Arbeidsnotater

STATISTISK SENTRALBYRÅ

IO 73/17 16. mai 1973

FRAFALLETS BETYDNING I UTVALGS-UNDERSØKELSER VURDERT VED EN PROBABILISTISK MODELL

Av

Steinar Tamsfoss[1)]

INNHOLD

1) Jeg vil takke cand.real. O.J. Skaugen for de verdifulle forslag han har kommet med under skrivingen av dette notatet.

1. INNLEDNING

"Frafall" i utvalgsundersøkelser forekommer når en ikke oppnår observasjoner for en utvalgt enhet. Vi skal bruke betegnelsen "frafall" både om selve fenomenet, og om samlingen av enheter som faller fra.

I dette notatet vil frafall bli betraktet som tilfeldig forekommende hendelser som kan beskrives meningsfylt i sannsynlighetsteoretiske (frekventistiske) termer og dermed inkorporeres i bl.a. estimeringsprosedyren. Cochran, [1], behandler frafallsproblemet ved å inndele populasjonen i to strata; ett hvorfra observasjoner oppnås og ett der dette ikke er tilfelle. Imidlertid er denne betraktningsmåten nokså forenklet, noe Cochran understreker: "In a more complete specification of the problem, we would attach to each unit a probability representing the chance that it would be measured by a given field method if it fell in the sample" (s. 293). En slik spesifisering av modellen vil bli gjort i neste avsnitt.

G. Elofsson, [3], har gjort et studium av tre ulike korreksjonsmetoder for frafall. Den modellen hun legger til grunn for frafallets opptreden, er på visse punkter nokså lik modellen i dette notatet. Korreksjonsmetodene som blir sammenlignet innbyrdes i [3], er såkalt "veiing", "middelverdiimputering" og "dubblering". Elofsson viser bl.a. at den første metoden er minst like god som de øvrige, samt at alle metodene (naturligvis) gir dårligere resultater enn hva en ville fått dersom frafall ikke hadde forekommet.

I de fleste praktiske situasjoner er problemstillingen følgende: Er frafallet av en slik karakter at det er påkrevd å foreta korrigeringer av estimatorene? Mer konkret er dette et spørsmål om ikke-korrigerte estimatorer har større varians (evt. bruttovarians) enn korrigerte prosedyrer. Formålet med dette notatet er i hovedsak å belyse denne problemstillingen. De resultater vi kommer fram til, er selvsagt avhengige av både frafallsmodellen som legges til grunn, og korreksjonsmetoden vi velger. Når det gjelder det valg av modell som her er gjort, er begrunnelsen ikke noen annen enn at den synes å beskrive frafalls-fenomenet noenlunde adekvat. Den korrigerte estimatoren som blir brukt, er en veieprosedyre som virker rimelig i relasjon til modellen.

Situasjonen som behandles her, er en utvalgsplan med enkel tilfeldig utvelging der oppgaven er å estimere det relative antall enheter som har et bestemt kjennetegn i en populasjon ("prosent-fordelinger"). Populasjonen antas å være så stor sammenlignet med utvalgsstørrelsen, at vi kan se bort fra effekten av trekking uten tilbakelegging. Dette

innebærer at vi kan anvende binomiske "sannsynligheter" framfor hypergeometriske som er matematisk mindre komfortable.

Sett i relasjon til den faste utvalgsplanen Byrået bruker ved intervjuundersøkelser, er dette studiet på enkelte punkter begrenset. Det antas likevel at den enkle situasjonen som drøftes her, kan utvides til å dekke "intervjukontorets" utvalgsplan. Dette vil bli diskutert i slutt-ordet.

I litteraturlisten er det tatt med noen skrifter som på ulike måter behandler frafallsproblemet i skandinaviske intervjuundersøkelser ([2], [6], [7]).

2. FRAFALLSHOMOGENITET

La $V_1, \ldots, V_q$ være alle variable som inngår i eller har vesentlig innflytelse på ("forsøksbetingelser") en undersøkelse. Alle variablene antas å være diskrete slik at V_i kan ha m_i ulike kjennetegn. Vektoren $\underset{\sim}{V} = (V_1, \ldots, V_q)$ kan da anta verdier i mengden

$$\Omega^* = \{\underset{\sim}{v}^*_1, \ldots, \underset{\sim}{v}^*_m\} \quad ; \quad m = \prod_{i=1}^{q} m_i$$

Hver enhet i populasjonen kan tillegges akkurat én verdi, $\underset{\sim}{v}^*_j$, på $\underset{\sim}{V}$.

