2003/28 Notater 2003

Anna-Karin Mevik

Usikkerhet i konjunktur-barometeret

Notater

Statistics Norway

Statistisk sentralbyrå

Seksjon for statistiske metoder og standarder
Emnegruppe: 08.90

Innhold

1. Innledning

I dette notatet skal vi se hvordan vi kan måle og estimere usikkerheten i konjunkturbarometeret. I kapittel 2 til 7 tar vi for oss usikkerheten til nåværende estimator, mens vi i kapittel 8 og 9 skal se på usikkerheten til to alternative estimatorer. Vi vil se bort fra mulige feilkilder som målefeil, registerfeil og frafall, og kun se på usikkerheten som kommer av at vi estimerer en populasjonstotal på bakgrunn av et utvalg fra populasjonen. Denne usikkerheten angis vanligvis med det designbaserte standardavviket, men her skal vi også se på et modellbasert standardavvik.

Selv om vi har valgt å se bort fra effekten av frafall, tar estimatoren som benyttes i konjunkturbarometeret hensyn til visse typer frafall. Er f.eks. frafallet tilfeldig innen hvert stratum pga. sammenheng med sysselsettingen, vil dette ikke medføre frafallsskjevhet.

Konjunkturbarometeret er en kvartalsvis kartlegging av industriens vurderinger av konjunktursituasjonen og -utsikten for en del faste kjennetegn. Det gis blant annet en sysselsettingsveid indeks, kalt diffusjonsindeks, for industriens forventede produksjon i neste kvartal sammenlignet med inneværende kvartal. Det er for denne indeksen vi vil estimere usikkerheten (men estimatorene vi kommer frem til vil også passe for tilsvarende indekser i konjunkturbarometeret).

Vi starter i avsnitt 2 med å gi en beskrivelse av populasjonen og diffusjonsindeksen. I neste avsnitt, avsnitt 3, gir vi en kort beskrivelse av utvalget og hvordan det trekkes. Selve estimatoren som benyttes til å estimere diffusjonsindeksen presenterer vi i avsnitt 4. I avsnitt 5 og 6 gir vi en designbasert og en modellbasert analyse av denne estimatoren. Usikkerhetsmålene vi får fra disse analysene har vi så benyttet til å estimere usikkerheten i konjunkturbarometeret i perioden 1999, 2000 og 2002. Resultatene av dette gir vi i avsnitt 7. I avsnitt 8 og 9 ser vi på to alternative estimatorer for diffusjonsindeksen. Helt til slutt gir vi en oppsummering i avsnitt 10.

2. Populasjon

Populasjonen består av alle bransjeenheter innen industri og bergverksdrift. Med nærings-grupperingen i SN94 (standard for næringsgruppering, 1994) betyr det alle bransjeenheter innen næringene 10,13,14 og 15-37. Før vi forklarer hva vi her mener med en bransjeenhet, nevner vi at alle bedrifter er sortert under næringshovedgrupper: I SN94 er alle næringene delt opp i næringshovedgrupper, også kalt 3-sifrede næringer, og i BoF (Bedrifts- og foretaksregisteret) er alle bedriftene sortert under disse næringshovedgruppene. Et foretak kan inneholde flere bransjeenheter. En bransjeenhet er definert ved alle bedriftene til foretaket som tilhører samme næringshovedgruppe (se eksempel neste avsnitt). Bransjeenheten er selv sortert under den samme næringshovedgruppen som bedriftene. Det er BoF som til enhver tid bestemmer hvilke bransjeenheter som er med i populasjonen.

For å gi et eksempel på en bransjeenhet ser vi på Norsk Hydro. Norsk Hydro er et foretak som består av 44 bedrifter. Disse bedriftene er som nevnt sortert under næringshovedgruppene. F.eks. driver 8 av bedriftene med en virksomhet som gjør at de er gruppert under næringshovedgruppen for utvinning av råolje og naturgass, og disse 8 bedriftene utgjør således en bransjeenhet. Videre er 14 av bedriftene gruppert under næringshovedgruppen for produksjon av kjemiske råvarer, slik at disse bedriftene utgjør en bransjeenhet. De resterende 22 bedriftene til Norsk Hydro er sortert under 11 ulike næringshovedgrupper, og utgjør dermed 11 bransjeenheter.

2.1. Stratifisering av populasjonen

Populasjonen stratifiseres etter sysselsetting og næringshovedgruppe. Dvs. at bransjeenhetene grupperes etter hvor stor sysselsetting de har og hvilken næringshovedgruppe de tilhører. Inndelingsintervallene for sysselsetting er 0-99, 100-199, 200-299 og 300-∞.[1] Fordi det er ca. 100 næringshovedgrupper i populasjonen, skulle dette tilsi omtrent 400 strata. Men fordi mange av næringshovedgruppene ikke har bransjeenheter i alle sysselsettingsintervallene, blir det ca. 270 strata.

2.2. Diffusjonsindeksen til populasjonen

For hver bransjeenhet i populasjonen tenker vi oss at det fins en sysselsettingsveid indikator, kalt bransjeenhetens konjunktursysselsetting, som angir hvorvidt bransjeenheten forventer en økning, nedgang eller ingen forandring i neste kvartals produksjon sammenlignet med inneværende kvartal. Mer presist tenker vi oss at det for hver bransjeenhet i i hvert stratum h fins en konjunktursysselsetting y_{hi} gitt ved

$$y_{hi} = \begin{cases} x_{hi} & \text{, hvis bransjeenheten forventer en økning i sin egen produksjon} \\ \frac{1}{2}x_{hi} & \text{, hvis bransjeenheten ikke forventer noen endring i sin egen produksjon} \\ 0 & \text{, hvis bransjeenheten forventer en nedgang i sin egen produksjon} \end{cases}.$$

Her er x_{hi} bransjeenhetens sysselsetting.

For hvert stratum h har vi også en konjunktursysselsetting. Denne er gitt ved

$$Y_h = \sum_{i \in U_h} y_{hi} ,$$

der U_h er indeksmengden for bransjeenhetene i stratumet. Stratumets konjunktursysselsetting er med andre ord gitt ved den samlede sysselsettingen til de bransjeenhetene (i stratumet) som tror på en økning i egen produksjon, pluss halvparten av den samlede sysselsettingen til de bransjeenhetene som ikke forventer noen endring. Dermed er stratumets konjunktursysselsetting alltid mindre eller lik sysselsettingen i stratumet, dvs.

$$Y_h \leq X_h$$

der $X_h = \sum_{i \in U_h} x_{hi}$ er stratumets sysselsetting. Videre vil konjunktursysselsettingen Y_h være større jo flere av bransjeenhetene som tror på en økning i egen produksjon.

Ved å summere konjunktursysselsettingen til alle strata i populasjonen, får vi populasjonens konjunktursysselsetting Y. Dvs.

$$Y = \sum_h Y_h ,$$

[1] For noen næringshovedgrupper kan inndelingen etter sysselsetting være litt annerledes.

der summen $\sum_h$ er over alle strata h i populasjonen. Konjunktursysselsettingen Y er alltid mindre eller lik den totale sysselsettingen i populasjonen, og jo større optimisme det er blant bransjeenhetene, jo nærmere den totale sysselsettingen vil Y være.

Vi er interessert i populasjonens diffusjonsindeks. Denne er gitt ved

$$
\begin{aligned}
d &= \frac{Y}{X}\cdot 100 \\
&= \frac{\sum_h Y_h}{\sum_h X_h}\cdot 100\ ,
\end{aligned}
\tag{1}
$$

der $X = \sum_h X_h$ er den totale sysselsettingen i populasjonen. Dvs. at diffusjonsindeksen angir populasjonens konjunktursysselsetting Y i prosent av den totale sysselsettingen X.[2]

På grunn av ulikheten $0 \le Y \le X$ vil d ligge mellom 0 og 100. Hvis $d > 50$ betyr det at bransjeenhetene som forventer en økning i produksjonen sysselsetter flere arbeidstakere enn bransjeenhetene som forventer en nedgang. At $d < 50$ betyr det motsatte, dvs. at bransjeenhetene som forventer en økning i produksjon sysselsetter færre arbeidstakere enn bransjeenhetene som forventer en nedgang. Vi kan si at $d > 50$ indikerer vekst i produksjonen mens $d < 50$ indikerer reduksjon.

Vi er også interessert i diffusjonsindekser som gjelder for spesielle grupper av strata (dvs. delmengder av populasjonen). Diffusjonsindeksen til en slik gruppe er gitt som (1), men hvor summen $\sum_h$ bare går over de strataene som er med i gruppen.

3. Utvalg

Ideelt sett skal det trekkes et nytt utvalg hvert år, og så skal det ved behov foretas supplering og rullering av utvalget i løpet av de fire kvartalene utvalget benyttes. I praksis følges ikke denne regelen helt. F.eks. er det ikke trukket nye utvalg i løpet av 1999 eller 2000.

Hver gang det trekkes et nytt utvalg så oppdateres populasjonsdataene mot BoF. Dermed stemmer populasjonsdataene overens med den faktiske populasjonen på det tidspunktet utvalget trekkes. Men når det samme utvalget brukes til å estimere diffusjonsindeksen på et senere tidspunkt, kan det være avvik mellom populasjonsdataene og den faktiske populasjonen.

3.1. Trekking av nytt utvalg

I (brutto-) utvalget, som er på ca. 710 enheter, inngår alle bransjeenheter med 300 eller flere sysselsatte fast. De resterende utvalgsenhetene allokeres mellom strataene på en optimal måte for å sikre en viss dekningsgrad av den totale sysselsettingen og omsettingen i populasjonen. Allokeringen sikrer ikke at vi får utvalg fra alle strataene, og vi kan dermed oppleve å ha strata uten utvalg.

---

[2] Fordi konjunktursysselsettingen Y tilsvarer den samlede sysselsettingen til de bransjeenhetene (i populasjonen) som tror på en økning i egen produksjon, pluss halvparten av den samlede sysselsettingen til de bransjeenhetene som ikke forventer noen endring, tilsvarer d den prosentvise andelen av sysselsatte som jobber i bransjeenheter hvor det forventes en økning i produksjonen, pluss halvparten av den prosentvise andelen av sysselsatte som jobber i bransjeenheter hvor det ikke forventes noen endring.

Strataenes utvalg trekkes uavhengig av hverandre og uten tilbakelegging. Innen hvert stratum brukes det trekksannsynligheter som er proporsjonale med bransjeenhetenes sysselsetting. Bransjeenheter med færre enn 10 sysselsatte er ikke med i trekkpopulasjonen. Det eneste unntaket er for grafisk industri, der alle bransjeenheter er tatt med i trekkpopulasjonen. Grunnen til dette er at resultater fra grafisk industri brukes i et markedsoppdrag, og derfor er dekningsgraden i denne næringshovedgruppen økt for å styrke disse resultatene.

3.2. Supplering og rullering av utvalget

Ved supplering og rullering av utvalget tas det blant annet hensyn til hvor mange ganger en bransjeenhet har vært med i utvalget og hvor stort frafall den har. Det gjøres ingen ny allokering av utvalgsenhetene, og de nye bransjeenhetene trekkes med trekksannsynligheter som er proporsjonale med sysselsettingen.

3.3. Innsamling av data

Den kvartalsvise datainnsamlingen fra de utvalgte bransjeenhetene skjer ved hjelp av spørreskjema. Ett av spørsmålene på dette skjemaet er om hvorvidt bransjeenheten forventer en økning, nedgang eller ingen forandring i neste kvartals produksjon sammenlignet med inneværende kvartal. På bakgrunn av svaret som gis, beregnes så bransjeenhetens konjunktursysselsetting.

Vanligvis er det den største bedriften innen bransjeenheten som skal svare, men for noen bransjeenheter er det valgt ut flere bedrifter som alle skal svare. I sistnevnte tilfelle beregnes det en konjunktursysselsetting for hvert av svarene til de leverede bedriftene, og bransjeenhetens konjunktursysselsetting er gitt ved gjennomsnittet av disse.

4. Estimering av diffusjonsindeksen

I konjunkturbarometeret estimeres diffusjonsindeksen (1) med

(2) $$\hat{d} = \frac{\sum_h \hat{Y}_h}{\sum_h X_h} \cdot 100 ,$$

der

(3) $$\hat{Y}_h = \frac{\bar{y}_{s_h}}{\bar{x}_{s_h}} \cdot X_h$$

er en rateestimator for stratumets konjunktursysselsetting $Y_h = \sum_{i \in U_h} y_{hi}$. Her er $\bar{x}_{s_h}$ og $\bar{y}_{s_h}$ henholdsvis gjennomsnittlig sysselsetting og gjennomsnittlig konjunktursysselsetting i stratumets (netto-) utvalg, dvs.

$$\bar{x}_{s_h} = \frac{1}{n_h} \sum_{i \in s_h} x_{hi} \quad \text{og} \quad \bar{y}_{s_h} = \frac{1}{n_h} \sum_{i \in s_h} y_{hi}$$

der s_h er stratumets (netto-) utvalg og n_h er antall bransjeenheter i dette utvalget.

Rateestimatoren (3) krever at vi har et utvalg fra stratumet. Som nevnt kan vi ha strata uten utvalg. Antakeligvis blir slike strata fjernet fra populasjonen når $\hat{d}$ skal beregnes, slik at summen $\sum_h$ i (2) bare er over de strataene hvor vi har utvalg.

På grunn av ulikheten $0 \le y_{hi} \le x_{hi}$ har vi ulikheten $0 \le \bar{y}_{s_h} \le \bar{x}_{s_h}$ slik at $0 \le \hat{Y}_h \le X_h$. Dette gir at $\hat{d}$ ligger mellom 0 og 100, noe som er en nødvendig egenskap siden diffusjonsindeksen ligger mellom 0 og 100. I de neste to avsnittene skal vi se på flere egenskaper til $\hat{d}$.

Diffusjonsindeksen til grupper av strata estimeres også med (2), men igjen med summen $\sum_h$ kun over de strataene som er med i gruppen.

5. Designbasert analyse av estimert diffusjonsindeks

I dette avsnittet skal vi gjøre en designbasert analyse av $\hat{d}$. At analysen er designbasert betyr at vi ser på bransjeenhetenes konjunktursysselsettinger som ukjente parametere, mens utvalget av bransjeenheter er stokastisk. I dette avsnittet vil derfor alle forventninger og varianser gjelde for denne situasjonen. Vi vil anta at vi har utvalg fra alle strata, og vi vil se bort fra frafall, registerfeil og målefeil.

Vi ønsker ofte at en estimator skal være forventningsrett. For $\hat{d}$ betyr dette at forventningen skal være lik d for alle mulige kombinasjoner av de verdiene som y_{hi} 'ene kan ha. Men dette er ikke tilfelle.[3] For å se det skriver vi forventningen til $\hat{d}$ som

$$\mathrm{E}\hat{d} = \frac{\sum_h \mathrm{E}\hat{Y}_h}{\sum_h X_h} \cdot 100,$$

der $\mathrm{E}\hat{Y}_h$ er forventningen til $\hat{Y}_h$. Som vi skal se senere er $\hat{Y}_h$ generelt ikke forventningsrett: Noen konfigurasjoner av y_{hi} 'ene kan gjøre at $\mathrm{E}\hat{Y}_h$ er større enn Y_h, andre at $\mathrm{E}\hat{Y}_h$ er mindre enn Y_h. Dermed kan forventningen til $\hat{d}$ bli både større og mindre enn d avhenging av y_{hi} 'ene, og følgelig er $\hat{d}$ ikke forventningsrett.

Ved å bruke uavhengigheten mellom stratumsutvalgene finner vi at variansen til $\hat{d}$ er gitt ved

$$\mathrm{V}\left(\hat{d}\right) = \frac{\sum_h \mathrm{V}\left(\hat{Y}_h\right)}{\left(\sum_h X_h\right)^2} \cdot 100^2,$$

[3] Selv om $\hat{d}$ ikke er forventningsrett, finnes det mange konfigurasjoner av y_{hi} 'ene som gjør at $\mathrm{E}\hat{d} = d$. Hvis f.eks. forholdet y_{hi} / x_{hi} er konstant for alle bransjeenheter innen hvert stratum, så er $\hat{Y}_h = Y_h \ \forall\ h$. Dermed er $\mathrm{E}\hat{Y}_h = Y_h \ \forall\ h$ slik at $\mathrm{E}\hat{d} = d$.

der $\mathrm{V}(\hat{Y}_h)$ er variansen til $\hat{Y}_h$. Dermed kan standardavviket, som er kvadratroten av variansen, for $\hat{d}$ skrives som

(4) $$\mathrm{st}(\hat{d}) = \frac{\sqrt{\sum_h \mathrm{V}(\hat{Y}_h)}}{\sum_h X_h} \cdot 100 .$$

Som nevnt brukes ofte standardavviket for å angi usikkerheten til en estimator. Men når estimatoren ikke er forventningsrett, er kanskje bruttovariansen et vel så bra usikkerhetsmål. Bruttovariansen er forventet kvadratavvik mellom estimator og estimand, og kan skrives som variansen pluss kvadratet av forventningsskjevheten. (Forventningsskjevheten til en estimator er forventningen til estimatoren minus estimanden). Dvs. at bruttovariansen til $\hat{d}$ er

$$\begin{aligned}\mathrm{MSE}(\hat{\mathrm{d}}) &= \mathrm{E}\left[\left(\hat{d} - d\right)^2\right] \\ &= \mathrm{V}(\hat{\mathrm{d}}) + \mathrm{Bias}(\hat{\mathrm{d}})^2,\end{aligned}$$

der $\mathrm{Bias}(\hat{d}) = \mathrm{E}\left[\hat{d} - d\right]$ er forventningsskjevheten til $\hat{d}$.

Til tross for at $\hat{d}$ ikke er forventningsrett velger vi å bruke standardavviket som usikkerhetsmål. Grunnen til dette er at vi ikke klarer å estimere bruttovariansen. (I underavsnitt 5.1 gir vi en øvre og nedre grense for forventningsskjevheten til $\hat{d}$).

Fordi standardavviket (4) er ukjent må vi estimere det. Det gjør vi ved å estimere variansen til $\hat{Y}_h$'ene. Fordi vi ikke har noen standardestimator for $\mathrm{V}(\hat{Y}_h)$ skal vi i underavsnitt 5.1 estimere denne variansen ved hjelp av lineærisering. Vi har også prøvd å finne en Bootstrapmetode som er egnet til å estimere $\mathrm{V}(\hat{Y}_h)$, uten å finne det. (I underavsnitt 5.2 gir vi en kort presentasjon av Bootstrap).

Før vi går i gang med lineæriseringen skal vi se litt på trekksannsynlighetene til bransjeenhetene. Vi skal også se på situasjonen $n_h = 1$, dvs. en situasjon hvor det er umulig å estimere variansen på vanlig måte.

