at følges ad: Såfremt ét æble i en kasse har brune pletter, har naboæblerne det også. Der er åbenbart en positiv intraklassekorrelation. □

1.10.1 Udvælgelse af klynger med varierende sands.

Antag, at den i'te klynge vælges med sandsynlighed p_i

Et godt valg er at vælge p_i proportional med klyngestørrelse, men da den ikke nødvendigvis er kendt, kan man evt bruge en hjælpevariabel, som forhåbentlig udtrykker noget om størrelsen.

Man siger da, at man udvælger **Pro**Portionalt med **E**stimeret **S**tørrelse (PPES -sampling).

Sætning 1.10.3 ***Estimation ved udvælgelse af klynger proportionalt med estimeret størrelse***

Såfremt den i'te klynge vælges med en fastlagt sandsynlighed p_i er skønnet

$$\widehat{\zeta}_{\cdot\cdot}^{ppes} = \frac{1}{k}\sum_{\kappa=1}^{k}\frac{Z_\kappa}{p_\kappa} \tag{1.10.18}$$

en central estimator for populationstotalen ζ.

Variansen på estimatet er

$$\mathsf{V}\,[\widehat{\zeta}_{\cdot\cdot}^{ppes}] = \frac{1}{k}\sigma_{pp}^2 \tag{1.10.19}$$

med

$$\sigma_{pp}^2 = \sum_{i=1}^{K} p_i(\frac{\zeta_{i\cdot}}{p_i} - \zeta_{\cdot\cdot})^2$$

Variansen σ_{pp}^2 estimeres ved den tilsvarende stikprøvevarians for de sandsynlighedskorrigerede klyngetotaler

$$\widehat{\sigma}_{pp}^2 = \frac{1}{k-1}\sum_{\kappa=1}^{k}(\frac{Z_\kappa}{p_\kappa} - \widehat{\zeta}_{\cdot\cdot}^{ppes})^2$$

Bevis:
Resultatet følger af sætning 1.8.2 anvendt på udvælgelse af klyngerne og klyngetotaler. □

Vælges specielt klyngerne med sandsynligheder, der er proportionale med klyngestørrelsen, $p_i = N_i/N_{tot}$ siger man, at man udvælger **ProPortionalt** med **Størrelse** (PPS -sampling). I dette tilfælde gælder

Sætning 1.10.4 ***Estimation ved udvælgelse proportional med klyngestørrelse***

Såfremt den i'te klynge vælges med sandsynligheden $p_i = N_i/N_{tot}$, er skønnet

$$\widehat{\bar{\xi}}_{\cdot}^{pps} = \frac{1}{k}\sum_{\kappa=1}^{k}\overline{X}_\kappa \tag{1.10.20}$$

en central estimator for populationsgennemsnittet $\bar{\xi}$.

Variansen for dette estimat er

$$\mathsf{V}\,[\widehat{\bar{\xi}}_{\cdot}^{pps}] = \frac{1}{k}(\sigma_z^*)^2\,, \tag{1.10.21}$$

hvor

$$(\sigma_z^*)^2 = \sum_{i=1}^{K} \left(\frac{N_i}{N_{tot}} \right) (\xi_{i\cdot} - \bar{\xi}_{\cdot\cdot})^2 \qquad (1.10.22)$$

Variansen $(\sigma_z^*)^2$ estimeres centralt ved

$$(\widehat{\sigma}_z^*)^2 = \frac{1}{k-1} \sum_{\kappa=1}^{k} (\overline{X}_\kappa - \widehat{\bar{\xi}}_{\cdot}^{pps})^2 \qquad (1.10.23)$$

Et centralt estimat for populationstotalen $\zeta_{\cdot}$ fås da som

$$\widehat{\zeta}_{\cdot\cdot}^{pps} = N_{tot} \widehat{\bar{\xi}}_{\cdot}^{pps}$$

Skønnet har variansen

$$\mathsf{V}\,[\widehat{\zeta}_{\cdot\cdot}^{pps}] = \frac{N_{tot}^2}{k} (\sigma_z^*)^2$$

Bevis:
Resultatet følger af den foregående sætning □

Bemærkning 1 ***Fortolkning af estimat ved udvælgelse proportionalt med klyngestørrelse***

Vi bemærker, at estimatet $\widehat{\bar{\xi}}_{\cdot}^{pps}$ for den gennemsnitlige værdi i populationen givet ved (1.10.20) netop er det simple gennemsnit af de gennemsnitlige værdier $\overline{X}_\kappa$ i klyngerne. □

Eksempel 1.10.3 ***Undersøgelse af skoleelevers TV-forbrug (fortsat)***

Vi betragter atter situationen i eksempel 1.10.1.