La

$$\P^*_j = \{\text{enheter med } \underset{\sim}{V} = \underset{\sim}{v}^*_j\} \; ; \; j = 1,2,\ldots,m.$$

På grunnlag av partisjonen

$$\P^* = \{\P^*_1, \ldots, \P^*_m\}$$

kan alle (mulige) interessante delpopulasjoner dannes ved sammenslåing av ulike $\P^*_j$-er. F.eks. vil $\bigcup_{j=1}^{m} \P^*_j$ være hele populasjonen. En vilkårlig partisjon vil vi betegne med

$$\P = \{\P_1, \ldots, \P_r\} \; ; \; r \leq m$$

Vi antar at alle[1] enheter i en vilkårlig $\P^*_j$ har samme sannsynlighet for å falle fra i en undersøkelse. Denne egenskapen vil vi kalle <u>frafallshomogenitet</u>: $\P^*$ er en frafallshomogen partisjon. En frafallshomogen partisjon som er slik at ingen delpopulasjoner har samme frafallsannsynlighet, kaller vi en <u>maksimal</u> frafallshomogen partisjon, og betegner den med

$$L = \{L_1, \ldots, L_H\}.$$

---

1) Det forutsettes at ingen $\P^*_j$ består av mindre enn to enheter. (Se diskusjon av dette sist i avsnittet).

Delpopulasjonene L_j er altså homogene m.h.p. frafallsannsynligheten, λ_j, og ikke nødvendigvis m.h.p. noen andre egenskaper.

Framstillingen i avsnitt 3 og 4 er av matematiske grunner basert på antakelsen om at $\lambda_1, \ldots, \lambda_H$ er kjente. I praksis vil denne forutsetningen være urealistisk. For å kunne estimere frafallsannsynlighetene på grunnlag av utvalgsobservasjonene, trenger vi en modell for hvordan λ varierer med de ulike komponentene i $\underset{\sim}{V}$. En slik modell kan f.eks. være en funksjon $\phi: \Omega^* \to [0,1>$ slik at $\lambda = \phi(\underset{\sim}{V})$.

For enheter som faller fra, er (pr. definisjon) situasjonen den at vi ikke oppnår kunnskap om de fleste komponentene i $\underset{\sim}{V}$. Vi vil derfor eliminere alle komponenter som ikke lar seg observere for hele utvalget. Kaller vi restvektoren $\underset{\sim}{W}$ (p komponenter si, $p < q$), så kan vi danne oss en modell $\lambda = \phi'(\underset{\sim}{W})$ og håpe at bildet av denne funksjonen i størst mulig grad faller sammen med det av $\phi(\underset{\sim}{V})$; dvs. at komponentene som er eliminerte, ikke bidrar til å "forklare" hvorfor to enheter kan ha forskjellige frafallsannsynligheter.

En slik modell for λ fører selvsagt til at enhetene innen en delpopulasjon, L_j, blir homogene m.h.p. et sett av kjennetegn, nemlig $\underset{\sim}{W}$, eller bare de av komponentene i $\underset{\sim}{W}$ som har "forklaringskraft" ang. λ.

Vanligvis vil $\underset{\sim}{W}$ være vektoren av variable som finnes i registeret hvorfra utvalget trekkes (alder, kjønn, ekteskapelig status, bosted, husholdningens størrelse) og dessuten intervjuerens alder, kjønn osv.

Wahlstrøm, [8], peker på visse prinsippielle problemer i forbindelse med estimering av frafallsannsynligheter ([8], kap. "8.7 Skattning av svarsbenägenhet"). Hans frafallmodell setter ingen restriksjoner på størrelsen til de frafallshomogene delpopulasjonene. Det oppstår derfor et målingsproblem i de tilfellene der delpopulasjonen bare består av ett individ. Estimering av dette individets "svarbenägenhet" ville kreve nøyaktige gjentakelser av den samme undersøkelsen (på samme tidspunkt), hvilket selvsagt er vanskelig. Skal dette problemet unngås, må vi i frafallsmodellen forutsette at tallet på enheter i hver av delpopulasjonene, $L_1, \ldots, L_H$, er større enn 1. Denne forutsetningen er lite restriktiv og har for det meste teoretisk interesse (vi kan alltid "definere oss bort fra problemet" ved å slå sammen $\P_i^*$-er når frafallsannsynlighetene har en frekventistisk tolkning).

Vi skal her ikke gå nærmere inn på estimeringen av frafallsannsynlighetene. For i større grad å gjøre framstillingen i de to neste avsnittene realistisk, vil de konsekvenser estimering av λ_j-ene får for resultatene i avsnitt 3 og 4, bli vurdert i sluttordet.

3. TO ESTIMATORER

Definer B_{ij} ved

$$B_{ij} = \P_i \cap L_j \quad \text{der} \quad i = 1,2,\ldots,r \text{ og } j = 1,\ldots,H.$$

Alle enheter i B_{ij} har samme frafallsannsynlighet, λ_j.

La videre $N_{ij} = \# B_{ij}$ [1]. Idet

$$\P_i = \bigcup_{j=1}^{H} B_{ij} \quad \text{og} \quad L_j = \bigcup_{i=1}^{r} B_{ij},$$ - begge disjunkte, så blir:

$$\# \P_i = \sum_{j=1}^{H} N_{ij} = N_{i\cdot} \quad , \quad \# L_j = \sum_{i=1}^{r} N_{ij} = N_{\cdot j}$$

og $$\# \{\P\} = \sum_{i,j} N_{ij} = N \text{ (= tallet på enheter i populasjonen).}$$

Vi definerer dessuten:

$$\gamma_{ij} = \frac{N_{ij}}{N} \quad , \quad \gamma_{i\cdot} = \frac{N_{i\cdot}}{N} \quad \text{og} \quad \gamma_{\cdot j} = \frac{N_{\cdot j}}{N} \, .$$

Vi skal i det følgende estimere $\gamma_{i\cdot}$, dvs. det relative antall enheter i populasjonen som tilhører $\P_i$. Som nevnt i innledningen, vil vi betrakte stikkprøven som en Bernoulli forsøksrekke.