Trekksannsynligheten til en bransjeenhet er sannsynligheten for at bransjeenheten blir trukket ut til å være med i utvalget. Som notasjon for denne sannsynligheten bruker vi π_{hi} for bransjeenhet i i stratum h. Fordi bransjeenheter med færre enn 10 sysselsatte ikke er med i trekkpopulasjonen (med unntak av bransjeenhetene i grafisk industri), er $\pi_{hi} = 0$ for disse bransjeenhetene. For de øvrige bransjeenhetene er π_{hi} proporsjonal med sysselsettingen (innen hvert stratum). Med notasjonen

$$U_h^* = \{i \in U_h \text{ s.a. } \pi_{hi} > 0\}$$

har vi derfor at $\pi_{hi} = 0$ når $i \notin U_h^*$, og $\pi_{hi} = n_h \left(x_{hi} / \sum_{i \in U_h^*} x_{hi} \right)$ når $i \in U_h^*$ (forutsatt at $n_h \left(x_{hi} / \sum_{i \in U_h^*} x_{hi} \right) \leq 1$ for alle $i \in U_h^*$, noe som er tilfelle for de aller fleste strataene). Vi kan merke oss at $U_h^* = U_h$ når sysselsettingsintervallet i stratumet ikke er 0-99.

For stratum h der $n_h = 1$ er det umulig å estimere variansen. For slike strata velger vi derfor å bruke en øvre grense som en konservativ estimator for variansen. For å komme fram til en øvre grense benytter vi at når $n_h = 1$ er $\hat{Y}_h = \left(y_{hi_s} / x_{hi_s} \right) X_h$ og $\pi_{hi} = x_{hi} / \sum_{i \in U_h^*} x_{hi}$. (Her er i_s den utvalgte bransjeenheten i stratumet). Dette gir forventningen

$$
\begin{aligned}
\mathrm{E}\hat{Y}_h &= \sum_{i \in U_h^*} \frac{y_{hi}}{x_{hi}} X_h \pi_{hi} \\
&= \sum_{i \in U_h^*} y_{hi} + \frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} \sum_{i \notin U_h^*} x_{hi} \ ,
\end{aligned} \tag{5}
$$

som generelt ikke er lik Y_h. Dvs. at $\hat{Y}_h$ generelt ikke er forventningsrett. Men hvis $U_h^* = U_h$ får vi $\mathrm{E}\hat{Y}_h = Y_h$, dvs. forventningsrett. Variansen til $\hat{Y}_h$ kan nå skrives som

$$
\mathrm{V}\left(\frac{y_{hi_s}}{x_{hi_s}} X_h \right) = X_h^2 \left(\frac{\sum_{i \in U_h^*} y_{hi}^2 / x_{hi}}{\sum_{i \in U_h^*} x_{hi}} - \left(\frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} \right)^2 \right). \tag{6}
$$

Ved å benytte at $y_{hi} \leq x_{hi}$, slik at

$$
\frac{\sum_{i \in U_h^*} y_{hi}^2 / x_{hi}}{\sum_{i \in U_h^*} x_{hi}} - \left(\frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} \right)^2 \leq \frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} - \left(\frac{\sum_{i \in U_h^*} y_{hi}}{\sum_{i \in U_h^*} x_{hi}} \right)^2 \leq \frac{1}{2} - \left(\frac{1}{2} \right)^2 = \frac{1}{4},
$$

får vi

$$
\mathrm{V}\left(\frac{y_{hi_s}}{x_{hi_s}} X_h \right) \leq \frac{1}{4} X_h^2 .
$$

Dvs. at $\mathrm{V}\left(\hat{Y}_h \right) \leq X_h^2 / 4$ når $n_h = 1$, og det er denne grensen vi vil estimere variansen med. (Teoretisk sett kunne vi funnet en bedre grense ved å maksimere (6) med hensyn på de mulige konfigurasjonene y_{hi} 'ene kan ha. Men hvis N_h er stor kan denne maksimeringen være vanskelig å gjennomføre i praksis).

5.1. Lineærisering

Anta generelt at vi skal estimere variansen $\mathrm{V}\left(\hat{\theta}\right)$, der $\hat{\theta} = g\left(\bar{\boldsymbol{y}}\right)$ er en estimator for $\theta = g\left(\bar{\boldsymbol{Y}}\right)$. Her er $\bar{\boldsymbol{Y}} = \left(\bar{Y}_1, \ldots, \bar{Y}_p\right)$ og $\bar{\boldsymbol{y}} = \left(\bar{y}_1, \ldots, \bar{y}_p\right)$ vektorene av henholdsvis populasjons- og utvalgsgjennomsnitt. Ved å gjøre en 1. ordens Taylorutvikling av g rundt forventningen $\mathrm{E}\bar{\boldsymbol{y}} = \left(\mathrm{E}\bar{y}_1, \ldots., \mathrm{E}\bar{y}_p\right)$ får vi tilnærmingen

$$g\left(\bar{\boldsymbol{y}}\right) \approx g\left(\mathrm{E}\bar{\boldsymbol{y}}\right) + \sum_{j=1}^{p} a_j \left(\bar{y}_j - \mathrm{E}\bar{y}_j\right),$$

der $a_j = \dfrac{\partial g}{\partial y_j}\left(\mathrm{E}\bar{\boldsymbol{y}}\right)$. Dette gir oss følgende tilnærming av variansen til $\hat{\theta}$:

$$\mathrm{V}\left(\hat{\theta}\right) \approx \sum_{j=1}^{p} a_j^2 \mathrm{V}\left(\bar{y}_j\right) + \sum_{j=1}^{p} \sum_{\substack{i=1 \\ i \neq j}}^{p} a_j a_i \mathrm{C}\left(\bar{y}_j, \bar{y}_i\right),$$

der $\mathrm{C}\left(\bar{y}_i, \bar{y}_j\right)$ står for kovariansen til $\bar{y}_i$ og $\bar{y}_j$. Variansen $\mathrm{V}\left(\hat{\theta}\right)$ kan nå estimeres ved å estimere a_j med $\hat{a}_j = \partial g / \partial y_j \left(\bar{\boldsymbol{y}}\right)$ og $\mathrm{V}\left(\bar{y}_j\right)$ og $\mathrm{C}\left(\bar{y}_j, \bar{y}_i\right)$ med passende valgte estimatorer.

I vår situasjon er det variansen til $\hat{Y}_h = g\left(\bar{y}_{s_h}, \bar{x}_{s_h}\right)$, der $g\left(y_1, y_2\right) = \left(y_1 / y_2\right) X_h$, som skal estimeres. Tilnærmingen av $\hat{Y}_h$ blir

(7) $$\hat{Y}_h \approx \frac{\mathrm{E}\bar{y}_{s_h}}{\mathrm{E}\bar{x}_{s_h}} \cdot X_h + \frac{1}{\mathrm{E}\bar{x}_{s_h}} \cdot X_h \cdot \left(\bar{y}_{s_h} - \mathrm{E}\bar{y}_{s_h}\right) - \frac{\mathrm{E}\bar{y}_{s_h}}{\left(\mathrm{E}\bar{x}_{s_h}\right)^2} \cdot X_h \cdot \left(\bar{x}_{s_h} - \mathrm{E}\bar{x}_{s_h}\right),$$

og gir at

$$\mathrm{V}\left(\hat{Y}_h\right) \approx \frac{1}{\left(\mathrm{E}\bar{x}_{s_h}\right)^2} \cdot X_h^2 \cdot \mathrm{V}\left(\bar{y}_{s_h}\right) + \frac{\left(\mathrm{E}\bar{y}_{s_h}\right)^2}{\left(\mathrm{E}\bar{x}_{s_h}\right)^4} \cdot X_h^2 \cdot \mathrm{V}\left(\bar{x}_{s_h}\right) - 2\frac{\mathrm{E}\bar{y}_{s_h}}{\left(\mathrm{E}\bar{x}_{s_h}\right)^3} \cdot X_h^2 \cdot \mathrm{C}\left(\bar{x}_{s_h}, \bar{y}_{s_h}\right).$$

Den foreslåtte variansestimatoren for $\hat{Y}_h$ er dermed

$$\hat{\mathrm{V}}\left(\hat{Y}_h\right) = \frac{1}{\left(\bar{x}_{s_h}\right)^2} \cdot X_h^2 \cdot \hat{\mathrm{V}}\left(\bar{y}_{s_h}\right) + \frac{\left(\bar{y}_{s_h}\right)^2}{\left(\bar{x}_{s_h}\right)^4} \cdot X_h^2 \cdot \hat{\mathrm{V}}\left(\bar{x}_{s_h}\right) - 2\frac{\bar{y}_{s_h}}{\left(\bar{x}_{s_h}\right)^3} \cdot X_h^2 \cdot \hat{\mathrm{C}}\left(\bar{x}_{s_h}, \bar{y}_{s_h}\right),$$

der $\hat{\mathrm{V}}\left(\bar{y}_{s_h}\right)$, $\hat{\mathrm{V}}\left(\bar{x}_{s_h}\right)$ og $\hat{\mathrm{C}}\left(\bar{x}_{s_h}, \bar{y}_{s_h}\right)$ er estimatorer for henholdsvis $\mathrm{V}\left(\bar{y}_{s_h}\right)$, $\mathrm{V}\left(\bar{x}_{s_h}\right)$ og $\mathrm{C}\left(\bar{x}_{s_h}, \bar{y}_{s_h}\right)$.

Vi trenger altså å finne estimatorer for $\mathrm{V}\left(\bar{y}_{s_h}\right)$, $\mathrm{V}\left(\bar{x}_{s_h}\right)$ og $\mathrm{C}\left(\bar{x}_{s_h}, \bar{y}_{s_h}\right)$. Vi ser først på variansen til $\bar{y}_{s_h}$. Denne kan skrives som

$$\mathrm{V}\left(\bar{y}_{s_h}\right)=\frac{1}{n_h^2}\sum_{i\in U_h^*}\sum_{j\in U_h^*} y_{hi}y_{hj}\left(\pi_{hij}-\pi_{hi}\pi_{hj}\right)$$
$$=\frac{1}{n_h^2}\frac{1}{2}\sum_{i\in U_h^*}\sum_{\substack{j\in U_h^*\\ j\neq i}}\left(\pi_{hi}\pi_{hj}-\pi_{hij}\right)\left(y_{hi}-y_{hj}\right)^2,$$

der π_{hij} er sannsynligheten for at bransjeenhet i og j er med i utvalget og $\pi_{hii}=\pi_{hi}$. To estimatorer for denne variansen er

$$\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)=\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{j\in s_h}\frac{y_{hi}y_{hj}\left(\pi_{hij}-\pi_{hi}\pi_{hj}\right)}{\pi_{hij}}$$

og

$$\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)=\frac{1}{2}\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{\substack{j\in s_h\\ j\neq i}}\frac{\pi_{hi}\pi_{hj}-\pi_{hij}}{\pi_{hij}}\left(y_{hi}-y_{hj}\right)^2 .$$

(HT står for Horvitz-Thompson og SYG står for Sen-Yates-Grundy). Begge disse estimatorene har den gode egenskapen at de er forventningsrette hvis π_{hij} er positiv når π_{hi} og π_{hj} er det.[4] Men dessverre har de også den uheldige egenskapen at de kan bli negative.

Problemet med både $\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)$ og $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$ er at vi må kjenne π_{hij}'ene. Når $i=j$ har vi at $\pi_{hij}=\pi_{hii}=\pi_{hi}$, og når $i\neq j$ samtidig som π_{hi} eller $\pi_{hj}=1$ så er $\pi_{hij}=\pi_{hi}\pi_{hj}$. Dermed kjenner vi disse π_{hij}'ene (fordi π_{hi} er kjent). Men når $i\neq j$ samtidig som π_{hi} og $\pi_{hj}\neq 1$, kjenner vi ikke π_{hij}. For å kunne estimere $\mathrm{V}\left(\bar{y}_{s_h}\right)$ med $\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)$ eller $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$ velger vi derfor å tilnærme de ukjente π_{hij}'ene med $\hat{\pi}_{hij}=\frac{N_h^*\left(n_h-1\right)}{n_h\left(N_h^*-1\right)}\cdot\pi_{hi}\pi_{hj}$, der N_h^* er antall bransjeenheter i stratum h som har $\pi_{hi}>0$.[5]

Med denne tilnærmingen av de ukjente π_{hij}'ene får vi at $\pi_{hi}\pi_{hj}-\pi_{hij}\geq 0$ for alle $i\neq j$. Dette medfører at $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$ alltid er positiv. Videre kan det vises at hvis π_{hij} er proporsjonal med $x_{hi}x_{hj}$

---

[4] Hvis $\pi_{hij}>0$ når π_{hi} og $\pi_{hj}>0$, har vi at $\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)=\frac{1}{n_h^2}\sum_{i\in U_h^*}\sum_{j\in U_h^*}\frac{y_{hi}y_{hj}\left(\pi_{hij}-\pi_{hi}\pi_{hj}\right)}{\pi_{hij}}\cdot I_{hi}I_{hj}$, der $I_{hi}=1$ hvis bransjeenhet i er med i utvalget og 0 ellers. Dermed er $\mathrm{E}\left[\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)\right]=\frac{1}{n_h^2}\sum_{i\in U_h^*}\sum_{j\in U_h^*}\frac{y_{hi}y_{hj}\left(\pi_{hij}-\pi_{hi}\pi_{hj}\right)}{\pi_{hij}}\cdot$ $\mathrm{E}\left[I_{hi}I_{hj}\right]=\frac{1}{n_h^2}\sum_{i\in U_h^*}\sum_{j\in U_h^*}\frac{y_{hi}y_{hj}\left(\pi_{hij}-\pi_{hi}\pi_{hj}\right)}{\pi_{hij}}\cdot\pi_{hij}=\mathrm{V}\left(\bar{y}_{s_h}\right)$. En tilsvarende fremgangsmåte viser at også $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\hat{Y}_h\right)$ er forventningsrett.

[5] Denne tilnærmingen har vi fra kursmaterialet til et kurs i modellbasert utvalgsteori (forelest av Ray Chambers).

når $i \neq j$, $i, j \in U_h^*$, blir $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$ større når vi bruker $\hat{\pi}_{hij}$ enn når vi bruker π_{hij} . For $\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)$ får vi det motsatte, dvs. at $\hat{\mathrm{V}}_{\mathrm{HT}}\left(\bar{y}_{s_h}\right)$ blir mindre når vi bruker $\hat{\pi}_{hij}$ enn når vi bruker π_{hij} . Av disse grunner velger vi å estimerer $\mathrm{V}\left(\bar{y}_{s_h}\right)$ med $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$, dvs. vi estimerer $\mathrm{V}\left(\bar{y}_{s_h}\right)$ med

$$(8) \qquad \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right) = \frac{1}{2}\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{\substack{j\in s_h\\ j\neq i}}\frac{\pi_{hi}\pi_{hj}-\hat{\pi}_{hij}}{\hat{\pi}_{hij}}\left(y_{hi}-y_{hj}\right)^2$$

der

$$\hat{\pi}_{hij} = \begin{cases} \pi_{hi} & \text{, når } i = j \\ \dfrac{N_h^*\left(n_h-1\right)}{n_h\left(N_h^*-1\right)}\cdot\pi_{hi}\pi_{hj} & \text{, når } i\neq j,\ \pi_{hi}\neq 1 \text{ og } \pi_{hj}\neq 1 \\ \pi_{hi}\pi_{hj} & \text{, når } i\neq j \text{ og } \pi_{hi} \text{ eller } \pi_{hj}=1\ . \end{cases}$$

På tilsvarende måte som vi har kommet frem til estimatoren (8) for $\mathrm{V}\left(\bar{y}_{s_h}\right)$, kommer vi frem til estimatorene

$$(9) \qquad \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{x}_{s_h}\right) = \frac{1}{2}\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{\substack{j\in s_h\\ j\neq i}}\frac{\pi_{hi}\pi_{hj}-\hat{\pi}_{hij}}{\hat{\pi}_{hij}}\left(x_{hi}-x_{hj}\right)^2$$

og

$$(10) \qquad \hat{\mathrm{C}}_{\mathrm{SYG}}\left(\bar{x}_{s_h},\bar{y}_{s_h}\right) = \frac{1}{2}\frac{1}{n_h^2}\sum_{i\in s_h}\sum_{\substack{j\in s_h\\ j\neq i}}\frac{\pi_{hi}\pi_{hj}-\hat{\pi}_{hij}}{\hat{\pi}_{hij}}\left(x_{hi}-x_{hj}\right)\left(y_{hi}-y_{hj}\right)$$

for $\mathrm{V}\left(\bar{x}_{s_h}\right)$ og $\mathrm{C}\left(\bar{x}_{s_h},\bar{y}_{s_h}\right)$, respektivt. Variansestimatoren for $\hat{Y}_h$ blir dermed

$$(11) \qquad \hat{\mathrm{V}}\left(\hat{Y}_h\right) = \frac{1}{\left(\bar{x}_{s_h}\right)^2}\cdot X_h^2\cdot\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right) + \frac{\left(\bar{y}_{s_h}\right)^2}{\left(\bar{x}_{s_h}\right)^4}\cdot X_h^2\cdot\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{x}_{s_h}\right) - 2\frac{\bar{y}_{s_h}}{\left(\bar{x}_{s_h}\right)^3}\cdot X_h^2\cdot\hat{\mathrm{C}}_{\mathrm{SYG}}\left(\bar{x}_{s_h},\bar{y}_{s_h}\right),$$

der $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{y}_{s_h}\right)$, $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\bar{x}_{s_h}\right)$ og $\hat{\mathrm{C}}_{\mathrm{SYG}}\left(\bar{x}_{s_h},\bar{y}_{s_h}\right)$ er gitt ved henholdsvis (8), (9) og (10). Det kan vises at denne estimatoren alltid er positiv.

Estimatoren (11) gjelder bare for strata hvor $n_h > 1$. Når $n_h = 1$ skal vi som nevnt estimere variansen med den øvre grensen $X_h^2/4$. Ved så å benytte at $\mathrm{V}\left(\hat{Y}_h\right) = 0$ når $n_h = N_h^*$, ender vi opp med følgende estimator for standardavviket (4):

(12) $$\widehat{\mathrm{st}}\left(\hat{d}\right)=\frac{\sqrt{\sum_h \tilde{\mathrm{V}}\left(\hat{Y}_h\right)}}{\sum_h X_h}\cdot 100$$

der

$$\tilde{\mathrm{V}}\left(\hat{Y}_h\right)=\begin{cases} 0 & \text{, når } n_h = N_h^* \\ \frac{1}{4}X_h^2 & \text{, når } n_h = 1 < N_h^* \\ \hat{\mathrm{V}}\left(\hat{Y}_h\right) & \text{, når } 1 < n_h < N_h^* \end{cases}$$

og $\hat{\mathrm{V}}\left(\hat{Y}_h\right)$ er gitt ved (11).