For at illustrere beregningerne ved PPS-estimation vil vi antage, at de 6 klasser i stikprøven var udvalgt med sandsynligheder, der var proportionale med klassestørrelsen.

En måde, hvorpå dette kunne foregå, var at man benyttede en liste over samtlige 628 elever, og valgte elever ud fra listen ved simpel tilfældig udvælgelse. Når en elev blev udtaget, undersøgte man hele den klasse, der hørte til den pågældende elev.

Ved indsættelse i (1.10.20) finder man skønnet

$$\widehat{\bar{\xi}}_{\cdot}^{pps} = 16.63 \text{ [timer/elev]} .$$

Residualerne $\overline{X}_\kappa - \widehat{\bar{\xi}}_{\cdot}^{pps}$ er angivet i nedenstående tabel.

Klasse	1	2	3	4	5	6
$\overline{X}_\kappa$	16.40	15.30	18.10	14.70	16.30	19.00
$\overline{X}_\kappa - 16.63$	-0.23	-1.33	1.47	-1.93	-0.33	2.37

Man får skønnet over $(\sigma_z^*)^2$

$$(\widehat{\sigma}_z^*)^2 = \frac{13.4333}{5} = 2.6867 ,$$

hvorfor estimatet for variansen på skønnet over den gennemsnitlige TV-tid pr elev bliver

$$\begin{aligned} \widehat{\mathrm{V}}\,[\widehat{\bar{\xi}}_{\cdot}^{pps}] &= \frac{2.6867}{6} = 0.4478 \\ &= (0.669 \text{ [timer/elev]})^2 \end{aligned}$$

□

Metode	estimat	varians	variansestimat	Ref.
Centrale	$\frac{1}{k}\sum_{\kappa}\sum_{\nu} X_{\kappa\nu}$	$\frac{1}{k\overline{N}^2}\ \sigma_\zeta^2\left(1-\frac{k}{K}\right)$	$\widehat{\sigma}_\zeta^2 = S_z^2 = \frac{1}{k-1}\sum_{\kappa=1}^{k}(Z_\kappa - \overline{Z}_\cdot)^2$	Sætn. 1.10.1
Kvotient	$\widehat{\overline{\xi}}_{\cdot\cdot}^{K} = \frac{\sum_\kappa Z_\kappa}{\sum_\kappa N_\kappa}$	$\frac{1}{k\overline{N}^2}\sigma_d^2\left(1-\frac{k}{K}\right)$	$\widehat{\sigma}_d^2 = \frac{1}{k-1}\sum_{\kappa}(Z_\kappa - N_\kappa\overline{\overline{X}}_{\cdot\cdot})^2$	Sætn. 1.10.2
Prop med størr.	$\widehat{\overline{\xi}}_{\cdot}^{pps} = \frac{1}{k}\sum_{\kappa}\overline{X}_\kappa$	$\frac{1}{k}(\sigma_z^*)^2$	$(\widehat{\sigma}_z^*)^2 = \frac{1}{k-1}\sum_{\kappa}(\overline{X}_\kappa - \widehat{\overline{\xi}}_{\cdot}^{pps})^2$	Sætn. 1.10.4

Tabel 1.6. Oversigt over estimater for populationsmiddelværdi $\xi_\cdot$ ved klyngeudvælgelse

N_i Klyngestørrelse af i'te klynge
Z_i Klyngetotal i i i'te klynge,
$\overline{X}_{i\cdot}$ gennemsnit i i'te klynge, : $\overline{X}_{i\cdot} = \sum_\nu X_{i\nu}/N_i$

1.11 Totrinsudvælgelse

Teorien for klyngeudvælgelse forudsætter at de udvalgte klynger undersøges fuldstændigt, hvorfor klyngerne helst skal være af en nogenlunde begrænset størrelse.