Frafallsannsynlighetene, $\lambda_1,\ldots,\lambda_H$, forutsettes å tilfredsstille ulikhetene

$$0 \leq \lambda_j < 1 \quad \text{for alle } j.$$

Hvis $\lambda_j = 1$ for noen j, betyr dette at vi <u>ikke kan</u> oppnå observasjoner fra denne L_j. I så fall vil en ikke kunne finne noen forventningsrette estimatorer for $\gamma_{i\cdot}$ (se bl.a. Wahlstrøm, [8]: sats 4.1).

Estimeringsproblemet vil her bli behandlet i lys av to modeller: én modell <u>uten frafall</u> ($\lambda_1 = \ldots = \lambda_H = 0$) og én der frafall kan forekomme. Sistnevnte modell vil bli referert til som modellen <u>med mulig frafall</u> eller kortere <u>med frafall</u> ($\lambda_j \geq 0$ for alle j og $\lambda_j > 0$ for minst én j).

Notasjoner for <u>utvalgsstørrelser</u>:

	Uten frafall	Med frafall
$\# B_{ij}$	n_{ij}	$\tilde{n}_{ij}$
$\# \P_i$	$n_{i\cdot}$	$\tilde{n}_{i\cdot}$
$\# L_j$	$n_{\cdot j}$	$\tilde{n}_{\cdot j}$
Hele utvalget	n	$\tilde{n}$

---

1) Hvis A er en mengde, så betyr "$\# A$" tallet på elementer i A ("kardinaltallet" til A).

hvor $n_{i.}$, $\tilde{n}_{i.}$, $n_{.j}$ og $\tilde{n}_{.j}$ er definert på samme måte som de tilsvarende populasjonsstørrelser. Av størrelsene ovenfor er det bare n som er ikke-stokastisk under den gitte utvalgsplanen.

La "BU" være en forkortet skrivemåte for "brutto-utvalget" og "NU" for "netto-utvalget" (= BU-frafall). Vi trenger følgende sannsynligheter:

$$\left.\begin{aligned} P[\text{en vilkårlig enhet i BU skal tilhøre } B_{ij}] &= \gamma_{ij} \\ P[\text{en vilkårlig enhet i NU skal tilhøre } B_{ij}] &= \frac{\mu_{ij}}{\mu} \end{aligned}\right\} \quad (3.1)$$

der $\mu_{ij} = \gamma_{ij}(1-\lambda_j)$ og

$$\mu = \sum_{i=1}^{r} \mu_{i.} = \sum_{i,j} \mu_{ij} = \sum_{j=1}^{H} \mu_{.j} = \sum_{j=1}^{H} \gamma_{.j}(1-\lambda_j)$$

μ kan tolkes som sannsynligheten for at en vilkårlig enhet i brutto-utvalget også skal opptre i netto-utvalget, dvs. ikke falle fra.

Vi har følgende fordelinger for de ulike variable:

Uten frafall:

$$P[n_{ij} = x] = \binom{n}{x} \gamma_{ij}^{x} [1-\gamma_{ij}]^{n-x} \qquad (3.2)$$

$$P[n_{i.} = y] = \binom{n}{y} \gamma_{i.}^{y} [1-\gamma_{i.}]^{n-y} \qquad (3.3)$$

Med frafall:

$$P[\tilde{n}_{ij} = x] = \binom{n}{x} \mu_{ij}^{x} [1-\mu_{ij}]^{n-x} \qquad (3.4)$$

$$P[\tilde{n}_{i.} = y] = \binom{n}{y} \mu_{i.}^{y} [1-\mu_{i.}]^{n-y}$$

$$P[\tilde{n} = z] = \binom{n}{x} \mu^{z} (1-\mu)^{n-z} \qquad (3.6)$$

For å gjøre det klart under hvilken modell vi tar forventning, varians osv. vil vi i det følgende anvende denne symbolikken:

	Uten frafall	Med frafall
Forventning	E_U	E_M
Varians	Var	var
Kovarians	Cov	cov
Bruttovarians	Bvar	bvar
Sannsynlighetsmålet	P_U	P_M

Når fullstendig tekst skrives istedenfor symboler, vil modell-henvisningen bli satt i parentes bak de ulike begrepene. Vi bruker da forkortelsene (u.f.) og (m.f.) for "uten" og "med" frafall respektivt.