Vi har sett at $\hat{Y}_h$ generelt ikke er forventningsrett når $n_h = 1$, og vi skal nå vise at det samme gjelder for stratum der $n_h > 1$. Når $n_h > 1$ har vi ikke noe eksakt uttrykk for forventningen, men fra tilnærmingen (7) får vi at den kan tilnærmes med

(13) $$\begin{aligned}\mathrm{E}\hat{Y}_h &\approx \frac{\mathrm{E}\overline{y}_{s_h}}{\mathrm{E}\overline{x}_{s_h}}\cdot X_h \\ &= \frac{\sum_{i\in U_h^*} y_{hi}\pi_{hi}}{\sum_{i\in U_h^*} x_{hi}\pi_{hi}}\cdot X_h \ .\end{aligned}$$

Dette tilnærmede uttrykket er generelt ikke lik Y_h. Vi kan nemlig, for de fleste verdier av x_{hi} 'ene, finne konfigurasjoner av y_{hi} 'ene som gjør at det tilnærmede uttrykket enten er større eller mindre enn Y_h. Anta f.eks. at vi har et stratum der $N_h = 5$, $n_h = 2$, x_{hi} 'ene er gitt ved 100, 120, 130, 130, 170 og de respektive y_{hi} 'verdiene er 0, 120/2, 130/2, 130/2 og 170. Dette gir at $\left(\mathrm{E}\overline{y}_{s_h} / \mathrm{E}\overline{x}_{s_h}\right)\cdot X_h \approx 396$ mens $Y_h = 360$. Dvs. at den tilnærmede forventningen i dette tilfellet er større enn Y_h. Er derimot de respektive y_{hi} 'verdiene gitt ved 100, 120/2, 130/2, 130/2 og 0, får vi at $\left(\mathrm{E}\overline{y}_{s_h} / \mathrm{E}\overline{x}_{s_h}\right)\cdot X_h \approx 255$ og $Y_h = 290$. Dvs. at nå er den tilnærmede forventningen mindre enn Y_h. Under forutsettingen om at tilnærmingen (13) bra nok, har vi derfor at $\hat{Y}_h$ generelt ikke er forventningsrett.

Fordi $\hat{Y}_h$ ikke er forventningsrett er $\hat{d}$ ikke forventningsrett. Dette betyr at forventningsskjevheten til $\hat{d}$, $\mathrm{Bias}\left(\hat{d}\right)$, ikke er 0 for alle konfigurasjoner som y_{hi} 'ene kan ha. Vi skal nå finne en øvre og nedre grense for denne forventningsskjevheten. Vi finner at

$$\mathrm{Bias}\left(\hat{d}\right)=\frac{\sum_h \mathrm{Bias}\left(\hat{Y}_h\right)}{\sum_h X_h}\cdot 100,$$

der $\mathrm{Bias}\left(\hat{Y}_h\right)=\mathrm{E}\left[\hat{Y}_h - Y_h\right]$ er forventningsskjevheten til $\hat{Y}_h$. For strata der $n_h = 1$ får vi fra (5) at

$$\text{Bias}\left(\hat{Y}_h\right) = \sum_{i \in U_h} c_{hi} y_{hi} \,,$$

der $c_{hi} = \sum_{i \notin U_h^*} x_{hi} \Big/ \sum_{i \in U_h^*} x_{hi}$ når $i \in U_h^*$ og -1 når $i \notin U_h^*$. (Når $U_h^* = U_h$ blir $\text{Bias}\left(\hat{Y}_h\right) = 0$ som det skal). For strata der $n_h > 1$ benytter vi tilnærmingen (13) og får

$$\begin{aligned}
\text{Bias}\left(\hat{Y}_h\right) &\approx \frac{\sum_{i \in U_h^*} y_{hi} \pi_{hi}}{\sum_{i \in U_h^*} x_{hi} \pi_{hi}} \cdot X_h - Y_h \\
&= \sum_{i \in U_h} b_{hi} y_{hi} \,,
\end{aligned}$$

der $b_{hi} = \dfrac{\pi_{hi} X_h}{\sum_{i \in U_h^*} x_{hj} \pi_{hj}} - 1$. Fordi $\text{Bias}\left(\hat{Y}_h\right) = 0$ når $n_h = N_h$ gir dette oss

(14) $$\text{Bias}\left(\hat{d}\right) \approx \frac{\sum_{h \text{ s.a. } 1 < n_h < N_h} \left(\sum_{i \in U_h} b_{hi} y_{hi} \right) + \sum_{h \text{ s.a. } n_h = 1 < N_h} \left(\sum_{i \in U_h} c_{hi} y_{hi} \right)}{\sum_h X_h} \cdot 100.$$

Ved å benytte at $0 \le b_{hi} y_{hi} \le b_{hi} x_{hi}$ når $b_{hi} \ge 0$, $b_{hi} x_{hi} \le b_{hi} y_{hi} \le 0$ når $b_{hi} \le 0$, og $-\sum_{i \notin U_h^*} x_{hi} \le \sum_{i \in U_h} c_{hi} y_{hi} \le \sum_{i \notin U_h^*} x_{hi}$, får vi ulikheten

$$\text{L} \le \frac{\sum_{h \text{ s.a. } 1 < n_h < N_h} \left(\sum_{i \in U_h} b_{hi} y_{hi} \right) + \sum_{h \text{ s.a. } n_h = 1 < N_h} \left(\sum_{i \in U_h} c_{hi} y_{hi} \right)}{\sum_h X_h} \cdot 100 \le \text{U}$$

der

$$\text{L} = \frac{\sum_{h \text{ s.a. } 1 < n_h < N_h} \left(\sum_{i \in A_h^c} b_{hi} x_{hi} \right) - \sum_{h \text{ s.a. } n_h = 1 < N_h} \left(\sum_{i \notin U_h^*} x_{hi} \right)}{\sum_h X_h} \cdot 100 \,,$$

$$\text{U} = \frac{\sum_{h \text{ s.a. } 1 < n_h < N_h} \left(\sum_{i \in A_h} b_{hi} x_{hi} \right) + \sum_{h \text{ s.a. } n_h = 1 < N_h} \left(\sum_{i \notin U_h^*} x_{hi} \right)}{\sum_h X_h} \cdot 100 \,,$$

$A_h = \{i \text{ s.a. } b_{hi} \geq 0\}$ og $A_h^c = \{i \text{ s.a. } b_{hi} < 0\}$. Under forutsetingen at tilnærmingen (13) er bra nok kan vi dermed bruke L som en nedre grense og U som en øvre grense for forventningsskjevheten til $\hat{d}$, dvs. at

$$\mathrm{L} \leq \mathrm{Bias}\left(\hat{d}\right) \leq \mathrm{U}.$$

Vi vil bemerke at grensene L og U avhenger av sysselsettingen til bransjeenheten i populasjonen og av n_h'ene. Derfor kan grensene bli forskjellige fra et kvartal til et annet hvis populasjonsdataene eller n_h'ene har endret seg. Vi vil også presisere at forventningsskjevheten avhenger av hvilke verdier y_{hi}'ene har. Selv om det viser seg at grensene L og U er store, fins det veldig mange konfigurasjoner av y_{hi}'ene som bare gir liten forventningsskjevhet.

Hvis vi har en konfigurasjon av y_{hi}'ene slik at forventningsskjevheten til $\hat{d}$ er tilnærmet 0, kan vi lage til konfidensintervall for d. Vi tar da utgangspunkt i at

$$\frac{\hat{d} - d}{\mathrm{st}\left(\hat{d}\right)} \approx \frac{\sum_h \left\{\left(\hat{Y}_h - Y_h\right) - \mathrm{E}\left[\hat{Y}_h - Y_h\right]\right\}}{\sqrt{\sum_h \mathrm{V}\left(\hat{Y}_h - Y_h\right)}}$$

når $\mathrm{Bias}\left(\hat{d}\right) \approx 0$. Under visse betingelser på $\mathrm{V}\left(\hat{Y}_h - Y_h\right)$ får vi derfor fra Lindebergsetningen at $\left(\hat{d} - d\right) \big/ \mathrm{st}\left(\hat{d}\right)$ er tilnærmet standardnormalfordelt. Et tilnærmet 95% konfidensintervall for d er dermed gitt ved $\hat{d} \pm 1.96 \cdot \mathrm{st}\left(\hat{d}\right)$. Fordi $\mathrm{st}\left(\hat{d}\right)$ er ukjent putter vi inn estimatoren $\widehat{\mathrm{st}}\left(\hat{d}\right)$ for denne og bruker

(15) $$\hat{d} \pm 1.96 \cdot \widehat{\mathrm{st}}\left(\hat{d}\right)$$

som et 95% konfidensintervall for d.

5.2. Bootstrap

Enkelt sagt går Bootstrap ut på å trekke nye utvalg fra det opprinnelige utvalget. På bakgrunn av disse utvalgene beregnes det nye estimater som benyttes til å estimere variansen til den opprinnelige estimatoren. Ideen er at utvalgene skal trekkers på en slik måte at Bootstrapestimatoren for variansen konvergerer mot en tilfredsstillende variansestimator når vi trekker uendelig med utvalg.

Vi er interessert i en Bootstrapmetode for å estimere variansen til $\hat{Y}_h = \left(\bar{y}_{s_h} / \bar{x}_{s_h}\right) \cdot X_h$. Det finnes en del metoder for denne typen estimatorer, men de bygger alle på et enkelt tilfeldig utvalg. Hvis vi benytter en av disse metodene får vi antakeligvis en estimator som er lite egnet for å estimere variansen til $\hat{Y}_h$. Det er også beskrevet noen Bootstrapmetoder som tar for seg situasjonen med ulike trekksannsynligheter (Rao og Wu, 1988, og Sitter, 1992). Men disse metodene er beregnet på en annen type estimator en den vi har. Vi har ikke funnet en metode som kombinerer ulike trekksannsynligheter med vår type estimator.

6. Modellbasert analyse av estimert diffusjonsindeks

I dette avsnittet skal vi gjøre en modellbasert analyse av $\hat{d}$. Dvs. at vi nå ser på bransjeenhetenes konjunktursysselsettinger som stokastiske variabler, mens utvalget antas gitt. Dermed er det ikke bare $\hat{Y}_h$'ene og $\hat{d}$ som er stokastiske, men også Y_h'ene og d. Dette betyr at vi skal anslå, eller predikere, verdien til en stokastisk variabel. Av den grunn kalles gjerne $\hat{Y}_h$'ene og $\hat{d}$ for prediktorer. Som i forrige avsnitt ser vi bort fra frafall, målefeil og registerfeil, og vi tenker oss at vi har utvalg fra alle strata.

Fordi et utvalg vanligvis trekkes stokastisk, kan det kanskje virke litt rart at vi nå skal se på utvalget som gitt og i stedet anta en modell for populasjonsvariablene. Men å anta en modell for populasjonsvariablene er ikke noe nytt. Veldig ofte velges estimatoren og utvalgsplanen på bakgrunn av blant annet en slik antagelse om hvordan populasjonsvariablene fordeler seg. Videre har vi spesielt for konjunkturbarometeret at det samme utvalget benyttes over en lengre periode på minst fire kvartal. For hvert kvartal observeres det nye verdier for konjunktursysselsettingen til de utvalgte bransjeenhetene. Derfor er det meningsfylt å gjøre en analyse der vi ser på utvalget som gitt og konjunktursysselsettingene som stokastiske variabler som følger en fordeling. Vi kan også argumentere for modelleringen ved hjelp av Likelihoodprinsippet (LP). LP sier at to proporsjonale likelihoodfunksjoner skal gi samme statistiske analyse, og fører til at vi må foreta en modellering av populasjonen for å kunne utføre en informativ analyse. (For mer om LP, se Bjørnstad, 1995).

Nå ser vi altså på y_{hi} som en stokastisk variabel som kan anta verdiene x_{hi}, $x_{hi}/2$ og 0. Vi lar p_s, p_u og p_m være sannsynlighetene for at y_{hi} inntar disse verdiene, respektivt. Dvs. at p_s er sannsynligheten for at $y_{hi} = x_{hi}$, altså at bransjeenheten forventer en økning i produksjonen, p_u er sannsynligheten for at $y_{hi} = x_{hi}/2$, altså at bransjeenheten ikke forventer noen endring, og p_m er sannsynligheten for at $y_{hi} = 0$, altså at bransjeenheten forventer en nedgang. Vi finner at forventningen og variansen til y_{hi} henholdsvis kan skrives som

$$\mathrm{E}y_{hi} = \beta x_{hi}$$

og

$$\mathrm{V}(y_{hi}) = \sigma^2 x_{hi}^2 ,$$

der $\beta = p_s + (1/2)\,p_u$ og $\sigma^2 = p_s + (1/4)\,p_u - \left(p_s + (1/2)\,p_u\right)^2$.

Verdiene til p_s, p_u og p_m kan variere fra konjunktursysselsetting til konjunktursysselsetting. Men fordi bransjeenheter innen samme stratum driver innen samme type næring og har relativt lik sysselsetting, er det ikke urimelig å anta at disse sannsynlighetene er tilnærmet like for alle konjunktursysselsettingene i stratumet. Dermed vil også β og σ^2 være lik for alle y_{hi} i stratumet, la oss si β_h og σ_h^2 for stratum h. Dette gir oss følgende modell, som vi vil kalle modell ξ:

$$\mathrm{E}y_{hi} = \beta_h x_{hi} \quad , \forall \;\; i \in \text{stratum } h$$

$$\mathrm{V}(y_{hi}) = \sigma_h^2 x_{hi}^2 \;, \forall \;\; i \in \text{stratum } h .$$

I resten av dette avsnittet er det denne modellen vi vil anta. I tillegg vil vi anta at konjunktursysselsettingene er uavhengige av hverandre.

I populasjonen finnes det noen bransjeenheter som har sysselsetting $x_{hi} = 0$, dvs. $y_{hi} = x_{hi} = 0$. For slike bransjeenheter er $p_s = p_u = p_m = 1$, og disse sannsynligheten er ikke lik de tilsvarende sannsynlighetene for de andre bransjeenhetene i stratumet. Dermed holder ikke antagelsene som leder til modell ξ for disse bransjeenhetene. Men fordi $\mathrm{E}y_{hi} = 0 = \beta_h x_{hi}$ og $\mathrm{V}(y_{hi}) = 0 = \sigma_h^2 x_{hi}^2$ når $x_{hi} = 0$, omfatter modellen også disse konjunktursysselsettingene.
Vi kan se på $\bar{y}_{s_h} / \bar{x}_{s_h}$ som en estimator for β_h, og innfører derfor notasjonen

$$\hat{\beta}_h = \frac{\bar{y}_{s_h}}{\bar{x}_{s_h}},$$

slik at

$$\hat{Y}_h = \hat{\beta}_h X_h .$$

Ved å bruke at utvalget s_h er gitt, finner vi at

$$\mathrm{E}\hat{\beta}_h = \beta_h .$$

Dermed er

$$\mathrm{E}\left[\hat{Y}_h - Y_h\right] = \beta_h X_h - \beta_h X_h = 0$$

og

$$\begin{aligned}\mathrm{E}\left[\hat{d} - d\right] &= \mathrm{E}\left(\frac{\sum_h \left(\hat{Y}_h - Y_h\right)}{\sum_h X_h} \cdot 100\right) \\ &= \frac{\sum_h \mathrm{E}\left[\hat{Y}_h - Y_h\right]}{\sum_h X_h} \cdot 100 = 0 \ .\end{aligned}$$

Fordi dette gjelder for alle utvalg s_h og alle β_h og σ_h, har vi at $\hat{Y}_h$ og $\hat{d}$ er forventningsrette.

I stedet for variansen til $\hat{d}$ er vi interessert i prediksjonsvariansen, dvs. variansen til prediksjonsfeilen $\hat{d} - d$. Denne kan skrives som

$$\begin{aligned}\mathrm{V}\left(\hat{d} - d\right) &= \mathrm{V}\left(\frac{\sum_h \left(\hat{Y}_h - Y_h\right)}{\sum_h X_h} \cdot 100\right) \\ &= \frac{\sum_h \mathrm{V}\left(\hat{Y}_h - Y_h\right)}{\left(\sum_h X_h\right)^2} \cdot 100^2,\end{aligned}$$

der den siste likheten følger av at vi har uavhengighet mellom konjunktursysselsettingene. Vi angir så usikkerheten til $\hat{d}$ med standardavviket

$$\text{st}\left(\hat{d}-d\right)=\sqrt{\text{V}\left(\hat{d}-d\right)}$$

(16)
$$=\frac{\sqrt{\sum_h \text{V}\left(\hat{Y}_h-Y_h\right)}}{\sum_h X_h}\cdot 100\ .$$

Før vi estimerer dette standardavviket, skal vi gi følgende begrunnelse for å bruke prediksjonsvariansen i stedet for variansen når vi skal angi usikkerheten til en prediktor: Anta først at vi klarer å predikere d perfekt, dvs. har en prediktor $\tilde{d}=d$. Da bør vi ha et usikkerhetsmål som er null fordi vi er sikre på å predikere d eksakt. Men måler vi usikkerheten med variansen til $\tilde{d}$, dvs. med $\text{V}\left(\tilde{d}\right)=\text{V}(d)$, har vi et usikkerhetsmål som indikerer stor usikkerhet hvis variansen til d er stor. Bruker vi derimot prediksjonsvariansen $\text{V}\left(\tilde{d}-d\right)=0$, har vi et mål som indikerer at det ikke er noen usikkerhet, akkurat som vi ønsker. Anta nå at vi kjenner forventningen til d, og at vi predikerer d med denne, dvs. $\tilde{d}=\text{E}d$. I denne situasjonen ønsker vi et usikkerhetsmål som indikerer stor usikkerhet hvis variansen til d er stor, fordi da er vi ikke så sikre på å predikere d bra. Men variansen til $\tilde{d}$ er null, selv om variansen til d er stor. Med prediksjonsvariansen $\text{V}\left(\tilde{d}-d\right)=\text{V}(d)$ får vi derimot det vi ønsker.

Vi skal nå estimere standardavviket (16), og det gjør vi ved å estimere prediksjonsvariansene $\text{V}\left(\hat{Y}_h-Y_h\right)$. Ved å benytte at $\hat{Y}_h-Y_h=\hat{\beta}_h\sum_{i\in r_h}x_{hi}-\sum_{i\in r_h}y_{hi}$ finner vi at

(17)
$$\text{V}\left(\hat{Y}_h-Y_h\right)=\sigma_h^2\left\{\left(\frac{\sum_{i\in r_h}x_{hi}}{\sum_{i\in s_h}x_{hi}}\right)^2\sum_{i\in s_h}x_{hi}^2+\sum_{i\in r_h}x_{hi}^2\right\},$$

der r_h er alle bransjeenhetene i stratum h som ikke er med i utvalget. Dermed kan vi estimere prediksjonsvariansen ved å estimere σ_h^2.