I praksis bruger man ofte en tillempning af klyngeplanen, hvor man udvælger stikprøveenhederne i flere trin.

Vi vil her blot kort skitsere princippet i en totrinsudvælgelse.

I en totrinsplan har man opdelt populationen i en række primære stikprøveenheder, det kan feks. være kommuner, geografiske områder el. lign. Der udvælges så (i lighed med klyngeplanen) tilfældigt et antal af disse primærenheder, og i hver af de udvalgte primærenheder, udvælges nu et antal egentlige stikprøveenheder (subenheder).

Såfremt de primære enheder udvælges med sandsynligheder, der er proportionale med deres størrelse (antal subenheder), og der derefter udvælges et fast antal subenheder fra hver primærenhed ved simpel tilfældig udvælgelse, opnår man at stikprøven er selvvægtende således at man kan bruge en estimator af formen (1.9.6) idet alle subenheder har samme sandsynlighed for at blive udvalgt, og det er derfor ikke nødvendigt at foretage en yderligere vægtning for at opnå et centralt estimat.

En sådan totrinsplan har ofte en række praktiske fordele frem for andre selvvægtende strategier som simpel tilfældig udvælgelse af primære stikprøveenheder efterfulgt af en stikprøve af samme relative størrelse i hver af de primære stikprøveenheder. Denne sidste fremgangsmåde indebærer blandt andet, at man ikke på forhånd kan bestemme den totale stikprøvestørrelse. Ligeledes vil et fast antal stikprøver i hver udvalgt primærenhed være lettere at administrere, feks ved interviewundersøgelser.

Hvis de primære stikprøveenheder har en selvstændig interesse, vil den foreslåede totrinsplan med PPS-udvælgelse af primære stikprøveenheder og fast stikprovestørrelse indenfor disse ofte være at foretrække, da den faste stikprøvestørrelse i de primære enheder muliggør en simpel sammenligning mellem de primære enheder.

1.12 Referencer

C.-M. Cassel, C-E. Särndal, J.H. Wretman (1977): *Foundations of Inference in Survey Sampling*, Wiley, New York

W.G.Cochran (1963): *Sampling Techniques*, 2.ed. , Wiley, New York

W.E.Deming (1950): *Some Theory of Sampling*, Dover, New York

W.E. Deming (1960): *Sample Design in Business Research*, Wiley, New York

M.H.Hansen, W.N.Hurwitz and W.G. Madow (1953): *Sample Survey Methods and Theory I and II*, Wiley, New York

D.G.Horvitz and D.J.Thompson: (1952): A generalization of sampling without replacement from a finite universe *JASA* **47**, pp 663-685

N.J.Johnson and H.Smith (eds) (1969): *New Developments in Survey Sampling*, Wiley, New York

Kakwani, N. C. (1980). *Income Inequality and Poverty.* Oxford University Press, Oxford.

Kendall, M., Stuart, A. and Ord, J.K. (1983): *The Advanced Theory of Statistics, Vol. 3: Design and Analysis, and Time Series*, fourth edition, Charles Griffin & Company, London

L.Kish (1965): *Survey Sampling*, Wiley, New York

H.S.Konijn (1973): *Statistical theory of sample survey design and analysis*, North Holland

Lerman, R.I. and Yitzhaki, S (1984): A note on the calculation and interpretation of the Gini index, *Economics Letters*, **15**, pp. 363-368.

M.N.Murthy (1967): *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta

Des Raj (1968): *Sampling Theory*, McGraw-Hill, New York

Des Raj (1972): *The Design of Sampling Surveys*, McGraw-Hill, New York,

M.Tin (1965): Comparison of some ratio estimators, *Journ. Amer. Statist. Ass.*, **60**, p294 ff.

P.V.Sukhatme and B.V.Sukhatme (1970): *Sampling Theory of Surveys with Applications*, Asia Publishing House, Bombay

F.Yates (1960): *Sampling Methods for Censuses and Surveys*, 3.ed. Griffin, London

F.Yates og P.M.Grundy (1953): Selection without replacement from within strata with probability proportional to size. Journ. Roy Statist.Soc. B, **15**, pp. 253-261.

Indeks