Vi skal nå finne en estimator for $\gamma_{i.}$

Idet $\gamma_{ij} = \frac{\mu_{ij}}{1-\lambda_j}$; $(\lambda_j<1)$

fås $\gamma_{i.} = \sum_{j=1}^{H} \frac{\mu_{ij}}{1-\lambda_j}$ (3.7)

Av (3.4) følger at observatoren

$$\tilde{\mu}_{ij} = \frac{\tilde{n}_{ij}}{n} \qquad (3.8)$$

har forventning (m.f.) lik μ_{ij}. Dermed blir

$$\tilde{\gamma}_{i.} = \frac{1}{n} \sum_{j=1}^{H} \frac{\tilde{n}_{ij}}{1-\lambda_j} \qquad (3.9)$$

en forventningsrett (m.f.) estimator for $\gamma_{i.}$. Variansen (m.f.) til denne er:

$$\text{var}\, \tilde{\gamma}_{i.} = \frac{1}{n}\{\sum_{j=1}^{H} \frac{\gamma_{ij}}{1-\lambda_j} - \gamma_{i.}^2\} \qquad (3.10)$$

Før vi beviser (3.10), skal vi kort se hva frafallet har medført for variansens vedkommende:

Dersom alle $\lambda_j = 0$, ville $\tilde{\gamma}_{i.}$ være den "ordinære" estimator for $\gamma_{i.}$, $\gamma^*_{i.} = \frac{n_{i.}}{n}$, med varians lik $\gamma_{i.}(1-\gamma_{i.})n^{-1}$. Siden $\lambda_j>0$ for minst én j, vil Σ-leddet i (3.10) være større enn $\gamma_{i.}$, hvilket innebærer at frafallet fører med seg variansøkning. Dessuten ses at jo større λ_j-ene blir, jo større blir var $\tilde{\gamma}_{i.}$.

Bevis for (3.10):

$$\text{var}\, \tilde{\gamma}_{i.} = \frac{1}{n^2}\{\sum_{j=1}^{H} \frac{\text{var}\, \tilde{n}_{ij}}{(1-\lambda_j)^2} + 2\sum_{k<j} \frac{\text{cov}(\tilde{n}_{ik}, \tilde{n}_{ij})}{(1-\lambda_k)(1-\lambda_j)}\}$$

Nå er var $\tilde{n}_{ij} = n\mu_{ij}(1-\mu_{ij})$. Videre er $\tilde{n}_{ij}$ og $\tilde{n}_{ik}$ trinomisk fordelte (m.f.), og da er som kjent: cov $(\tilde{n}_{ij}, \tilde{n}_{ik}) = -n\mu_{ij}\mu_{ik}$

Ved innsetting i uttrykket for var $\tilde{\gamma}_{i.}$ fås:

$$\text{var}\,\tilde{\gamma}_{i.} = \frac{1}{n}\{\sum_j \frac{\mu_{ij}(1-\mu_{ij})}{(1-\lambda_j)^2} - 2\sum_{k<j} \frac{\mu_{ij}}{1-\lambda_j}\frac{\mu_{ik}}{1-\lambda_k}\}$$

Erstatter vi her μ_{ij} med $\gamma_{ij}(1-\lambda_j)$, fås (3.10) ved noen enkle sammentrekkinger. q.e.d.

En annen estimator for $\gamma_{i.}$ er

$$\hat{\gamma}_{i.} = \frac{\tilde{n}_{i.}}{\tilde{n}} \qquad (3.11)$$

(3.11) er det "praktiske resultat" i en modell uten frafall der en anvender estimatoren

$$\gamma^*_{i.} = \frac{n_{i.}}{n} \qquad (3.12)$$

som er forventningsrett (u.f.) og har varians (u.f.)

$$\text{Var}\,\gamma^*_{i.} = \frac{\gamma_{i.}(1-\gamma_{i.})}{n}$$

Ettersom (3.11) er den mest vanlige estimator for $\gamma_{i.}$, skal vi studere dens egenskaper noe nærmere under modellen med mulig frafall. Vi skal imidlertid endre litt på definisjonen av $\hat{\gamma}_{i.}$:

$$\bar{\gamma}_{i.} = \begin{cases} \hat{\gamma}_{i.} & \text{når } \tilde{n}>0 \\ 0 & \text{når } \tilde{n}=0 \end{cases} \qquad (3.13)$$

Vi har nemlig at $P_M(\tilde{n} = 0) > 0$, hvilket medfører at forventningen (m.f.) og variansen (m.f.) til (3.11) ikke eksisterer.