Fra regresjonsteori har vi at en forventningsrett estimator for σ_h^2 er

$$\hat{\sigma}_h^2=\frac{1}{n_h-1}\sum_{i\in s_h}\frac{1}{x_{hi}^2}\left(y_{hi}-\tilde{\beta}_h x_{hi}\right)^2,$$

der

$$\tilde{\beta}_h=\frac{1}{n_h}\sum_{i\in s_h}\frac{y_{hi}}{x_{hi}}.$$

Dermed er

$$\hat{\sigma}_h^2\left\{\left(\frac{\sum_{i\in r_h}x_{hi}}{\sum_{i\in s_h}x_{hi}}\right)^2\sum_{i\in s_h}x_{hi}^2+\sum_{i\in r_h}x_{hi}^2\right\}$$

en forventningsrett estimator for (17) (forutsatt at $n_h > 1$). Når $n_h = 1$ estimerer vi σ_h^2 med

$$\tilde{\sigma}_h^2 = \frac{1}{n_g - |g|} \sum_{l \in g} \sum_{i \in s_l} \frac{1}{x_{li}^2} \left(y_{li} - \tilde{\beta}_l x_{li} \right)^2,$$

der g betegner gruppen av strata som har samme sysselsettingsintervall som stratum h, n_g et antall utvalgte bransjeenheter i denne gruppen, og $|g|$ er antall strata i gruppen. Denne estimatoren vil være forventningsrett for σ_h^2 dersom σ_l^2 er lik for alle strata l som er med i gruppen.

Estimatoren for standardavviket (16) er dermed

$$\text{(18)} \qquad \widehat{\mathrm{st}}\left(\hat{d} - d\right) = \frac{\sqrt{\sum_h \widehat{\mathrm{V}}\left(\hat{Y}_h - Y_h\right)}}{\sum_h X_h} \cdot 100,$$

der

$$\text{(19)} \qquad \widehat{\mathrm{V}}\left(\hat{Y}_h - Y_h\right) = \begin{cases} \hat{\sigma}_h^2 \left\{ \left(\dfrac{\sum_{i \in r_h} x_{hi}}{\sum_{i \in s_h} x_{hi}} \right)^2 \sum_{i \in s_h} x_{hi}^2 + \sum_{i \in r_h} x_{hi}^2 \right\}, & \text{hvis } n_h > 1 \\ \tilde{\sigma}_h^2 \left\{ \left(\sum_{i \in r_h} x_{hi} \right)^2 + \sum_{i \in r_h} x_{hi}^2 \right\}, & \text{hvis } n_h = 1 \end{cases}$$

er en estimator for prediksjonsvariansen (17).

I tillegg til å estimere d med estimatoren $\hat{d}$ kan vi gi et konfidensintervall for d. Vi har at

$$\frac{\hat{d} - d}{\mathrm{st}\left(\hat{d} - d\right)} = \frac{\sum_h \left(\hat{Y}_h - Y_h\right)}{\sqrt{\sum_h \mathrm{V}\left(\hat{Y}_h - Y_h\right)}}.$$

Under visse betingelser på $\mathrm{V}\left(\hat{Y}_h - Y_h\right)$ får vi derfor fra Lindebergsetningen at $\left(\hat{d} - d\right) / \mathrm{st}\left(\hat{d} - d\right)$ er tilnærmet standardnormalfordelt. Et tilnærmet 95% konfidensintervall for d er dermed gitt ved $\hat{d} \pm 1.96 \cdot \mathrm{st}\left(\hat{d} - d\right)$. Fordi $\mathrm{st}\left(\hat{d} - d\right)$ er ukjent putter vi inn estimatoren $\widehat{\mathrm{st}}\left(\hat{d} - d\right)$ for denne og bruker

$$\text{(20)} \qquad \hat{d} \pm 1.96 \cdot \widehat{\mathrm{st}}\left(\hat{d} - d\right)$$

som et 95% konfidensintervall for d.

7. Talleksempel

Vi skal nå benytte usikkerhetsmålene fra avsnitt 5 og 6 til å estimere usikkerheten i konjunkturbarometeret. Perioden vi ser på er 2., 3., 4. kvartal 1999, 1., 2., 3., 4. kvartal 2000 og 1. og 2. kvartal 2002. Dvs. at vi for disse kvartalene skal angi usikkerheten til den estimerte diffusjonsindeksen med det designbaserte standardavviket $\widehat{\mathrm{st}}\left(\hat{d}\right)$, gitt ved (12), og det modellbaserte standardavviket $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$, gitt ved (18).

Som nevnt skal det ideelt trekkes ett nytt utvalg hvert år. Dette er ikke gjort for 2000, hvilket betyr at utvalget fra 1999 også brukes i 2000. I denne tiden det er kun foretatt supplering/rullering av noen få bransjeenheter i enkelte kvartal. F.eks. er utvalget for 1. kvartal 2000 supplert med 6 nye bransjeenheter mens 7 bransjeenheter fra forrige kvartals utvalg er fjernet.

Frafallet varierer fra kvartal til kvartal og er større i 1999 og 2000 enn i 2002. I 1999 og 2000 varierer frafallet fra 125 til 158 bransjeenheter, mens det i 2002 er på 84 og 97 bransjeenheter for henholdsvis 1. og 2. kvartal. Fordi bruttoutvalget ligger på ca. 720 i 1999 og 2000 og på ca. 660 for 2002, tilsvarer dette et kvartalsvis frafall på omtrent 19% i 1999 og 2000 mot ca. 14% i 2002.[6]

Populasjonen består av ca. 30000 bransjeenheter. Bransjeenhetene er fordelt på ca. 270 strata, hvorav det skal være fulltelling i ca. 50 (dvs. i de strataene som består av bransjeenhetene som har 300 eller flere sysselsatte). Blant de resterende strataene hvor det ikke er fulltelling, blir utvalgsenhetene fordelt etter en optimal metode. Som nevnt sikrer ikke denne allokeringen av utvalgsenhetene at vi får utvalg fra alle strata, og i den perioden vi ser på er det ca. 70 strata som ikke har bruttoutvalg. Hvis vi også teller med de strataene hvor vi har frafall av alle utvalgsenhetene, har vi ca. 85 strata uten utvalg (dvs. uten nettoutvalg).

Når vi har beregnet den estimerte diffusjonsindeksen $\hat{d}$ og standardavvikene $\widehat{\mathrm{st}}\left(\hat{d}\right)$ og $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$, har vi fjernet fra populasjonen de strataene hvor vi ikke har utvalg. Strengt tatt betyr dette at vi estimerer diffusjonsindeksen til den delmengden av populasjonen hvor vi har utvalg fra. Dette bety også at $\widehat{\mathrm{st}}\left(\hat{d}\right)$ og $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$ er usikkerhetsmål for $\hat{d}$ som en estimator for diffusjonsindeksen til denne delmengden av populasjonen, og ikke usikkerhetsmål for $\hat{d}$ som en estimator for diffusjonsindeksen til hele populasjonen. Som stratifiseringsintervall for sysselsettingen bruker vi 0-99, 100-199, 200-299 og 300-∞ for alle næringshovedgrupper, selv om noen kan ha en litt annen inndeling.

Populasjonsdataene for 1999 og 2000, som er holdt konstant i disse to årene, er aggregert på stratumnivå. Dette betyr at vi ikke får oppgitt hver enkelt bransjeenhets sysselsetting, men kun den totale sysselsettingen innen hvert stratum. Vi kan derfor ikke beregne det modellbaserte standardavviket $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$ for 1999 og 2000, fordi dette standardavviket krever at vi kjenner sysselsettingen til hver enkelt bransjeenhet i populasjonen. For 2002 derimot er populasjonsdataene på mikronivå, dvs. for hver bransjeenhet. Vi kan derfor beregne standardavviket $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$ for 1. og 2. kvartal 2002.

[6] Her har vi regnet en bransjeenhet som frafall hvis bransjeenheten ikke har svart på spørsmålet om forventet endring i neste kvartals produksjon, selv om bransjeenheten har svart på andre spørsmål.

I tillegg til å estimere diffusjonsindeksen for hele populasjonen, har vi estimert diffusjonsindeksen for noen spesielle grupper av strata. De gruppene vi har sett på kalles for E1, E2 og E5. Gruppen E1 omfatter innsatsvarer, E2 investeringsvarer og E5 konsumvarer.

Vi skulle også ha estimert diffusjonsindeksen for gruppen E6. Denne gruppen omfatter energivarer og består av 7 strata som til sammen har ca. 35 bransjeenheter. Av disse 7 strataene har vi utvalg fra kun 2 og 3 strata. (3 strata i 2., 3., 4. kvartal 1999 og 1., 2. kvartal 2002, og 2 strata i 1., 2., 3., 4. kvartal 2000). Fordi de strataene hvor vi har utvalg fra består av kun en bransjeenhet hver, dvs. $N_h = 1$, har vi fulltelling i disse strataene. Dermed er den estimerte diffusjonsindeksen $\hat{d}$ identisk med diffusjonsindeksen for de strataene hvor vi har utvalg fra. Videre får vi at $\widehat{\mathrm{st}}\left(\hat{d}\right) = \widehat{\mathrm{st}}\left(\hat{d} - d\right) = 0$, og disse usikkerhetsmålet er helt korrekt å bruke hvis vi er ute etter å estimere diffusjonsindeksen til bare de strataene i gruppen hvor vi har utvalg fra. Men om vi er ute etter å estimere diffusjonsindeksen til hele gruppen, vil det være misvisende å angi usikkerheten med 0. Av denne grunn har vi unnlatt å presentere tall for gruppen E6.

Estimatene vi har fått for diffusjonsindeksen, det designbaserte standardavviket og konfidensintervallet (15), er gitt i Tabell 1. Hvis vi først ser på de estimerte diffusjonsindeksene, og sammenligner disse fra kvartal til kvartal, ser vi at samtlige estimater er blitt større fra 2. til 3. kvartal 1999, mindre fra 3. til 4. kvartal 1999, og så videre ut 2000. Eneste unntak er for gruppen E2 der estimatet er blitt større fra 3. til 4. kvartal 2000 mens estimatene for de øvrige gruppene er blitt mindre. Vi ser også at av de i alt 28 estimerte diffusjonsindeksene i 1999 og 2000 er 19 estimater større enn 50 mens 9 er mindre enn 50. Spesielt har vi for gruppen E1 at estimatene er større enn 50 gjennom hele denne perioden. Dette kan tyde på at det gjennomgående har vært mer optimisme enn pessimisme i populasjonen gjennom 1999 og 2000, og at særlig næringene i gruppen E1 har hatt gode utsikter.

Ser vi nå på de estimerte standardavvikene $\widehat{\mathrm{st}}\left(\hat{d}\right)$ i Tabell 1, finner vi at standardavvikene som gjelder for hele populasjonen ligger mellom 1.43 og 1.59. Ut fra disse tallene kan vi nok si at det er relativt liten usikkerhet i estimatoren for diffusjonsindeksen til hele populasjonen. For gruppe E1 og E5 er de estimerte standardavvikene noe større, og ligger rundt 2.40 for E1 og 2.25 for E5. (Med unntak av 2. kvartal 2002 så er $\widehat{\mathrm{st}}\left(\hat{d}\right)$ større for gruppen E1 enn for E5). De største estimerte standardavvikene har vi for gruppen E2. Disse standardavvikene varierer fra 3.02 til 3.64, og tyder på at det er en del usikkerhet i estimeringen av diffusjonsindeksen til denne gruppen. Som en følge av at $\widehat{\mathrm{st}}\left(\hat{d}\right)$ er blitt minst for hele populasjonen og størst for E2, er konfidensintervallene i Tabell 1 smalest for populasjonen og videst for gruppe E2.[7]

Som nevnt er det estimerte standardavviket $\widehat{\mathrm{st}}\left(\hat{d}\right)$ blitt mindre for hele populasjonen enn for de andre gruppene. Det er bare gruppen E6 som har fått et estimert standardavvik som er mindre enn det estimerte standardavviket til hele populasjonen. (Som nevnt er $\widehat{\mathrm{st}}\left(\hat{d}\right) = 0$ for gruppen E6). Faktisk er det en sammenheng mellom $\widehat{\mathrm{st}}\left(\hat{d}\right)$ til hele populasjonen og til de andre gruppene. Dette følger av at gruppene E1, E2, E5 og E6 er en partisjon av populasjonen, dvs. at gruppene til sammen utgjør hele populasjonen og at det ikke er noen overlapp av strata mellom gruppene. For å se sammenhengen lar

[7] Fordi vi ikke vet om y_{hi} 'ene i populasjonen er slik at den designbaserte forventningsskjevheten til $\hat{d}$ er tilnærmet 0, vet vi heller ikke om disse konfidensintervallene kan sees på som 95% konfidensintervall.

vi d_p, d_1, d_2, d_5 og d_6 stå for diffusjonsindeksen til hele populasjonen og til gruppe E1, E2, E5 og E6, respektivt. Vi lar så $\hat{d}_p$, $\hat{d}_1$, $\hat{d}_2$, $\hat{d}_5$ og $\hat{d}_6$ stå for de tilsvarende estimatorene. Med denne notasjonen kan kvadratet av det estimerte standardavviket til populasjonen skrives som

$$(21)\qquad \widehat{\mathrm{st}}\left(\hat{d}_p\right)^2 = \left(\frac{X_1}{X_p}\right)^2 \widehat{\mathrm{st}}\left(\hat{d}_1\right)^2 + \left(\frac{X_2}{X_p}\right)^2 \widehat{\mathrm{st}}\left(\hat{d}_2\right)^2 + \left(\frac{X_5}{X_p}\right)^2 \widehat{\mathrm{st}}\left(\hat{d}_5\right)^2 + \left(\frac{X_6}{X_p}\right)^2 \widehat{\mathrm{st}}\left(\hat{d}_6\right)^2,$$

der X_p, X_1, X_2, X_5 og X_6 står for den total sysselsetting i henholdsvis populasjonen, gruppe E1, E2, E5 og E6. Fordi X_i / X_p, $i = 1, 2, 5, 6$, alltid er mindre enn 1, vil vi ofte få at det estimerte standardavviket for hele populasjonen er mindre en det estimerte standardavviket for en gruppe.

En tilsvarende sammenheng som (21) gjelder også for det modellbaserte estimerte standardavviket $\widehat{\mathrm{st}}\left(\hat{d} - d\right)$. (Men fordi vi slår sammen flere strata ved estimeringen av prediksjonsvariansen til $\hat{Y}_h$ når $n_h = 1 < N_h$, og dermed bruker vi flere strata når vi ser på hele populasjonen enn når vi ser på en gruppe, er sammenhengen ikke helt eksakt for $\widehat{\mathrm{st}}\left(\hat{d} - d\right)$). Dermed vil vi ofte få at også $\widehat{\mathrm{st}}\left(\hat{d} - d\right)$ er mindre for hele populasjonen enn for gruppene.

Tallene vi har fått for det modellbaserte standardavviket $\widehat{\mathrm{st}}\left(\hat{d} - d\right)$ og det modellbaserte konfidensintervallet (20) er vist i Tabell 2. I 1. kvartal 2002 er de estimerte standardavvikene blitt 1.17, 1.91, 2.63 og 1.69 for henholdsvis populasjonen, gruppen E1, E2 og E5. I 2. kvartal 2002 er de tilsvarende estimatene blitt 1.28, 1.81, 2.79 og 2.17. Det estimerte standardavviket er som vi ser mindre for hele populasjonen enn for gruppene, som ventet. (Det eneste unntaket er gruppen E6, som har et estimert standardavvik på 0). Videre er $\widehat{\mathrm{st}}\left(\hat{d} - d\right)$ størst for gruppen E5. Som en følge av dette har vi fått de smaleste konfidensintervallene for diffusjonsindeksen til hele populasjonen og de videste konfidensintervallene for diffusjonsindeksen til gruppen E5.

Det designbaserte standardavviket $\mathrm{st}\left(\hat{d}\right)$ og det modellbaserte standardavviket $\mathrm{st}\left(\hat{d} - d\right)$ måler ikke den samme usikkerheten. Usikkerheten som $\mathrm{st}\left(\hat{d}\right)$ måler kommer av at det er flere mulige utvalg som kan trekkes, mens usikkerheten som $\mathrm{st}\left(\hat{d} - d\right)$ måler kommer av at y_{hi}'ene antas å være stokastiske variabler som kan anta flere verdier. Hvis vi sammenligner estimatene vi har fått for disse to standardavvikene, ser vi at estimatene er mindre for $\mathrm{st}\left(\hat{d} - d\right)$ enn for $\mathrm{st}\left(\hat{d}\right)$. Dette kan tyde på at usikkerheten som måles med $\mathrm{st}\left(\hat{d} - d\right)$ er mindre enn usikkerheten som måles med $\mathrm{st}\left(\hat{d}\right)$. Men fordi vi har brukt konservative estimatorer for $\mathrm{V}\left(\hat{Y}_h\right)$ ved estimeringen av $\mathrm{st}\left(\hat{d}\right)$, er det også mulig at usikkerhetene som måles med de to standardavvikene relativt like.

Som vi har nevnt er ikke $\hat{d}$ designforventningsrett. For å få tall på hvor stor forventningsskjevheten kan bli på det verste, har vi beregnet den nedre og øvre grensen fra underavsnitt 5.1. Dette gav oss den nedre grensen -19.23 og den øvre grensen 19.23 i 1. og 2. kvartal 2002.[8] (Fordi vi trenger sysselsettingen til hver bransjeenhet for å beregne grensene, kan vi bare beregne disse for 2002). Disse grensene er veldig store. Men det betyr ikke nødvendigvis at forventningsskjevhetene for 1. og 2. kvartal 2002 er store, det kan godt tenkes at y_{hi} 'verdiene for de to kvartalene er slik at forventningsskjevhetene er små.

For å se hvordan forventningsskjevheten blir for ulike konfigurasjoner av y_{hi} 'ene har vi simulert 10000 tilfeldige konfigurasjoner. For hver av disse har vi beregnet den tilnærmede forventningsskjevheten (14) for 1. kvartal 2002. Dette gav oss 10000 forventningsskjevheter som alle lå mellom -1.664 og 1.656. Bare 106 av disse var mindre enn -1 eller større enn 1, mens hele 8030 lå mellom -0.5 og 0.5. Dette indikerer at bare en veldig liten andel av alle mulige konfigurasjoner gir en stor forventningsskjevhet, og hvis de faktiske y_{hi} 'ene ikke er en slik konfigurasjon så er forventningsskjevheten veldig liten.