Følgende betingede fordelinger utledes enkelt ved å anvende (3.1):

$$P_M[\tilde{n}_{ij} = x|\tilde{n}] = \binom{\tilde{n}}{x}\left[\frac{\mu_{ij}}{\mu}\right]^x\left[1 - \frac{\mu_{ij}}{\mu}\right]^{\tilde{n}-x} \qquad (3.14)$$

$$P_M[\tilde{n}_{i.} = y|\tilde{n}] = \binom{\tilde{n}}{y}\left[\frac{\mu_{i.}}{\mu}\right]^y\left[1 - \frac{\mu_{i.}}{\mu}\right]^{\tilde{n}-y} \qquad (3.15)$$

Forventningen til (3.13) blir da:

$$E_M\bar{\gamma}_{i.} = E_M\frac{1}{\tilde{n}}E_M[\tilde{n}_{i.}|\tilde{n}] = \frac{\mu_{i.}}{\mu} \qquad (2.16)^{1)}$$

1) Siste likhet i (3.16) er egentlig en tilnærmelse, fordi:

$$E_M\bar{\gamma}_{i.} = 0\cdot P_M(\tilde{n} = 0) + \sum_{z=1}^{n}\frac{\mu_{i.}}{\mu}P_M(\tilde{n} = z) < \frac{\mu_{i.}}{\mu} \text{ når } P_M(\tilde{n} = 0) > 0.$$

Tilnærmelsen (3.16) gjelder når $P_M(\tilde{n} = 0) \simeq 0$.

som også kan skrives

$$E_M\bar{\gamma}_{i.} = \frac{\gamma_{i.} - \sum_{j=1}^{H}\gamma_{ij}\lambda_j}{1 - \sum_{j=1}^{H}\gamma_{.j}\lambda_j} \qquad (3.17)$$

Bruttovariansen (m.f.) til $\bar{\gamma}_{i.}$ er:

$$\text{bvar}\,\bar{\gamma}_{i.} \pm \text{var}\,\bar{\gamma}_{i.} + \Psi_i^2 \qquad (3.18)$$

der

$$\Psi_i^2 = \left[E_M\bar{\gamma}_{i.} - \gamma_{i.}\right]^2 = \left\{\frac{\sum_{j=1}^{H}\lambda_j\left[\gamma_{i.}\gamma_{.j} - \gamma_{ij}\right]}{1 - \sum_{j=1}^{H}\gamma_{.j}\lambda_j}\right\}^2 \qquad (3.19)$$

"Skjevheten" Ψ_i^2 forsvinner hvis for alle j: $\gamma_{i.}\gamma_{.j}$. Nå er

$$\gamma_{ij} = P_U(\P_i|L_j)\,P_U(L_j) =$$

$$= P_U(\P_i|L_j)\,\gamma_{.j}$$

Av dette følger at $\bar{\gamma}_{i.}$ er forventningsrett (m.f.) (dvs. $\Psi_i = 0$) når

$$P_U(\P_i|L_j) = \gamma_{i.} \quad \text{for alle } j^{1)}.$$

Altså vil $\bar{\gamma}_{i.}$ være (tilnærmet) forventningsrett (m.f.) hvis en enhet med den søkte egenskap ($\P_i$) har like stor sannsynlighet (u.f.) for å bli valgt ut ved en enkelttrekking i alle delpopulasjoner L_j; $j = 1,\ldots, H$.

Av (3.17) følger dessuten at $\bar{\gamma}_{i.}$ er forventningsrett (m.f.) hvis alle λ_j-ene er like. Dette er imidlertid et spesialtilfelle av det foranstående idet vi nå har bare én delpopulasjon L_j, nemlig hele populasjonen.

Vi skal så finne variansen (m.f.) til $\bar{\gamma}_{i.}$:

$$\text{var}\,\bar{\gamma}_{i.} = E_M\text{var}\left[\bar{\gamma}_{i.}|\tilde{n}\right] + \text{var}\,E_M\left[\bar{\gamma}_{i.}|\tilde{n}\right] =$$

$$= E_M\text{var}\left[\bar{\gamma}_{i.}|\tilde{n}\right] =$$

$$= \sum_{z=0}^{n}\text{var}\left[\bar{\gamma}_{i.}|\tilde{n}=z\right]\cdot P_M(\tilde{n}=z) =$$

$$= \sum_{z=1}^{n}\text{var}\left[\bar{\gamma}_{i.}|\tilde{n}=z\right]P_M(\tilde{n}=z) = \quad \left[\text{var}(\bar{\gamma}_{i.}|\tilde{n}=0) = 0\right]$$

$$= \sum_{z=1}^{n}\frac{1}{z}\frac{\mu_{i.}}{\mu}\left[1 - \frac{\mu_{i.}}{\mu}\right]\binom{n}{z}\mu^z(1-\mu)^{n-z} =$$

$$= \frac{\mu_{i.}}{\mu}\left[1 - \frac{\mu_{i.}}{\mu}\right]\sum_{z=1}^{n}\frac{1}{z}\binom{n}{z}\mu^z(1-\mu)^{n-z} \qquad (3.20)$$

---

1) Det er tilstrekkelig at $P_U(\P_i|L_j) = \gamma_{i.}$ bare for de j hvor $\lambda_j > 0$.