Tabell 1: Designbasert standardavvik

2. KVARTAL 1999

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	47.73	1.59	44.61	50.85
Gruppe E1	52.40	2.47	47.56	57.24
Gruppe E2	38.86	3.64	31.73	46.00
Gruppe E5	49.16	2.40	44.46	53.87

3. KVARTAL 1999

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	55.03	1.54	52.01	58.05
Gruppe E1	56.07	2.48	51.22	60.93
Gruppe E2	39.07	3.43	32.35	45.79
Gruppe E5	66.04	2.29	61.55	70.53

4. KVARTAL 1999

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	42.99	1.44	40.18	45.80
Gruppe E1	51.01	2.33	46.44	55.59
Gruppe E2	38.30	3.13	32.17	44.44
Gruppe E5	38.07	2.17	33.82	42.32

[8] At den øvre og nedre grensen er blitt like hverandre i tallverdi er ikke tilfeldig. Med notasjonen fra underavsnitt 5.1 har vi at $\sum_{i \in U_h} b_{hi} x_{hi} = 0$. Dette fører til at $\sum_{i \in A_h^c} b_{hi} x_{hi} = -\sum_{i \in A_h} b_{hi} x_{hi}$, som igjen gir $\mathrm{L} = -\mathrm{U}$.

1. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	63.40	1.52	60.43	66.37
Gruppe E1	68.71	2.37	64.07	73.35
Gruppe E2	52.55	3.53	45.63	59.48
Gruppe E5	66.20	2.18	61.93	70.47

2. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	53.50	1.45	50.65	56.34
Gruppe E1	59.94	2.49	55.06	64.81
Gruppe E2	50.28	3.02	44.37	56.20
Gruppe E5	49.35	2.16	45.11	53.60

3. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	60.36	1.53	57.36	63.35
Gruppe E1	62.41	2.47	57.56	67.26
Gruppe E2	54.71	3.23	48.38	61.05
Gruppe E5	62.67	2.39	57.98	67.35

4. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	53.59	1.43	50.78	56.40
Gruppe E1	60.26	2.39	55.58	64.95
Gruppe E2	59.46	3.14	53.31	65.61
Gruppe E5	42.13	2.07	38.07	46.19

1. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	63.43	1.45	60.59	66.27
Gruppe E1	65.72	2.34	61.12	70.32
Gruppe E2	56.82	3.27	50.42	63.23
Gruppe E5	66.09	2.06	62.05	70.12

2. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	49.70	1.43	46.89	52.51
Gruppe E1	54.73	2.17	50.48	58.98
Gruppe E2	45.31	3.17	39.09	51.52
Gruppe E5	47.85	2.29	43.36	52.35

Tabell 2: Modellbasert standardavvik og konfidensintervall

1. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	63.43	1.17	61.13	65.73
E-gruppe 1	65.72	1.91	61.98	69.46
E-gruppe 2	56.82	2.63	51.67	61.98
E-gruppe 5	66.09	1.69	62.78	69.40

2. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	49.70	1.28	47.18	52.21
E-gruppe 1	54.73	1.81	51.18	58.28
E-gruppe 2	45.31	2.79	39.84	50.77
E-gruppe 5	47.85	2.17	43.61	52.10

8. Alternativ estimator: Basert på en korrigert Horvitz-Thompson-estimator

Som vi har sett er $\hat{d}$ ikke designforventningsrett. Vi skal nå se på en alternativ estimator som for de aller fleste y_{hi} -konfigurasjonene har en mindre forventningsskjevhet enn $\hat{d}$, nemlig estimatoren

$$\tilde{d} = \frac{\sum_h \tilde{Y}_h}{\sum_h X_h} \cdot 100 ,$$

der

$$\tilde{Y}_h = \sum_{i \in s_h} \frac{y_{hi}}{\pi_{hi}} + \tilde{\beta}_h \sum_{i \notin U_h^*} x_{hi} \qquad \text{og} \qquad \tilde{\beta}_h = \frac{1}{n_h} \sum_{i \in s_h} \frac{y_{hi}}{x_{hi}} .$$

Vi kan merke oss at hvis $\pi_{hi} = n_h \left(x_{hi} / \sum_{i \in U_h^*} x_{hi} \right)$ for alle $i \in U_h^*$, så er $\tilde{Y}_h = \tilde{\beta}_h X_h$. Selv om det er ganske få strata hvor dette ikke er tilfelle (ca. 9 strata), gjør disse få strataene et stort utslag. F.eks. blir den designbaserte forventningsskjevheten til $\tilde{d}$ merkbart større hvis vi bytter ut $\tilde{Y}_h$ med $\tilde{\beta}_h X_h$ i disse strataene.

Når $U_h^* = U_h$ blir $\tilde{Y}_h = \sum_{i \in s_h} y_{hi} / \pi_{hi}$, dvs. Horvitz-Thompson-estimatoren for Y_h. Når $U_h^* \neq U_h$ kan vi se på $\tilde{Y}_h$ som en korrigert Horvitz-Thompson-estimator, der vi korrigerer med $\tilde{\beta}_h \sum_{i \notin U_h^*} x_{hi}$ for å redusere den designbaserte forventningsskjevheten.

For strata der $n_h = 1$ har vi at $\pi_{hi} = x_{hi} / \sum_{i \in U_h^*} x_{hi}$ slik at $\tilde{Y}_h = \left(y_{hi_s} / x_{hi_s} \right) X_h$. For de samme strataene har vi fra avsnitt 5 at $\hat{Y}_h = \left(y_{hi_s} / x_{hi_s} \right) X_h$. Følgelig er $\tilde{Y}_h = \hat{Y}_h$ når $n_h = 1$.

Som ved analysene av $\hat{d}$ tar vi ikke hensyn til frafall, registerfeil eller målefeil, og vi antar at vi har utvalg fra alle strata.

8.1. Designbasert analyse

I dette underavsnittet skal vi se på $\tilde{d}$ fra et designsynspunkt. Vi starter med å gi en begrunnelse for hvorfor vi velger å gjøre en korrigering av Horvitz-Thompson-estimatoren når vi skal estimerer Y_h.

Vi vet at Horvitz-Thompson-estimatoren er forventningsrett når $\pi_{hi} > 0$ for alle $i \in U_h$. Men når $U_h^* \neq U_h$ fins det $i \in U_h$ slik at $\pi_{hi} = 0$. For disse strataene er Horvitz-Thompson-estimatoren generelt ikke forventningsrett. Forventningen kan skrives som

$$\mathrm{E}\left[\sum_{i \in s_h} \frac{y_{hi}}{\pi_{hi}} \right] = \sum_{i \in U_h^*} y_{hi} ,$$

mens forventningsskjevheten er gitt ved

$$\text{Bias}\left(\sum_{i\in s_h}\frac{y_{hi}}{\pi_{hi}}\right)=-\sum_{i\notin U_h^*}y_{hi}\ .$$

Som vi ser er forventningsskjevheten alltid negativ når $U_h^* \neq U_h$ (med mindre $\sum_{i\notin U_h^*} y_{hi}=0$). Vi kan si at Horvitz-Thompson-estimatoren systematisk underestimerer Y_h i disse tilfellene. For å rette opp denne skjevheten estimerer vi $\sum_{i\notin U_h^*} y_{hi}$ med $\tilde{\beta}_h \sum_{i\notin U_h^*} x_{hi}$, og korrigerer Horvitz-Thompson-estimatoren med denne. Dette gir oss vår estimator $\tilde{Y}_h$.

Vi har altså at $\tilde{Y}_h$ er forventningsrett når $U_h^* = U_h$. Generelt for alle strata kan forventningen til $\tilde{Y}_h$ skrives som

$$\mathrm{E}\left[\tilde{Y}_h\right]=\sum_{i\in U_h^*}y_{hi}+\frac{1}{n_h}\sum_{i\in U_h^*}\frac{y_{hi}}{x_{hi}}\pi_{hi}\sum_{i\notin U_h^*}x_{hi}\ .$$

Forventningsskjevheten er dermed

$$\begin{aligned}\text{Bias}\left(\tilde{Y}_h\right)&=\frac{1}{n_h}\sum_{i\in U_h^*}\frac{y_{hi}}{x_{hi}}\pi_{hi}\sum_{i\notin U_h^*}x_{hi}-\sum_{i\notin U_h^*}y_{hi}\\&=\sum_{i\in U_h}b_{hi}y_{hi}\ ,\end{aligned}$$

der $b_{hi}=\left(\pi_{hi}/n_h x_{hi}\right)\sum_{i\notin U_h^*}x_{hi}$ når $i\in U_h^*$ og -1 ellers. (Fordi $\sum_{i\notin U_h^*}x_{hi}=\sum_{i\notin U_h^*}y_{hi}=0$ når $U_h^*=U_h$, er $\mathrm{E}\tilde{Y}_h=Y_h$ og $\text{Bias}\left(\tilde{Y}_h\right)=0$ som det skal i denne situasjonen).

Når $U_h^*\neq U_h$ er $\text{Bias}\left(\tilde{Y}_h\right)$ generelt ikke lik 0, dvs. at $\tilde{Y}_h$ generelt ikke er forventningsrett. (Men for de fleste y_{hi} -konfigurasjonene har $\tilde{Y}_h$ en mindre forventningsskjevhet enn Horvitz-Thompson-estimatoren). Dermed er $\tilde{d}$ ikke forventningsrett. Men det viser seg, ved simulering av tilfeldige y_{hi} -konfigurasjoner, at forventningsskjevheten til $\tilde{d}$ som oftest er veldig liten (vi skal se tall på dette i underavsnitt 8.3). Det viser seg også at forventningsskjevheten til $\tilde{d}$ som oftest er mindre enn forventningsskjevheten til $\hat{d}$.

Vi skal nå finne en grense for hvor stor forventningsskjevheten til $\tilde{d}$ kan bli. Fordi $\tilde{Y}_h=Y_h$ når $n_h=N_h$, kan forventningsskjevheten skrives som

$$\text{Bias}\left(\tilde{d}\right) = \frac{\sum_{h \text{ s.a. } n_h < N_h} \text{Bias}\left(\tilde{Y}_h\right)}{\sum_h X_h} \cdot 100$$

$$= \frac{\sum_{h \text{ s.a. } n_h < N_h} \sum_{i \in U_h} b_{hi} y_{hi}}{\sum_h X_h} \cdot 100 \ .$$

Ved å benytte ulikheten $-\sum_{i \notin U_h^*} x_{hi} \leq \text{Bias}\left(\tilde{Y}_h\right) \leq \sum_{i \notin U_h^*} x_{hi}$ får vi

$$\text{L} \leq \text{Bias}\left(\tilde{d}\right) \leq \text{U} \,,$$

der

$$\text{L} = \frac{-\sum_{h \text{ s.a. } n_h < N_h} \sum_{i \notin U_h^*} x_{hi}}{\sum_h X_h} \cdot 100$$

og

$$\text{U} = \frac{\sum_{h \text{ s.a. } n_h < N_h} \sum_{i \notin U_h^*} x_{hi}}{\sum_h X_h} \cdot 100 \ .$$

Dermed er L en nedre og U en øvre grense for $\text{Bias}\left(\tilde{d}\right)$. Som vi skal se i underavsnitt 8.3 er disse grensene en god del mindre enn de tilsvarende grensene for $\text{Bias}\left(\hat{d}\right)$, men de er allikevel store.

Standardavviket til $\tilde{d}$ er gitt ved

$$\text{st}\left(\tilde{d}\right) = \frac{\sqrt{\sum_h \text{V}\left(\tilde{Y}_h\right)}}{\sum_h X_h} \cdot 100 \,,$$

der $\text{V}\left(\tilde{Y}_h\right)$ er variansen til $\tilde{Y}_h$. Ved å benytte at $\tilde{Y}_h = \sum_{i \in s_h} \left(1/\pi_{hi} + \left(\sum_{i \notin U_h^*} x_{hi}\right) / n_h x_{hi}\right) y_{hi}$ får vi at denne variansen kan skrives som

$$\text{V}\left(\tilde{Y}_h\right) = \frac{1}{2} \sum_{i \in U_h^*} \sum_{\substack{j \in U_h^* \\ j \neq i}} \left(\pi_{hi}\pi_{hj} - \pi_{hij}\right)\left(v_{hi} - v_{hj}\right)^2 \ ,$$

der $v_{hi} = \left(1/\pi_{hi} + \left(\sum_{i \notin U_h^*} x_{hi}\right) / n_h x_{hi}\right) y_{hi}$.

Vi estimerer standardavviket til $\tilde{d}$ ved å estimere variansen til $\tilde{Y}_h$'ene. Når $n_h = 1$ klarer vi ikke estimere variansen $\mathrm{V}\left(\tilde{Y}_h\right)$ på vanlig måte. I disse tilfellene benytter vi derfor en øvre grense som en konservativ estimator, og vi har valgt å bruke den samme grensen som vi benyttet for variansen til $\hat{Y}_h$ i avsnitt 5, dvs. $X_h^2/4$. (Husk at $\tilde{Y}_h = \hat{Y}_h$ når $n_h = 1$).

Når $n_h > 1$ estimerer vi variansen med

(22) $$\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\tilde{Y}_h\right) = \frac{1}{2}\sum_{i \in s_h}\sum_{\substack{j \in s_h \\ j \neq i}} \frac{\pi_{hi}\pi_{hj} - \hat{\pi}_{hij}}{\hat{\pi}_{hij}}\left(v_{hi} - v_{hj}\right)^2,$$

der

$$\hat{\pi}_{hij} = \begin{cases} \pi_{hi} & \text{, når } i = j \\ \dfrac{N_h^*\left(n_h - 1\right)}{n_h\left(N_h^* - 1\right)} \cdot \pi_{hi}\pi_{hj} & \text{, når } i \neq j,\ \pi_{hi} \neq 1 \text{ og } \pi_{hj} \neq 1 \\ \pi_{hi}\pi_{hj} & \text{, når } i \neq j \text{ og } \pi_{hi} \text{ eller } \pi_{hj} = 1\,. \end{cases}$$

Estimatoren for standardavviket til $\tilde{d}$ er dermed gitt ved

(23) $$\widehat{\mathrm{st}}\left(\tilde{d}\right) = \frac{\sqrt{\sum_h \hat{\mathrm{V}}\left(\tilde{Y}_h\right)}}{\sum_h X_h} \cdot 100,$$

der

$$\hat{\mathrm{V}}\left(\tilde{Y}_h\right) = \begin{cases} 0 & \text{, når } n_h = N_h^* \\ \dfrac{1}{4}X_h^2 & \text{, når } n_h = 1 < N_h^* \\ \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\tilde{Y}_h\right) & \text{, når } 1 < n_h < N_h^* \end{cases}$$

og $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\tilde{Y}_h\right)$ er gitt ved (22).

Hvis y_{hi}'ene er slik at forventningsskjevheten til $\tilde{d}$ er tilnærmet 0, kan vi finne et tilnærmet 95% konfidensintervall for d. Fremgangsmåten er den samme som for det designbaserte konfidensintervallet i underavsnitt 5.1, og konfidensintervallet vi ender opp med er

(24) $$\tilde{d} \pm 1.96 \cdot \widehat{\mathrm{st}}\left(\tilde{d}\right).$$

Hvis vi gjør en sammenligning av estimatorene $\tilde{d}$ og $\hat{d}$ ut fra et designsynspunkt, har vi at ingen av estimatorene er forventningsrette. Men det viser seg at $\tilde{d}$ for de fleste y_{hi}-konfigurasjonene har en noe mindre forventningsskjevhet enn $\hat{d}$. Hvis standardavviket til $\tilde{d}$ alltid er mindre enn standardavviket til $\hat{d}$, vil vi derfor foretrekke estimatoren $\tilde{d}$ fremfor $\hat{d}$. Men vi har dessverre ikke klart å sammenligne standardavvikene til $\tilde{d}$ og $\hat{d}$. Hvis vi i stedet sammenligner de estimerte verdiene vi har fått for disse standardavvikene, finner vi at de er ganske like hverandre. (Talleksempel med estimater for $\mathrm{st}\left(\tilde{d}\right)$ er gitt i underavsnitt 8.3). Dette kan tyde på at $\mathrm{st}\left(\tilde{d}\right)$ og $\mathrm{st}\left(\hat{d}\right)$ er mer eller mindre like hverandre, uten at vi kan si dette sikkert. Det ser dermed ikke ut til å være noen stor gevinst ved å bytte ut den etablerte estimatoren $\hat{d}$ med $\tilde{d}$.

8.2. Modellbasert analyse

Vi skal nå gjøre en modellbasert analyse av $\tilde{d}$ under modellen ξ fra avsnitt 6. Dvs. at vi nå ser på utvalget som gitt mens vi antar følgende:

$$\mathrm{E}y_{hi} = \beta_h x_{hi} \quad , \forall \;\; i \in \text{stratum } h ,$$

$$\mathrm{V}(y_{hi}) = \sigma_h^2 x_{hi}^2 \;\; , \forall \;\; i \in \text{stratum } h$$

og uavhengighet mellom y_{hi} 'ene.

Med disse antagelsene viser det seg at

$$\mathrm{E}\left[\tilde{Y}_h - Y_h\right] = 0$$

slik at

$$\mathrm{E}\left[\tilde{d} - d\right] = \frac{\sum_h \mathrm{E}\left[\tilde{Y}_h - Y_h\right]}{\sum_h X_h} \cdot 100 = 0 \; ,$$

dvs. at $\tilde{d}$ er en forventningsrett estimator for d .[9]

Som begrunnet i avsnitt 6 måles usikkerheten til en prediktor med standardavviket til prediksjonsfeilen fremfor standardavviket til selve prediktoren. Vi måler derfor usikkerheten til $\tilde{d}$ med standardavviket til $\tilde{d} - d$. Dette standardavviket er gitt ved

[9] At $\tilde{Y}_h$ er forventningsrett kan vises på følgende måte: For strata h hvor $\pi_{hi} = n_h x_{hi} / \sum_{i \in U_h^*} x_{hi} \;\; \forall \; i \in U_h^*$, er $\tilde{Y}_h = \left(X_h / n_h\right) \sum_{i \in s_h} y_{hi} / x_{hi}$. Av dette følger det at $\mathrm{E}\left[\tilde{Y}_h - Y_h\right] = 0$. For strata h hvor $n_h x_{hi} / \sum_{i \in U_h^*} x_{hi} > 1$ for noen i , la oss si for $i \in s1_h$, er $\pi_{hi} = 1$ for $i \in s1_h$ og $\pi_{hi} = \left(n_h - |s1_h|\right) x_{hi} / \sum_{i \in U_h^* \setminus s1_h} x_{hi}$ for $i \in U_h^* \setminus s1_h$. Dermed er $\tilde{Y}_h = \sum_{i \in s1_h} y_{hi} + \left(n_h - |s1_h|\right)^{-1} \left(\sum_{i \in s_h \setminus s1_h} y_{hi} / x_{hi}\right) \sum_{i \in U_h^* \setminus s1_h} x_{hi} + n_h^{-1} \left(\sum_{i \in s_h} y_{hi} / x_{hi}\right) \sum_{i \notin U_h^*} x_{hi}$ (fordi $i \in s1_h$ alltid er med i utvalget). Av dette følger det at $\mathrm{E}\left[\tilde{Y}_h - Y_h\right] = 0$.

$$\mathrm{st}\left(\tilde{d}-d\right)=\frac{\sqrt{\sum_h \mathrm{V}\left(\tilde{Y}_h-Y_h\right)}}{\sum_h X_h}\cdot 100\,,$$

og vi estimerer det ved å estimere prediksjonsvariansene $\mathrm{V}\left(\tilde{Y}_h-Y_h\right)$.