I den siste summen vil vi nå erstatte $\frac{1}{z}$ med $\frac{1}{z+1}$. Summen blir da noe mindre. På den annen side vil den bli større igjen ved å ta z=0 som nedre grense i summasjonen (dette er nå "tillatt"). For store n vil vi ha:

$$\sum_{z=1}^{n} \frac{1}{z}\binom{n}{z}\mu^z(1-\mu)^{n-z} \approx \sum_{z=0}^{n} \frac{1}{z+1}\binom{n}{z}\mu^z(1-\mu)^{n-z}$$

Summen på høyre side er beregnet av J.M. Hoem ([4] s. 3):

$$\sum_{z=0}^{n} \frac{1}{z+1}\binom{n}{z}\mu^z(1-\mu)^{n-z} = \frac{1-(1-\mu)^{n+1}}{\mu(n+1)} \approx \frac{1}{n\mu} \qquad (3.21)$$

Ved innsetting i (3.20), fås:

$$\text{var}\ \bar{\gamma}_{i.} \approx \frac{1}{n\mu}\ \frac{\mu_{i.}}{\mu}\left[1 - \frac{\mu_{i.}}{\mu}\right] \qquad (3.22)$$

som altså gjelder for "store" n.

Bruttovariansen (m.f.) til $\bar{\gamma}_{i.}$ blir nå (tilnærmet):

$$\text{bvar}\ \bar{\gamma}_{i.} = \frac{1}{n\mu}\,\frac{\mu_{i.}}{\mu}\left[1 - \frac{\mu_{i.}}{\mu}\right] + \left[\frac{\mu_{i.}}{\mu} - \gamma_{i.}\right]^2 \qquad (3.23)$$

4. SAMMENLIGNING AV ESTIMATORENE

Vi skal nå sammenligne estimatorene $\bar{\gamma}_{i.}$ og $\tilde{\gamma}_{i.}$. Som mål på kvalitets-forskjellen vil vi bruke

$$\Delta_i = \text{bvar}\ \bar{\gamma}_{i.} - \text{var}\ \tilde{\gamma}_{i.} \qquad (4.1)$$

$\bar{\gamma}_{i.}$ vil betegnes som bedre evt. dårligere enn $\tilde{\gamma}_{i.}$ ettersom $\Delta_i < 0$ evt. $\Delta_i > 0$. Før vi drøfter verdien av Δ_i, skal vi finne det mulige variasjons-området for ψ_i.

Vi har:

$$\mu_{i.} = \begin{cases} \sum_j \gamma_{ij}(1-\lambda_j) \le \gamma_{i.} \\ \mu - \sum_{k\neq i}\mu_{k.} \le \mu \end{cases}$$

Da disse ulikhetene alltid gjelder, vil

$$\mu_{i.} \le \min\{\mu;\ \gamma_{i.}\}$$

Dessuten er:

$$\mu_{i.} = \mu - \sum_{k\neq i}\mu_{k.} \ge \begin{cases} \mu - [1-\gamma_{i.}] \\ 0 \end{cases}$$

Altså vil alltid $\mu_{i.} \geq \max\{0;\ \mu - [1-\gamma_{i.}]\}$

som også kan skrives

$$\mu_{i.} \geq \mu - \min\{\mu;\ 1-\gamma_{i.}\}$$

Samlet har vi dermed at

$$\mu - \min\{\mu;\ 1-\gamma_{i.}\} \leq \mu_{i.} \leq \min\{\mu;\ \gamma_{i.}\}$$

hvilket medfører at variasjons-området for Ψ_i må ligge i følgende intervaller:

Skjema 1: Mulige verdiområder for Ψ_i når $\gamma_{i.}$ og μ varierer:

$\mu \leq \min\{\gamma_{i.}, 1-\gamma_{i.}\}$	$\gamma_{i.} \leq \frac{1}{2}$ og $\gamma_{i.} \leq \mu \leq 1-\gamma_{i.}$	$\gamma_{i.} \geq \frac{1}{2}$ og $1-\gamma_{i.} \leq \mu \leq \gamma_{i.}$	$\mu \geq \max\{\gamma_{i.}, 1-\gamma_{i.}\}$
$-\gamma_{i.} \leq \Psi_i \leq 1-\gamma_{i.}$	$-\gamma_{i.} \leq \Psi_i \leq \frac{1-\mu}{\mu}\gamma_{i.}$	$-\frac{1-\mu}{\mu}[1-\gamma_{i.}] \leq \Psi_i \leq 1-\gamma_{i.}$	$-\frac{1-\mu}{\mu}[1-\gamma_{i.}] \leq \Psi_i \leq \frac{1-\mu}{\mu}\gamma_{i.}$

Det ses umiddelbart at alle intervallene inneholder punktet $\Psi_i = 0$.

Vi går så over til drøftingen av Δ_i, og betrakter først det tilfellet at alle λ_j-ene er like, $\lambda_j = \lambda$. I så fall er $\mu_{i.} = \gamma_{i.}(1-\lambda)$ og $\mu = 1-\lambda$. Dessuten er $\bar{\gamma}_{i.}$ (tilnærmet) forventningsrett, og vi finner at

$$\Delta_i = -\frac{\lambda}{n(1-\lambda)}\gamma_{i.}^2 = -\frac{1-\mu}{n\mu}\gamma_{i.}^2$$

Da dette uttrykket er negativt, vil $\bar{\gamma}_{i.}$ være å foretrekke framfor $\tilde{\gamma}_{i.}$. Vi ser dessuten at $\bar{\gamma}_{i.}$ blir relativt bedre jo større λ er.