Ved å benytte at $\tilde{Y}_h-Y_h=\sum_{i\in s_h}\left(\frac{1}{\pi_{hi}}+\frac{\sum_{i\notin U_h^*}x_{hi}}{n_h x_{hi}}-1\right)y_{hi}-\sum_{i\in r_h}y_{hi}$ finner vi at

$$\mathrm{V}\left(\tilde{Y}_h-Y_h\right)=\sigma_h^2\left\{\sum_{i\in s_h}\left(\frac{1}{\pi_{hi}}+\frac{\sum_{i\notin U_h^*}x_{hi}}{n_h x_{hi}}-1\right)^2 x_{hi}^2+\sum_{i\in r_h}x_{hi}^2\right\}.$$

Dermed kan vi estimere prediksjonsvariansen til $\tilde{Y}_h$ ved å estimere σ_h^2. Fordi vi bruker samme modell som i avsnitt 6, kan vi bruke den samme estimatoren for σ_h^2 som i avsnitt 6. Vi estimerer dermed σ_h^2 med

$$\hat{\sigma}_h^2=\frac{1}{n_h-1}\sum_{i\in s_h}\frac{1}{x_{hi}^2}\left(y_{hi}-\tilde{\beta}_h x_{hi}\right)^2$$

når $n_h>1$ og

$$\tilde{\sigma}_h^2=\frac{1}{n_g-|g|}\sum_{l\in g}\sum_{i\in s_l}\frac{1}{x_{li}^2}\left(y_{li}-\tilde{\beta}_l x_{li}\right)^2$$

når $n_h=1$. (Her er g gruppen av strata som har samme sysselsettingsintervall som stratum h, og n_g er antall utvalgte bransjeenheter i denne gruppen). Som nevnt i avsnitt 6 er $\hat{\sigma}_h^2$ en forventningsrett estimator for σ_h^2 under den antatte modellen, mens $\tilde{\sigma}_h^2$ er forventningsrett dersom σ_l^2 er lik for alle strata l som har samme sysselsettingsintervall som stratum h.

Estimatoren for standardavviket blir dermed

(25) $$\widehat{\mathrm{st}}\left(\tilde{d}-d\right)=\frac{\sqrt{\sum_h \hat{\mathrm{V}}\left(\tilde{Y}_h-Y_h\right)}}{\sum_h X_h}\cdot 100\,,$$

der

$$\hat{\mathrm{V}}\left(\tilde{Y}_h-Y_h\right)=\begin{cases}\hat{\sigma}_h^2\left\{\sum_{i\in s_h}\left(\frac{1}{\pi_{hi}}+\frac{\sum_{i\notin U_h^*}x_{hi}}{n_h x_{hi}}-1\right)^2 x_{hi}^2+\sum_{i\in r_h}x_{hi}^2\right\}, & \text{hvis } n_h>1\\ \tilde{\sigma}_h^2\left\{\sum_{i\in s_h}\left(\frac{1}{\pi_{hi}}+\frac{\sum_{i\notin U_h^*}x_{hi}}{n_h x_{hi}}-1\right)^2 x_{hi}^2+\sum_{i\in r_h}x_{hi}^2\right\}, & \text{hvis } n_h=1\,.\end{cases}$$

Et tilnærmet 95% konfidensintervall for d basert på $\tilde{d}$ er gitt ved

(26) $\tilde{d} \pm 1.96 \cdot \widehat{\mathrm{st}}\left(\tilde{d} - d\right)$.

Gjør vi en modellbasert sammenligning av $\tilde{d}$ og $\hat{d}$, har vi at begge estimatorene er forventningsrette. Spørsmålet er om standardavviket $\mathrm{st}\left(\tilde{d} - d\right)$ er større eller mindre enn standardavviket $\mathrm{st}\left(\hat{d} - d\right)$. Ved en sammenligning av $\mathrm{V}\left(\tilde{Y}_h - Y_h\right)$ og $\mathrm{V}\left(\hat{Y}_h - Y_h\right)$ viser det seg at $\mathrm{V}\left(\tilde{Y}_h - Y_h\right)$ kan bli både større og mindre enn $\mathrm{V}\left(\hat{Y}_h - Y_h\right)$, avhenging av utvalget og sysselsettingen i populasjonen. Derfor er det mulig at $\mathrm{st}\left(\tilde{d} - d\right)$ kan bli både større og mindre enn $\mathrm{st}\left(\hat{d} - d\right)$ (dette avhenger også av σ_h'ene). Av estimatene vi har fått for 1. og 2. kvartal 2002 (se Tabell 4), virker det som om $\mathrm{st}\left(\tilde{d} - d\right)$ er noe mindre enn $\mathrm{st}\left(\hat{d} - d\right)$ for disse to kvartalene.

8.3. Talleksempel

Til å beregne estimatoren $\tilde{d}$ og de tilhørende usikkerhetsmålene $\widehat{\mathrm{st}}\left(\tilde{d}\right)$ og $\widehat{\mathrm{st}}\left(\tilde{d} - d\right)$, gitt ved henholdsvis (23) og (25), har vi brukt samme data som i avsnitt 7. Dermed kan vi bare beregne $\widehat{\mathrm{st}}\left(\tilde{d} - d\right)$ for 1. og 2. kvartal 2002, fordi dette standardavviket krever at vi kjenner sysselsettingen til alle bransjeenhetene i populasjonen. Tallene vi fikk for $\widehat{\mathrm{st}}\left(\tilde{d}\right)$ og konfidensintervallet (24) er gitt i Tabell 3, mens tallene for $\widehat{\mathrm{st}}\left(\tilde{d} - d\right)$ og konfidensintervallet (26) er gitt i Tabell 4. (For lettere å kunne sammenligne $\tilde{d}$ med $\hat{d}$, har vi gitt tallene for $\hat{d}$, $\widehat{\mathrm{st}}\left(\hat{d}\right)$ og $\widehat{\mathrm{st}}\left(\hat{d} - d\right)$ i parentes etter tallene for $\tilde{d}$, $\widehat{\mathrm{st}}\left(\tilde{d}\right)$ og $\widehat{\mathrm{st}}\left(\tilde{d} - d\right)$, respektivt).

Hvis vi sammenligner estimatene vi har fått for diffusjonsindeksen når den er estimert med $\tilde{d}$ og $\hat{d}$, finner vi at disse er relativt like hverandre. Noen ganger er $\tilde{d}$ blitt større enn $\hat{d}$, andre ganger mindre. Men oftest er $\tilde{d}$ større enn $\hat{d}$ (28 av de 36 estimatene med $\tilde{d}$ er større enn det tilsvarende estimatet med $\hat{d}$). Den største forskjellen mellom estimatene har i 4. kvartal 2000, der vi for gruppen E5 har fått $\tilde{d} = 43.95$ mens $\hat{d} = 42.13$.

Tallene vi har fått for det designbaserte standardavviket $\widehat{\mathrm{st}}\left(\tilde{d}\right)$ er minst for hele populasjonen og størst for gruppen E2. For populasjonen ligger estimatene rundt 1.45, for gruppen E1 ligger de rundt 2.40, for E2 rundt 3.20 og for E5 rundt 2.20. Ut fra disse tallene kan vi nok si at usikkerheten er relativt liten ved estimering av diffusjonsindeksen til populasjonen, en del større ved estimeringen av diffusjonsindeksen til gruppene E1 og E5, og litt stor ved estimeringen av diffusjonsindeksen til gruppen E2.

Sammenligner vi estimatene for $\mathrm{st}\left(\tilde{d}\right)$ med estimatene for $\mathrm{st}\left(\hat{d}\right)$, ser vi at disse er ganske like hverandre. Noen ganger er $\widehat{\mathrm{st}}\left(\tilde{d}\right)$ større enn $\widehat{\mathrm{st}}\left(\hat{d}\right)$, andre ganger mindre. Dette tyder på at de designbaserte standardavvikene $\mathrm{st}\left(\tilde{d}\right)$ og $\mathrm{st}\left(\hat{d}\right)$ er relativt like hverandre.

I Tabell 4 ser vi at de modellbaserte standardavvikene $\widehat{\mathrm{st}}\left(\tilde{d}-d\right)$ for 1. kvartal 2002 er blitt 1.11, 1.81, 2.49 og 1.61 for henholdsvis populasjonen, gruppen E1, E2 og E5. De tilsvarende estimatene for 2. kvartal er 1.23, 1.72, 2.66 og 2.10. Dvs. at vi også her har fått de minste estimatene for populasjonen og de største for gruppen E2.

Sammenligner vi estimatene for $\mathrm{st}\left(\tilde{d}-d\right)$ med estimatene for $\mathrm{st}\left(\hat{d}-d\right)$, ser vi at $\widehat{\mathrm{st}}\left(\tilde{d}-d\right)$ er litt mindre enn $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$. Dette kan bety at $\mathrm{st}\left(\tilde{d}-d\right)$ er litt mindre enn $\mathrm{st}\left(\hat{d}-d\right)$, i hvert fall med det utvalget og den populasjonen vi har i 2002.

For å se hvor stor den designbaserte forventningsskjevheten til $\tilde{d}$ kan bli, har vi beregnet grensene L og U fra underavsnitt 8.1. Dette gav oss den nedre grensen -11.00 og den øvre grensen 11.00 for 1. og 2. kvartal 2002. Sammenlignet med grensene for $\mathrm{Bias}\left(\hat{d}\right)$, som var ± 19.23, er disse grensene en god del mindre. Men grensene er allikevel store. Heldigvis tyder en simulering av tilfeldige y_{hi} -konfigurasjoner at det bare er et fåtall av alle mulige konfigurasjoner som gir en stor forventningsskjevhet. Vi har nemlig simulert 10000 y_{hi} -konfigurasjoner, og beregnet $\mathrm{Bias}\left(\tilde{d}\right)$ for 1. kvartal 2002 med disse konfigurasjonene. Dette gav oss 10000 forventningsskjevheter som alle lå mellom -0.479 og 0.428.

Vi har også gjort en sammenligning av $\mathrm{Bias}\left(\hat{d}\right)$ og $\mathrm{Bias}\left(\tilde{d}\right)$ for tilfeldige simuleringer av y_{hi} 'ene. Dette viste at $\left|\mathrm{Bias}\left(\tilde{d}\right)\right|$ kan bli både større og mindre enn $\left|\mathrm{Bias}\left(\hat{d}\right)\right|$, men at $\left|\mathrm{Bias}\left(\tilde{d}\right)\right|$ som oftest er mindre enn $\left|\mathrm{Bias}\left(\hat{d}\right)\right|$. (Fordi vi ikke kjenner den eksakte forventningsskjevheten til $\hat{d}$, har vi benyttet tilnærmingen (14) for $\mathrm{Bias}\left(\hat{d}\right)$).

Tabell 3: Designbasert standardavvik og konfidensintervall

Tallene i parentes gjengir $\hat{d}$ og $\widehat{\mathrm{st}}\left(\hat{d}\right)$ fra Tabell 1.

2. KVARTAL 1999

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	48.56 (47.73)	1.58 (1.59)	45.47	51.66
Gruppe E1	52.23 (52.40)	2.48 (2.47)	47.37	57.09
Gruppe E2	40.36 (38.86)	3.68 (3.64)	33.15	47.58
Gruppe E5	50.54 (49.16)	2.27 (2.40)	46.09	54.99

3. KVARTAL 1999

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	55.86 (55.03)	1.51 (1.54)	52.89	58.83
Gruppe E1	57.20 (56.07)	2.51 (2.48)	52.29	62.12
Gruppe E2	40.76 (39.07)	3.24 (3.43)	34.41	47.12
Gruppe E5	65.89 (66.04)	2.28 (2.29)	61.43	70.36

4. KVARTAL 1999

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	43.93 (42.99)	1.45 (1.44)	41.10	46.77
Gruppe E1	51.68 (51.01)	2.35 (2.33)	47.07	56.29
Gruppe E2	39.00 (38.30)	3.14 (3.13)	32.84	45.15
Gruppe E5	39.50 (38.07)	2.20 (2.17)	35.18	43.82

1. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	63.44 (63.40)	1.46 (1.52)	60.58	66.29
Gruppe E1	69.08 (68.71)	2.36 (2.37)	64.45	73.71
Gruppe E2	52.78 (52.55)	3.32 (3.53)	46.28	59.28
Gruppe E5	65.77 (66.20)	2.07 (2.18)	61.71	69.83

2. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	54.26 (53.50)	1.48 (1.45)	51.35	57.16
Gruppe E1	59.73 (59.94)	2.48 (2.49)	54.88	64.58
Gruppe E2	51.21 (50.28)	3.11 (3.02)	45.11	57.31
Gruppe E5	50.99 (49.35)	2.26 (2.16)	46.56	55.42

3. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	60.60 (60.36)	1.52 (1.53)	57.62	63.59
Gruppe E1	63.12 (62.41)	2.46 (2.47)	58.31	67.93
Gruppe E2	54.63 (54.71)	3.26 (3.23)	48.23	61.02
Gruppe E5	62.71 (62.67)	2.37 (2.39)	58.06	67.36

4. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	54.52 (53.59)	1.41 (1.43)	51.76	57.29
Gruppe E1	60.73 (60.26)	2.37 (2.39)	56.09	65.37
Gruppe E2	59.88 (59.46)	3.05 (3.14)	53.91	65.86
Gruppe E5	43.95 (42.13)	2.06 (2.07)	39.90	48.00

1. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	63.74 (63.43)	1.32 (1.45)	61.15	66.32
Gruppe E1	66.01 (65.72)	2.27 (2.34)	61.57	70.46
Gruppe E2	58.23 (56.82)	2.84 (3.27)	52.66	63.79
Gruppe E5	65.53 (66.09)	1.84 (2.06)	61.92	69.14

2. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	49.63 (49.70)	1.41 (1.43)	46.87	52.38
Gruppe E1	54.04 (54.73)	2.12 (2.17)	49.89	58.18
Gruppe E2	45.75 (45.31)	3.15 (3.17)	39.57	51.93
Gruppe E5	48.04 (47.85)	2.22 (2.29)	43.69	52.38

Tabell 4: Modellbasert standardavvik og konfidensintervall

Tallene i parentes gjengir $\hat{d}$ og $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$ fra Tabell 2.

1. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	63.74 (63.43)	1.11 (1.17)	61.56	65.92
Gruppe E1	66.01 (65.72)	1.81 (1.91)	62.47	69.56
Gruppe E2	58.23 (56.82)	2.49 (2.63)	53.35	63.10
Gruppe E5	65.53 (66.09)	1.61 (1.69)	62.38	68.68

2. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	49.63 (49.70)	1.23 (1.28)	47.22	52.03
Gruppe E1	54.04 (54.73)	1.72 (1.81)	50.66	57.41
Gruppe E2	45.75 (45.31)	2.66 (2.79)	40.54	50.96
Gruppe E5	48.04 (47.85)	2.10 (2.17)	43.93	52.14

9. Alternativ estimator: Basert på beste lineære prediktor

Estimatoren vi skal se på nå er gitt ved

$$\hat{\hat{d}} = \frac{\sum_h \hat{\hat{Y}}_h}{\sum_h X_h} \cdot 100 ,$$

der

$$\hat{\hat{Y}}_h = \sum_{i \in s_h} y_{hi} + \tilde{\beta}_h \sum_{i \in r_h} x_{hi} \qquad \text{og} \qquad \tilde{\beta}_h = \frac{1}{n_h} \sum_{i \in s_h} \frac{y_{hi}}{x_{hi}} .$$

Mens estimatoren i forrige avsnitt var motivert ut fra et designsynspunkt, er denne estimatoren motivert ut fra et modellsynspunkt. Under modell ξ fra avsnitt 6 kan det nemlig vises at $\hat{\hat{Y}}_h$ er den beste lineære forventningsrette prediktor for Y_h (Bjørnstad, 1995). Dvs. at $\hat{\hat{Y}}_h$ er den prediktoren (estimatoren), blant alle lineære og forventningsrette prediktorer for Y_h, som har minst prediksjonsvarians.

Når $n_h = 1$ kan $\hat{\hat{Y}}_h$ skrives som $\hat{\hat{Y}}_h = \left(y_{hi_s} / x_{hi_s} \right) X_h$. Dermed har vi at $\hat{\hat{Y}}_h = \hat{Y}_h = \tilde{Y}_h$ når $n_h = 1$.

Som ved analysene av $\hat{d}$ og $\tilde{d}$ ser vi også her bort fra frafall, registerfeil og målefeil, og antar at vi har utvalg fra alle strata.

9.1. Designbasert analyse

Ut fra et designsynspunkt, som vi skal ha i dette underavsnittet, klarer vi ikke finne eksakt forventning og varians for $\hat{\hat{Y}}_h$ (bortsett fra når $n_h = 1$). Vi skriver derfor om $\hat{\hat{Y}}_h$ som

$$\hat{\hat{Y}}_h = \sum_{i \in s_h} \left(1 + \frac{X_h}{n_h x_{hi}} \right) \cdot y_{hi} - n_h \overline{\left(\frac{y}{x} \right)}_{s_h} \overline{x}_{s_h} ,$$

der $\overline{(y/x)}_{s_h} = (1/n_h)\sum_{i\in s_h} y_{hi}/x_{hi}$, og bruker en 1.ordens Taylortilnærming av det siste leddet. Dvs. vi gjør tilnærmingen

$$n_h\overline{\left(\frac{y}{x}\right)}_{s_h}\overline{x}_{s_h} \approx n_h\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}\mathrm{E}\overline{x}_{s_h} + n_h\mathrm{E}\overline{x}_{s_h}\left(\overline{\left(\frac{y}{x}\right)}_{s_h}\text{-E}\overline{\left(\frac{y}{x}\right)}_{s_h}\right) + n_h\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}\left(\overline{x}_{s_h}\text{-E}\overline{x}_{s_h}\right),$$

og får at

$$\hat{\hat{Y}}_h \approx \sum_{i\in s_h}\left(1+\frac{X_h}{n_h x_{hi}}\right)\cdot y_{hi} - n_h\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}\mathrm{E}\overline{x}_{s_h} - n_h\mathrm{E}\overline{x}_{s_h}\left(\overline{\left(\frac{y}{x}\right)}_{s_h}\text{-E}\overline{\left(\frac{y}{x}\right)}_{s_h}\right) - n_h\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}\left(\overline{x}_{s_h}\text{-E}\overline{x}_{s_h}\right). \tag{27}$$

Dette gir oss at

$$\begin{aligned}
\mathrm{E}\hat{\hat{Y}}_h &\approx \mathrm{E}\left[\sum_{i\in s_h}\left(1+\frac{X_h}{n_h x_{hi}}\right)\cdot y_{hi}\right] - n_h\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}\mathrm{E}\overline{x}_{s_h} \\
&= \sum_{i\in U_h^*}\left(1+\frac{X_h}{n_h x_{hi}}\right)y_{hi}\pi_{hi} - \frac{1}{n_h}\sum_{i\in U_h^*}\frac{y_{hi}}{x_{hi}}\pi_{hi}\sum_{i\in U_h^*}x_{hi}\pi_{hi} \\
&= Y_h + \sum_{i\in U_h}\left(y_{hi} - \frac{Y_h}{X_h}x_{hi}\right)\pi_{hi},
\end{aligned}$$

der siste likhet gjelder for strata hvor $\pi_{hi} = n_h\left(x_{hi}/X_h\right)$ for alle $i\in U_h$. Forventningsskjevheten kan dermed tilnærmes med

$$\begin{aligned}
\mathrm{Bias}\left(\hat{\hat{Y}}_h\right) &\approx \sum_{i\in U_h^*}\left(\pi_{hi} + \frac{X_h - \sum_{j\in U_h^*}x_{hj}\pi_{hj}}{n_h x_{hi}}\cdot\pi_{hi} - 1\right)y_{hi} - \sum_{i\notin U_h^*}y_{hi} \\
&= \sum_{i\in U_h} b_{hi}y_{hi}\ ,
\end{aligned}$$

der $b_{hi} = \pi_{hi} + \dfrac{X_h - \sum_{j\in U_h^*}x_{hj}\pi_{hj}}{n_h x_{hi}}\cdot\pi_{hi} - 1$ når $i\in U_h^*$ og -1 ellers. Denne tilnærmingen av forventningsskjevheten til $\hat{\hat{Y}}_h$ er generelt ikke lik 0.

For strata der $n_h = 1$ kan vi finne et eksakt uttrykk for $\mathrm{Bias}\left(\hat{\hat{Y}}_h\right)$. For disse strataene har vi fra avsnitt 5 at

$$\begin{aligned}
\mathrm{Bias}\left(\hat{\hat{Y}}_h\right) &= \frac{\sum_{i\in U_h^*}y_{hi}}{\sum_{i\in U_h^*}x_{hi}}\sum_{i\notin U_h^*}x_{hi} - \sum_{i\notin U_h^*}y_{hi} \\
&= \sum_{i\in U_h} c_{hi}y_{hi}\ ,
\end{aligned}$$

der $c_{hi} = \sum_{i \notin U_h^*} x_{hi} \Big/ \sum_{i \in U_h^*} x_{hi}$ når $i \in U_h^*$ og -1 ellers. (Husk at $\hat{\hat{Y}}_h = \hat{Y}_h$ når $n_h = 1$). For denne situasjonen ser vi at $\text{Bias}\left(\hat{\hat{Y}}_h\right) = 0$ når $U_h^* = U_h$, men ellers er $\hat{\hat{Y}}_h$ generelt ikke forventningsrett.

Fordi forventningsskjevheten til $\hat{\hat{d}}$ er gitt ved

$$\text{Bias}\left(\hat{\hat{d}}\right) = \frac{\sum_h \text{Bias}\left(\hat{\hat{Y}}_h\right)}{\sum_h X_h} \cdot 100 ,$$

får vi av dette at $\hat{\hat{d}}$ ikke er en forventningsrett estimator for d .

For å finne en øvre og en nedre grense for forventningsskjevheten til $\hat{\hat{d}}$ merker vi oss at

(28) $$\text{Bias}\left(\hat{\hat{d}}\right) \approx \frac{\sum\limits_{h \text{ s.a. } 1<n_h<N_h} \left(\sum\limits_{i \in U_h} b_{hi} y_{hi}\right) + \sum\limits_{h \text{ s.a. } n_h=1<N_h} \left(\sum\limits_{i \in U_h} c_{hi} y_{hi}\right)}{\sum\limits_h X_h} \cdot 100 .$$

Ved å benytte at $0 \le b_{hi} y_{hi} \le b_{hi} x_{hi}$ når $b_{hi} \ge 0$, $b_{hi} x_{hi} \le b_{hi} y_{hi} \le 0$ når $b_{hi} \le 0$, og $-\sum_{i \notin U_h^*} x_{hi} \le \sum_{i \in U_h} c_{hi} y_{hi} \le \sum_{i \notin U_h^*} x_{hi}$, får vi ulikheten

$$\text{L} \le \frac{\sum\limits_{h \text{ s.a. } 1<n_h<N_h} \left(\sum\limits_{i \in U_h} b_{hi} y_{hi}\right) + \sum\limits_{h \text{ s.a. } n_h=1<N_h} \left(\sum\limits_{i \in U_h} c_{hi} y_{hi}\right)}{\sum\limits_h X_h} \cdot 100 \le \text{U} ,$$

der

$$\text{L} = \frac{\sum\limits_{h \text{ s.a. } 1<n_h<N_h} \left(\sum\limits_{i \in A_h^c} b_{hi} x_{hi}\right) - \sum\limits_{h \text{ s.a. } n_h=1<N_h} \left(\sum\limits_{i \notin U_h^*} x_{hi}\right)}{\sum\limits_h X_h} \cdot 100 ,$$

$$\text{U} = \frac{\sum\limits_{h \text{ s.a. } 1<n_h<N_h} \left(\sum\limits_{i \in A_h} b_{hi} x_{hi}\right) + \sum\limits_{h \text{ s.a. } n_h=1<N_h} \left(\sum\limits_{i \notin U_h^*} x_{hi}\right)}{\sum\limits_h X_h} \cdot 100 ,$$

$A_h = \{i \text{ s.a. } b_{hi} \ge 0\}$ og $A_h^c = \{i \text{ s.a. } b_{hi} < 0\}$. Under forutsetingen at tilnærmingen (28) er bra nok kan vi dermed bruke L som en nedre grense og U som en øvre grense for forventningsskjevheten til $\hat{\hat{d}}$, dvs. at

$$\mathrm{L} \leq \mathrm{Bias}\left(\hat{\hat{d}}\right) \leq \mathrm{U}\,.$$

Som vi skal se i underavsnitt 9.3 så er disse grensene store. Men fra et simuleringseksperiment virker det som om forventningsskjevheten allikevel er relativt liten for de fleste y_{hi} -konfigurasjonene.

For variansen til $\hat{\hat{Y}}_h$ finner vi av (27) at

$$\begin{aligned}
\mathrm{V}\left(\hat{\hat{Y}}_h\right) &\approx \mathrm{V}\left(\sum_{i\in s_h}\left(1+\frac{X_h}{n_h x_{hi}}\right)\cdot y_{hi} - n_h\left(\mathrm{E}\overline{x}_{s_h}\right)\overline{\left(\frac{y}{x}\right)}_{s_h} - n_h\left(\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}\right)\overline{x}_{s_h}\right)\\
&= \mathrm{V}\left(\sum_{i\in s_h}\left(y_{hi}+\frac{X_h}{n_h x_{hi}}y_{hi}-\frac{y_{hi}}{x_{hi}}\mathrm{E}\overline{x}_{s_h} - x_{hi}\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}\right)\right)\\
&= \mathrm{V}\left(\sum_{i\in s_h} z_{hi}\right)\\
&= \frac{1}{2}\sum_{i\in U_h^*}\sum_{\substack{j\in U_h^*\\ j\neq i}}\left(\pi_{hi}\pi_{hj}-\pi_{hij}\right)\left(z_{hi}-z_{hj}\right)^2,
\end{aligned}$$

der $z_{hi} = y_{hi} + \frac{X_h}{n_h x_{hi}} y_{hi} - \frac{y_{hi}}{x_{hi}}\mathrm{E}\overline{x}_{s_h} - x_{hi}\mathrm{E}\overline{\left(\frac{y}{x}\right)}_{s_h}$. Dermed kan vi estimerer variansen til $\hat{\hat{Y}}_h$ med

$$\text{(29)}\qquad \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\hat{\hat{Y}}_h\right) = \frac{1}{2}\sum_{i\in s_h}\sum_{\substack{j\in s_h\\ j\neq i}}\frac{\pi_{hi}\pi_{hj}-\hat{\pi}_{hij}}{\hat{\pi}_{hij}}\left(\hat{z}_{hi}-\hat{z}_{hj}\right)^2,$$

der

$$\hat{\pi}_{hij} = \begin{cases} \pi_{hi} & \text{, når } i=j \\ \dfrac{N_h^*\left(n_h-1\right)}{n_h\left(N_h^*-1\right)}\cdot\pi_{hi}\pi_{hj} & \text{, når } i\neq j,\ \pi_{hi}\neq 1 \text{ og } \pi_{hj}\neq 1 \\ \pi_{hi}\pi_{hj} & \text{, når } i\neq j \text{ og } \pi_{hi} \text{ eller } \pi_{hj}=1 \end{cases}$$

og $\hat{z}_{hi} = y_{hi} + \frac{X_h}{n_h x_{hi}} y_{hi} - \frac{y_{hi}}{x_{hi}}\overline{x}_{s_h} - x_{hi}\overline{\left(\frac{y}{x}\right)}_{s_h}$. Når $n_h = 1$ gjør vi som for $\hat{Y}_h$ og estimerer variansen med den øvre grensen $X_h^2/4$.

Vi estimere nå standardavviket til $\hat{\hat{d}}$, som er gitt ved

$$\mathrm{st}\left(\hat{\hat{d}}\right) = \frac{\sqrt{\sum_h \mathrm{V}\left(\hat{\hat{Y}}_h\right)}}{\sum_h X_h}\cdot 100\,,$$

med

(30) $$\widehat{\mathrm{st}}\left(\hat{\hat{d}}\right)=\frac{\sqrt{\sum_h \hat{\mathrm{V}}\left(\hat{\hat{Y}}_h\right)}}{\sum_h X_h}\cdot 100\ .$$

Her er

$$\hat{\mathrm{V}}\left(\hat{\hat{Y}}_h\right)=\begin{cases} 0 & \text{, når } n_h=N_h \\ \frac{1}{4}X_h^2 & \text{, når } n_h=1<N_h \\ \hat{\mathrm{V}}_{\mathrm{SYG}}\left(\hat{\hat{Y}}_h\right) & \text{, når } 1<n_h<N_h \end{cases}$$

og $\hat{\mathrm{V}}_{\mathrm{SYG}}\left(\hat{\hat{Y}}_h\right)$ er gitt ved (29).

Et tilnærmet 95% konfidensintervall for $\hat{\hat{d}}$ er gitt ved

(31) $$\hat{\hat{d}}\pm 1.96\cdot\widehat{\mathrm{st}}\left(\hat{\hat{d}}\right),$$

forutsatt at y_{hi} 'ene er slik at forventningsskjevheten er tilnærmet 0.

Sammenligner vi estimatoren $\hat{\hat{d}}$ med $\hat{d}$, har vi at ingen av estimatorene er forventingsrette. For å avgjøre om den ene estimatoren er å foretrekke fremfor den andre, må vi derfor sammenligne både standardavvikene og forventningsskjevhetene. Dessverre har vi ikke klart å sammenligne standardavvikene, men av estimatene vi har fått for disse kan det virke som om det ikke er så stor forskjell. For forventningsskjevhetene viser det seg at $\mathrm{Bias}\left(\hat{\hat{d}}\right)$ kan bli både større og mindre enn $\mathrm{Bias}\left(\hat{d}\right)$, avhengig av y_{hi} 'ene (men det virker som om $\mathrm{Bias}\left(\hat{\hat{d}}\right)$ er litt mindre enn $\mathrm{Bias}\left(\hat{d}\right)$ for de fleste konfigurasjonene). Vi kan derfor ikke si, ut fra et designsynspunkt, at den ene estimatoren alltid er bedre enn den andre.

9.2. Modellbasert analyse

I dette underavsnittet skal vi igjen anta at konjunktursysselsettingene følger modell ξ. Dvs. vi antar at

$$\mathrm{E}y_{hi}=\beta_h x_{hi}\quad ,\ \forall\ \ i\in \text{stratum } h\,,$$

$$\mathrm{V}(y_{hi})=\sigma_h^2 x_{hi}^2\ ,\ \forall\ \ i\in \text{stratum } h$$

og at y_{hi} 'ene er uavhengige av hverandre, mens vi ser på utvalget som gitt.

Under disse antagelsene finner vi at $\mathrm{E}\left[\hat{\hat{Y}}_h - Y_h\right] = 0$ slik at

$$\mathrm{E}\left[\hat{\hat{d}} - d\right] = \frac{\sum_h \mathrm{E}\left[\hat{\hat{Y}}_h - Y_h\right]}{\sum_h X_h} \cdot 100 = 0 \ .$$

Dvs. at $\hat{\hat{d}}$ er forventningsrett for d .

Ved å benytte at $\hat{\hat{Y}}_h - Y_h = (1/n_h) \sum_{i \in s_h} (y_{hi} / x_{hi}) \sum_{i \in r_h} x_{hi} - \sum_{i \in r_h} y_{hi}$ finner vi at prediksjonsvariansen til $\hat{\hat{Y}}_h$ kan skrives som

$$\mathrm{V}\left(\hat{\hat{Y}}_h - Y_h\right) = \sigma_h^2 \left\{ \frac{1}{n_h} \left(\sum_{i \in r_h} x_{hi} \right)^2 + \sum_{i \in r_h} x_{hi}^2 \right\} .$$

Av dette uttrykket ser vi at prediksjonsvariansen er minst når det er bransjeenhetene med størst sysselsetting som er med i utvalget.

Det kan vises at prediksjonsvariansen til $\hat{\hat{Y}}_h$ alltid er mindre eller lik prediksjonsvariansen til alle mulige lineære og forventningsrette prediktorer for Y_h. Fordi $\hat{Y}_h$ og $\tilde{Y}_h$ er lineære og forventningsrette innebærer dette at prediksjonsvariansen til $\hat{\hat{Y}}_h$ er mindre eller lik prediksjonsvariansen til $\hat{Y}_h$ og $\tilde{Y}_h$. Dermed får vi at standardavviket til $\hat{\hat{d}} - d$, som er gitt ved

$$\text{(32)} \qquad \mathrm{st}\left(\hat{\hat{d}} - d\right) = \frac{\sqrt{\sum_h \mathrm{V}\left(\hat{\hat{Y}}_h - Y_h\right)}}{\sum_h X_h} \cdot 100 \, ,$$

er mindre eller lik de tilsvarende standardavvikene til $\hat{d}$ og $\tilde{d}$.

Vi estimerer standardavviket (32) ved å estimere prediksjonsvariansen til $\hat{\hat{Y}}_h$, og vi estimerer prediksjonsvariansen til $\hat{\hat{Y}}_h$ ved å estimere σ_h^2. Som estimator for σ_h^2 bruker vi samme estimator som i avsnitt 6 og underavsnitt 8.2, dvs.

$$\hat{\sigma}_h^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} \frac{1}{x_{hi}^2} \left(y_{hi} - \tilde{\beta}_h x_{hi} \right)^2$$

når $n_h > 1$ og

$$\tilde{\sigma}_h^2 = \frac{1}{n_g - |g|} \sum_{l \in g} \sum_{i \in s_l} \frac{1}{x_{li}^2} \left(y_{li} - \tilde{\beta}_l x_{li} \right)^2$$

når $n_h = 1$. (Her er g gruppen av strata som har samme sysselsettingsintervall som stratum h, og n_g er antall utvalgte bransjeenheter i denne gruppen). Estimatoren for standardavviket (32) er dermed

$$(33)\qquad \widehat{\mathrm{st}}\left(\hat{\hat{d}} - d\right) = \frac{\sqrt{\sum_h \widehat{\mathrm{V}}\left(\hat{\hat{Y}}_h - Y_h\right)}}{\sum_h X_h} \cdot 100,$$

der

$$\widehat{\mathrm{V}}\left(\hat{\hat{Y}}_h - Y_h\right) = \begin{cases} \hat{\sigma}_h^2 \left\{ \frac{1}{n_h}\left(\sum_{i\in r_h} x_{hi}\right)^2 + \sum_{i\in r_h} x_{hi}^2 \right\}, & \text{hvis } n_h > 1 \\ \tilde{\sigma}_h^2 \left\{ \frac{1}{n_h}\left(\sum_{i\in r_h} x_{hi}\right)^2 + \sum_{i\in r_h} x_{hi}^2 \right\}, & \text{hvis } n_h = 1 \end{cases}$$

er estimator for $\mathrm{V}\left(\hat{\hat{Y}}_h - Y_h\right)$.

Vi kan bemerke at fordi prediksjonsvariansen til $\hat{\hat{Y}}_h$ er mindre eller lik prediksjonsvariansen til $\hat{Y}_h$ og $\tilde{Y}_h$ for alle mulige verdier av $\sigma_h^2 > 0$, og fordi σ_h^2 estimeres med samme estimator i alle tre prediksjonsvariansene, så er $\widehat{\mathrm{V}}\left(\hat{\hat{Y}}_h - Y_h\right)$ mindre eller lik $\widehat{\mathrm{V}}\left(\hat{Y}_h - Y_h\right)$ og $\widehat{\mathrm{V}}\left(\tilde{Y}_h - Y_h\right)$. Dette medfører igjen at $\widehat{\mathrm{st}}\left(\hat{\hat{d}} - d\right)$ er mindre eller lik $\widehat{\mathrm{st}}\left(\hat{d} - d\right)$ og $\widehat{\mathrm{st}}\left(\tilde{d} - d\right)$.

Et (tilnærmet) 95% konfidensintervall for d basert på $\hat{\hat{d}}$ er gitt ved

$$(34)\qquad \hat{\hat{d}} \pm 1.96 \cdot \widehat{\mathrm{st}}\left(\hat{\hat{d}} - d\right).$$

Fordi $\widehat{\mathrm{st}}\left(\hat{\hat{d}} - d\right)$ er mindre eller lik $\widehat{\mathrm{st}}\left(\hat{d} - d\right)$ og $\widehat{\mathrm{st}}\left(\tilde{d} - d\right)$, er dette konfidensintervallet smalere enn de tilsvarende konfidensintervallene basert på $\hat{d}$ og $\tilde{d}$ (dvs. smalere enn konfidensintervallene (20) og (26)).

Vi har sett at under modell ξ så er $\hat{d}$, $\tilde{d}$ og $\hat{\hat{d}}$ alle forventningsrette estimatorer. Men fordi $\mathrm{st}\left(\hat{\hat{d}} - d\right)$ er mindre eller lik $\mathrm{st}\left(\hat{d} - d\right)$ og $\mathrm{st}\left(\tilde{d} - d\right)$, kan vi si at $\hat{\hat{d}}$ er en bedre estimator enn $\hat{d}$ og $\tilde{d}$. Men om $\hat{\hat{d}}$ er så mye bedre att det er verdt å bytte ut $\hat{d}$ med denne, avhenger av hvor mye mindre standardavviket er. (Som vi skal se i neste underavsnitt er estimatene for $\mathrm{st}\left(\hat{\hat{d}} - d\right)$ bare litt mindre enn estimatene for $\mathrm{st}\left(\hat{d} - d\right)$).