Anta nå at λ_j-ene er innbyrdes ulike. Da har vi for Δ_i:

$$\Delta_i = \frac{1}{n\mu}\frac{\mu_{i.}}{\mu}\left[1-\frac{\mu_{i.}}{\mu}\right] + \left[\frac{\mu_{i.}}{\mu} - \gamma_{i.}\right]^2 - \frac{1}{n}\left[\sum_{j=1}^{H}\frac{\gamma_{ij}}{1-\lambda_j} - \gamma_{i.}^2\right]$$

Setter vi her inn for $\left[\frac{\mu_{i.}}{\mu} - \gamma_{i.}\right] = \Psi_i$, fås ved noen enkle omforminger:

$$\Delta_i = \frac{n\mu-1}{n\mu}\Psi_i^2 + \frac{1-2\gamma_{i.}}{n\mu}\Psi_i + \frac{1}{n\mu}\left[\sum_{j=1}^{H}\gamma_{ij}\left(1-\frac{\mu}{1-\lambda_j}\right) - (1-\mu)\gamma_{i.}^2\right] \qquad (4.2)$$

Anta at n og μ er så store at $(n\mu-1)/n\mu \tilde{\approx} 1$. Dessuten vil - når variasjonene i λ_j-ene ikke er "altfor" store - følgende tilnærmelse gjelde:

$\frac{1}{n\mu}\sum_j \gamma_{ij}(1-\frac{\mu}{1-\lambda_j}) \approx 0$. (4.2) kan da skrives:

$$\Delta_i \approx \psi_i^2 + \frac{1-2\gamma_{i.}}{n\mu}\psi_i - \frac{1-\mu}{n\mu}\gamma_{i.}^2 \qquad (4.3)$$

(Den siste tilnærmelsen før (4.3) kan begrunnes ved å sette $\psi_i = 0$ i (4.3). Da fås resultatet vi utledet direkte foran ved å sette alle $\lambda_j = \lambda$, - et tilfelle der $\psi_i = 0$).

Vi finner at Δ_i som funksjon av ψ_i har sitt matematiske minimum ("matmin")[1] for

$$\psi_i = \frac{1}{n\mu}\left[\gamma_{i.} - \tfrac{1}{2}\right] = \psi_i^* \qquad (4.4)$$

Da er

$$\underset{\psi_i}{\text{matmin}}\ \Delta_i = -\frac{1}{(n\mu)^2}\{[1+n\mu(1-\mu)]\gamma_{i.}^2 - \gamma_{i.} + \tfrac{1}{4}\} \qquad (4.5)$$

For tilstrekkelig stor n vil (4.4) alltid være et indre punkt i ψ_i-intervallene gitt i skjema 1. I motsatt fall vil Δ_i anta sin minimumsverdi i et av endepunktene for intervallene. De forskjellige tilfellene er satt opp i skjema 2.

Vi skal i det følgende basere drøftingene på den forutsetning at ψ_i^* er et indre punkt i ψ_i-intervallene, dvs. at n er relativt stor. For å vurdere rekkevidden av denne forutsetningen, skal vi bestemme de minste verdier n kan ha for at forutsetningen skal gjelde. "Den minste n" er i tabell 1 definert som den største nedre skranke for n ifølge skjema 2 for ulike verdier av $\gamma_{i.}$ og μ.

Tabell 1: Den minste verdi n kan ha for at ψ_i^* skal være et indre punkt i ψ_i-intervallene

μ \ $\gamma_{i.}$	0,05	0,10	0,30	0,50	0,70	0,90	0,95
0,90	10	4	3	0	3	4	10
0,80	12	5	2	0	2	5	12
0,70	13	6	1	0	1	6	13
0,60	15	7	2	0	2	7	15

1) Vi skiller her mellom det matematiske minimum og "det praktiske" som er den minste verdi Δ_i antar når ψ_i varierer i et begrenset, "tillatt" område (skjema 1).