9.3. Talleksempel

Med de samme dataene som vi beregnet usikkerhetsmålene til estimatorene $\hat{d}$ og $\tilde{d}$, har vi beregnet usikkerhetsmålene for estimatoren $\hat{\hat{d}}$. I Tabell 5 ser vi tallene vi fikk for det designbaserte standardavviket $\widehat{\mathrm{st}}\left(\hat{\hat{d}}\right)$, gitt ved (30), og konfidensintervallet (31). I Tabell 6 ser vi tallene vi fikk for det modellbaserte standardavviket $\widehat{\mathrm{st}}\left(\hat{\hat{d}}-d\right)$, gitt ved (33), og konfidensintervallet (34).

Estimatene vi har fått for $\hat{\hat{d}}$ avviker ikke mye fra $\hat{d}$. Som $\tilde{d}$ er $\hat{\hat{d}}$ oftest blitt større enn $\hat{d}$ (26 av de 36 estimatene med $\hat{\hat{d}}$ er større enn de tilsvarende estimatene med $\hat{d}$). Den største forskjellen har vi i 1. kvartal 2002, der vi for gruppen E2 har fått $\hat{\hat{d}} = 58,48$ mens $\hat{d} = 56,82$. Sammenligner vi også med estimatene vi fikk for $\tilde{d}$, finner vi at $\hat{\hat{d}}$ ligger mellom $\hat{d}$ og $\tilde{d}$ i 23 tilfeller (og som oftest har vi sammenhengen $\hat{d} \le \hat{\hat{d}} \le \tilde{d}$).

Av Tabell 5 finner vi at estimatene for det designbaserte standardavviket $\mathrm{st}\left(\hat{\hat{d}}\right)$ ligger mellom 1.32 og 1.58 for populasjonen. For gruppe E1 ligger estimatene rundt 2.40, for gruppe E2 rundt 3.20 og for gruppe E5 rundt 2.20. Usikkerheten ser med andre ord ut til å være relativt liten for populasjonen, noe større for gruppene E1 og E5 og litt stor for gruppen E2.

Sammenligner vi estimatene for $\mathrm{st}\left(\hat{\hat{d}}\right)$ med estimatene for $\mathrm{st}\left(\hat{d}\right)$, ser vi at disse er ganske like hverandre. Noen ganger er $\widehat{\mathrm{st}}\left(\hat{\hat{d}}\right)$ større enn $\widehat{\mathrm{st}}\left(\hat{d}\right)$, andre ganger mindre, og ved ett tilfelle er estimatene blitt (tilnærmet) like. Dette tyder på at $\mathrm{st}\left(\hat{\hat{d}}\right)$ og $\mathrm{st}\left(\hat{d}\right)$ er relativt like hverandre. (Hvis vi sammenligner $\widehat{\mathrm{st}}\left(\hat{\hat{d}}\right)$ og $\widehat{\mathrm{st}}\left(\tilde{d}\right)$ finner vi ved hele 24 tilfeller at estimatene er (tilnærmet) like).

For det modellbaserte standardavviket $\mathrm{st}\left(\hat{\hat{d}}-d\right)$ har vi for 1. kvartal 2002 fått estimatene 1.11, 1.80, 2.47 og 1.59 for henholdsvis populasjonen, gruppen E1, E2 og E5. For 2. kvartal 2002 er de tilsvarende estimatene blitt 1.22, 1.72, 2.63 og 2.07. Dette tyder på at usikkerheten ved estimering av diffusjonsindeksen er liten for populasjonen og gruppen E1, og noe større for gruppene E2 og E5.

Vi vet at standardavviket $\mathrm{st}\left(\hat{\hat{d}}-d\right)$ skal være mindre eller lik de tilsvarene standardavvikene for $\hat{d}$ og $\tilde{d}$. Vi vet også at det samme gjelder for estimatorene til disse standardavvikene, dvs. at $\widehat{\mathrm{st}}\left(\hat{\hat{d}}-d\right)$ skal være mindre eller lik $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$ og $\widehat{\mathrm{st}}\left(\tilde{d}-d\right)$. Av Tabell 4 og 6 ser vi at $\widehat{\mathrm{st}}\left(\hat{\hat{d}}-d\right)$ er blitt nesten lik $\widehat{\mathrm{st}}\left(\tilde{d}-d\right)$, og bare er litt mindre enn $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$. Dette kan bety at $\mathrm{st}\left(\hat{\hat{d}}-d\right)$ er nesten lik $\mathrm{st}\left(\tilde{d}-d\right)$, og bare litt mindre enn $\mathrm{st}\left(\hat{d}-d\right)$.

For den designbaserte fornventningsskjevheten til $\hat{\hat{d}}$ finner vi at den nedre grenser er -12.59 og den øvre grensen er 12.59 (for 1. og 2. kvartal 2002). De tilsvarende grensene for $\hat{d}$ og $\tilde{d}$ er henholdsvis ± 19.23 og ± 11.00. Dvs. at grensene til $\text{Bias}\left(\hat{\hat{d}}\right)$ er en del mindre enn grensene til $\text{Bias}\left(\hat{d}\right)$, og litt større enn grensene til $\text{Bias}\left(\tilde{d}\right)$.

For å få flere tall for den designbaserte fornventningsskjevheten til $\hat{\hat{d}}$, har vi simulert 10000 tilfeldige y_{hi} -konfigurasjoner. For hver av disse har vi beregnet $\text{Bias}\left(\hat{\hat{d}}\right)$ for 1. kvartal 2002. (Vi har benyttet tilnærmingen (28) for forventningsskjevheten). Resultatet ble at alle forventningsskjevhetene lå mellom -0.708 og 0.774, og at hele 9903 av forventningsskjevhetene lå mellom -0.5 og 0.5. Dette tyder på at forventningsskjevheten til $\hat{\hat{d}}$ er ganske liten for de fleste av alle mulige y_{hi} -konfigurasjoner.

Vi har også gjort en sammenligning av $\text{Bias}\left(\hat{d}\right)$ og $\text{Bias}\left(\hat{\hat{d}}\right)$ for tilfeldige simuleringer av y_{hi} 'ene. Dette gav samme resultat som ved sammenligningen av $\text{Bias}\left(\hat{d}\right)$ og $\text{Bias}\left(\tilde{d}\right)$. Dvs. at $\left|\text{Bias}\left(\hat{\hat{d}}\right)\right|$ kan bli både større og mindre enn $\left|\text{Bias}\left(\hat{d}\right)\right|$, men at $\left|\text{Bias}\left(\hat{\hat{d}}\right)\right|$ som oftest er mindre enn $\left|\text{Bias}\left(\hat{d}\right)\right|$.

Tabell 5: Designbasert standardavvik

Tallene i parentes gjengir $\hat{d}$ og $\widehat{\text{st}}\left(\hat{d}\right)$ fra Tabell 1.

2. KVARTAL 1999

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	48.30 (47.73)	1.58 (1.59)	45.20	51.41
Gruppe E1	52.04 (52.40)	2.48 (2.47)	47.18	56.89
Gruppe E2	40.31 (38.86)	3.68 (3.64)	33.10	47.51
Gruppe E5	50.06 (49.16)	2.31 (2.40)	45.53	54.58

3. KVARTAL 1999

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	55.48 (55.03)	1.51 (1.54)	52.51	58.44
Gruppe E1	57.13 (56.07)	2.51 (2.48)	52.21	62.04
Gruppe E2	39.87 (39.07)	3.24 (3.43)	33.52	46.22
Gruppe E5	65.58 (66.04)	2.27 (2.29)	61.13	70.03

4. KVARTAL 1999

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	43.72 (42.99)	1.44 (1.44)	40.90	46.53
Gruppe E1	51.62 (51.01)	2.35 (2.33)	47.01	56.22
Gruppe E2	38.49 (38.30)	3.09 (3.13)	32.43	44.55
Gruppe E5	39.34 (38.07)	2.19 (2.17)	35.04	43.64

1. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	63.58 (63.40)	1.46 (1.52)	60.72	66.43
Gruppe E1	69.05 (68.71)	2.36 (2.37)	64.42	73.68
Gruppe E2	53.01 (52.55)	3.30 (3.53)	46.53	59.49
Gruppe E5	66.01 (66.20)	2.08 (2.18)	61.94	70.09

2. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	54.13 (53.50)	1.48 (1.45)	51.23	57.03
Gruppe E1	59.72 (59.94)	2.48 (2.49)	54.86	64.57
Gruppe E2	50.92 (50.28)	3.10 (3.02)	44.84	57.01
Gruppe E5	50.85 (49.35)	2.26 (2.16)	46.43	55.27

3. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	60.42 (60.36)	1.52 (1.53)	57.45	63.40
Gruppe E1	63.03 (62.41)	2.45 (2.47)	58.22	67.83
Gruppe E2	54.50 (54.71)	3.24 (3.23)	48.15	60.85
Gruppe E5	62.40 (62.67)	2.37 (2.39)	57.75	67.04

4. KVARTAL 2000

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	54.36 (53.59)	1.41 (1.43)	51.60	57.13
Gruppe E1	60.70 (60.26)	2.37 (2.39)	56.06	65.34
Gruppe E2	60.04 (59.46)	3.06 (3.14)	54.05	66.02
Gruppe E5	43.41 (42.13)	2.06 (2.07)	39.37	47.45

1. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	63.90 (63.43)	1.32 (1.45)	61.31	66.49
Gruppe E1	66.04 (65.72)	2.27 (2.34)	61.59	70.49
Gruppe E2	58.48 (56.82)	2.84 (3.27)	52.90	64.06
Gruppe E5	65.78 (66.09)	1.85 (2.06)	62.16	69.40

2. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	49.69 (49.70)	1.41 (1.43)	46.93	52.45
Gruppe E1	54.22 (54.73)	2.12 (2.17)	50.06	58.37
Gruppe E2	46.12 (45.31)	3.15 (3.17)	39.94	52.30
Gruppe E5	47.72 (47.85)	2.22 (2.29)	43.37	52.07

Tabell 6: Modellbasert standardavvik og konfidensintervall

Tallene i parentes gjengir $\hat{d}$ og $\widehat{\mathrm{st}}\left(\hat{d}-d\right)$ fra Tabell 2.

1. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	63.90 (63.43)	1.11 (1.17)	61.74	66.07
Gruppe E1	66.04 (65.72)	1.80 (1.91)	62.51	69.57
Gruppe E2	58.48 (56.82)	2.47 (2.63)	53.64	63.32
Gruppe E5	65.78 (66.09)	1.59 (1.69)	62.65	68.90

2. KVARTAL 2002

Aggregerings-nivå	Estimert diffusjonsindeks	Estimert standardavvik	Nedre grense i 95% konfidensintervall	Øvre grense i 95% konfidensintervall
Populasjon	49.69 (49.70)	1.22 (1.28)	47.30	52.07
Gruppe E1	54.22 (54.73)	1.72 (1.81)	50.85	57.58
Gruppe E2	46.12 (45.31)	2.63 (2.79)	40.96	51.28
Gruppe E5	47.72 (47.85)	2.07 (2.17)	43.67	51.77

10. Oppsummering

I dette notatet har vi sett hvordan vi kan måle usikkerheten i konjunkturbarometeret ved først å gjøre en designbasert analyse av $\hat{d}$ og så en modellbasert analyse. Dette har gitt oss to standardavvik som kan brukes som usikkerhetsmål. Det designbaserte standardavviket måler usikkerheten som kommer av at det er flere mulige utvalg som kan trekkes. Det modellbaserte standardavviket måler usikkerheten som kommer av at konjunktursysselsettingene antas å være stokastiske variabler som kan anta flere verdier.

Hvis vi måler usikkerheten med det designbaserte standardavviket, er usikkerheten relativt liten når vi estimerer diffusjonsindeksen til hele populasjonen (estimatene for standardavviket varierer fra 1.43 til 1.59). Ved estimering av diffusjonsindeksen til gruppene E1 og E5 er det noe større usikkerhet, og for gruppen E2 ser det ut til å være en del usikkerhet (de estimerte standardavvikene vi fikk for gruppen E2 ligger mellom 3.02 og 3.64). For både populasjonen og gruppene ser det ut til at usikkerheten ikke varierer så mye fra kvartal til kvartal.

Med det modellbaserte usikkerhetsmålet ser det ut til at det er relativt liten usikkerhet når vi estimerer diffusjonsindeksen til både populasjonen og gruppen E1 (for populasjonen fikk vi estimatene 1.17 og 1.28). Usikkerheten virker litt større når vi estimerer diffusjonsindeksen til gruppen E5. Den største usikkerheten har vi for gruppen E2, der de estimerte standardavvikene er blitt 2.63 og 2.79.

Hvilket av disse to usikkerhetsmålene som skal velges avhenger av hvilken usikkerhet som vil angis. Innen utvalgsundersøkelser brukes gjerne den designbaserte usikkerheten, men vi mener nok at den modellbaserte usikkerheten er bedre for konjunkturbarometeret. En begrunnelse for dette er at når vi har det samme utvalget over flere kvartal, og for hvert kvartal observerer nye verdier til de utvalgte bransjeenhetene, er det nærliggende å tenke på utvalget som konstant og i stedet se på konjunktursysselsettingene som stokastiske variabler.

I tillegg til å måle usikkerheten til $\hat{d}$, har vi sett på de to alternative estimatorene $\tilde{d}$ og $\hat{\hat{d}}$ for diffusjonsindeksen. Hensikten med dette var å se om vi klarer å estimere diffusjonsindeksen mer presist med en av disse estimatorene. Hvorvidt det er tilfelle, avhenger om vi ser det fra et design- eller modellsynspunkt.

Om vi velger et designsynspunkt, kan vi ikke si at en av estimatorene alltid er bedre enn de andre to. Det virker som om en estimator kan være bedre for noen y_{hi} -konfigurasjoner, men ikke for alle. (Iallfall når vi ser på forventningsskjevheten).

Mener vi derimot at det er mer riktig å se på estimatorene fra et modellsynspunkt, er $\hat{\hat{d}}$ den beste estimatoren fordi denne har minst standardavvik. (Alle estimatorene er modellforventningsrette). Men ut fra estimatene vi har fått for standardavvikene, ser det ikke ut som om $\hat{\hat{d}}$ er så veldig mye bedre. Derfor er det antakelig ikke verdt å bytte ut estimatoren som benyttes i dag med $\hat{\hat{d}}$.

Referanser

Bjørnstad, J.F. (1995): *Utvalgsundersøkelser og prediksjon*.

Chambers, R.: Kursmaterialet til kurset ST640 (Survey Sampling and Estimation II) fra MSc in Official Statistics.

Rao, J.N.K. og Wu, C.F.J. (1988): *Resampling Inference With Complex Survey Data.* JASA vol. 83 no. 401, 231-241.

Sitter, R.R. (1992): *A Resampling Procedure for Complex Survey Data.* JASA, vol. 87, no. 419, 755-765.

De sist utgitte publikasjonene i serien Notater

2003/1 G. Dahl: Arbeidsmarkedstiltak blant sosialhjelpsmottakere. 25s.

2003/2 C. Nordseth og T. Sandnes: FD - Trygd. Dokumentasjonsrapport . Pensjonsgivende inntekt, 1992-2000. Omsorgspoeng, 1992-1998. 25s.

2003/3 B. Otnes: Tidsbruk blant uførepensjonister med barn. 56s.

2003/4 L.H. Thingstad: Endringer i lov om merverdiavgift i 2001. Konsekvenser for terminvise og kvartalsvise omsetningsstatistikker. 81s.

2003/5 Y. Bergstrøm, J.H. Wang, S. Bakke og G. Haraldsen: Dokumentasjon og veiledning for implementering av Web-skjema i SSBs Web-portal. Utvikling av et rapporteringssystem via Internett for kvartalsvis investeringsstatistikk og detaljomsetningsindeksen innenfor rammen av IDUN-prosjektet. 69s.

2003/6 T. Dahle, A.K. Johnsen og D. Roll-Hansen: Utvikling av informasjonsmateriell til undersøkelsen som livserfaring og leseforståelse (Adult Literacy and Life Skills Survey - ALL) ved hjelp av fokusgrupper. 39s.

2003/7 H.C. Hougen og G.E. Wangen: WHOs Vekststudie av sped- og småbarn. Dokumentasjonrapport. 12s.

2003/8 T. Smith: Vann- og avløpsgebyrer- en gjennomgang av kommunenes praksis. 65s.

2003/9 T.M. Normann: Omnibusundersøkelsen november/desember 2002. Dokumentasjonsrapport. 51s.

2003/10 E.Engelien og M. Steinnes: Tilgang til friluftsområder - metode og resultater 2002. 59s.

2003/11 Y. Dyvi: Virkningsberegninger på MODAG. 66s.

2003/12 A.K Johnsen og T.M. Normann: Evaluering av informasjonstiltak og Internettilbud i Folke- og boligtellingen 2001. Dokumentasjonsrapport. 30s.

2003/16 I. Kvalstad: SEDA - Sentrale data fra allmennlegetjenesten. Teknisk dokumentasjon. 136s.

2003/17 K.I. Bøe og T. Sandnes: FD - Trygd. Dokumentasjonsrapport. Statsansatte. 1992-2000. 28s.

2003/18 C. Nordseth og T. Sandnes: FD - Trygd. Dokumentasjonsrapport. Inntekt og formue, 1992-2000. 42s.

2003/19 A. Rolland (red.): Borger- og bruker-undersøkelser i en modernisert offentlig sektor. 112s.

2003/20 A-K.Brændvang, E. Evensen, P. Løkkevik og H. Sande Olsen: Næringene hotell, restaurant og samferdsel. Dokumentasjon av beregningene i nasjonalregnskapet. 53s.

2003/21 I. Håland, T. Köber og S.Lyby: Kvalitetssikring av driftsrutinene AKU. 14s.

2003/22 H. Hartvedt og E. Frisvoll: Kobling av adresseregistrene i DSF og GAB 2002. Dokumentasjon av samsvar og avik. 34s.

2003/23 A. Akselsen og T. Sandnes: FD - Trygd. Dokumentasjonsrapport. Stønader til enslig forsørger. 1992-2001. 46s.

2003/24 C. Nordseth og T. Sandnes: FD - Trygd. Dokumentasjonsrapport. Foreløpig uførestønad. 1992-2001. 39s.

2003/25 S. Derakhshanfar og T. Sandnes: FD - Trygd. Dokumentasjonsrapport. Økonomisk sosialhjelp. 1992-2001. 35s.

2003/26 A. Akselsen, S. Lien og T. Sandnes: FD - Trygd. Dokumentasjonsrapport. Pensjoner. Grunn og hjelpestønader. 1992-2001. 113s.

2003/27 E. Eng Eikebak og R. Johannessen: Forventningsindikator - konsumprisene. November-mai 2003. 17s.