Skjema 2: Minimerende ψ_i-verdier når ψ_i^* ikke minimerer Δ_i

Intervall for ψ_i	$-\gamma_{i.} \le \psi_i \le 1-\gamma_{i.}$	$-\gamma_{i.} \le \psi_i \le \frac{1-\mu}{\mu}\gamma_{i.}$	$-\frac{1-\mu}{\mu}\lceil 1-\gamma_{i.}\rceil \le \psi_i \le 1-\gamma_{i.}$	$-\frac{1-\mu}{\mu}\lceil 1-\gamma_{i.}\rceil \le \psi_i \le \frac{1-\mu}{\mu}\gamma_{i.}$
Betingelser som må oppfylles for at ψ_i^* skal falle til <u>høyre</u> for intervallet	(i): $\gamma_{i.} > \frac{1}{2}$, $\mu \le 1-\gamma_{i.}$ (ii): $n < \frac{\gamma_{i.}^{-\frac{1}{2}}}{\mu(1-\gamma_{i.})}$ Minimerende ψ_i: $\psi_i = 1-\gamma_{i.}$	ψ_i^* kan ikke falle til høyre for dette intervallet idet intervallet gjelder bare når $\gamma_{i.} \le \frac{1}{2}$	(i): $\gamma_{i.} > \frac{1}{2}$ og $1-\gamma_{i.} \le \mu \le \gamma_{i.}$ (ii): $n < \frac{\gamma_{i.}^{-\frac{1}{2}}}{\mu(1-\gamma_{i.})}$ Minimerende ψ_i: $\psi_i = 1-\gamma_{i.}$	(i): $\gamma_{i.} > \frac{1}{2}$, $\mu \ge \gamma_{i.}$ (ii): $n < \frac{\gamma_{i.}^{-\frac{1}{2}}}{(1-\mu)\gamma_{i.}}$ Minimerende ψ_i: $\psi_i = \frac{1-\mu}{\mu}\gamma_{i.}$
Betingelser som må oppfylles for at ψ_i^* skal falle til <u>venstre</u> for intervallet	(i): $\gamma_{i.} < \frac{1}{2}$, $\mu \le \gamma_{i.}$ (ii): $n < \frac{\frac{1}{2}-\gamma_{i.}}{\mu\gamma_{i.}}$ Minimerende ψ_i: $\psi_i = -\gamma_{i.}$	(i): $\gamma_{i.} < \frac{1}{2}$, $\mu \ge \gamma_{i.}$ (ii): $n < \frac{\frac{1}{2}-\gamma_{i.}}{\mu\gamma_{i.}}$ Minimerende ψ_i: $\psi_i = -\gamma_{i.}$	ψ_i^* kan ikke falle til venstre for dette intervallet idet intervallet gjelder bare når $\gamma_{i.} \ge \frac{1}{2}$	(i): $\gamma_{i.} < \frac{1}{2}$, $\mu \ge 1-\gamma_{i.}$ (ii): $n < \frac{\frac{1}{2}-\gamma_{i.}}{(1-\mu)(1-\gamma_{i.})}$ Minimerende ψ_i: $\psi_i = -\frac{1-\mu}{\mu}\lceil 1-\gamma_{i.}\rceil$

Med Ψ_i^* som indre punkt i Ψ_i-intervallene, vil min Δ_i være gitt ved (4.5). Det innses lett at denne verdien alltid er negativ. Δ_i antar nå sine maksimale verdier i intervall-endepunktene. Lar vi $\Delta_i(p)$ være den verdien Δ_i antar for $\Psi_i = p$, så finner vi at de ulike maksimumspunktene[1] blir:

$$\Delta_i(-\gamma_{i.}) = \frac{[(n+1)\mu + 1]\gamma_{i.}^2 - \gamma_{i.}}{n\mu} \tag{4.6}$$

$$\Delta_i(-\frac{1-\mu}{\mu}(1-\gamma_{i.})) = \frac{1-\mu}{n\mu^2}\{[n(1-\mu) - 1](1-\gamma_{i.})^2 + \gamma_{i.}[1-\gamma_{i.}(1-\mu)]\} \tag{4.7}$$

$$\Delta_i(1-\gamma_{i.}) = \frac{[(n+1)\mu + 1]\gamma_{i.}^2 - (2n\mu + 3)\gamma_{i.} + n\mu + 1}{n\mu} \tag{4.8}$$

$$\Delta_i(\frac{1-\mu}{\mu}\gamma_{i.}) = \frac{[n-2-(2n-1)\mu + (n+1)\mu^2]\gamma_{i.}^2 + (1-\mu)\gamma_{i.}}{n\mu^2} \tag{4.9}$$

Betrakter vi fortegnet til disse Δ_i-verdiene og samtidig tar hensyn til skjema 1, får vi de resultatene som er oppstilte i skjema 3.

Av skjema 3 ser vi at negative Δ_i-verdier i endepunktene kan forekomme bare for relativt små n. Det framgår dessuten at ingen av endepunktene alltid er negative.

Løser vi ligningen $\Delta_i = 0$ m.h.p. Ψ_i, finner vi for hvilke Ψ_i de to estimatorene er like gode. Løsningene blir:

$$\Psi_i = \begin{cases} a = \frac{1}{n\mu}\{\gamma_{i.} - \frac{1}{2} - \sqrt{n\mu(1-\mu)\gamma_{i.}^2 + (\gamma_{i.} - \frac{1}{2})^2}\} \\ b = \frac{1}{n\mu}\{\gamma_{i.} - \frac{1}{2} + \sqrt{n\mu(1-\mu)\gamma_{i.}^2 + (\gamma_{i.} - \frac{1}{2})^2}\} \end{cases} \tag{4.10}$$

Når n vokser, går begge punktene (4.10) mot 0. Det ses forøvrig at for alle n, μ og $\gamma_{i.}$, så er $a \leq 0$ og $b \geq 0$.

I diagram 1 er Δ_i framstilt grafisk som funksjon av Ψ_i for det tilfellet at $\Delta_i > 0$ i begge endepunktene (h.h.v. α og β) på Ψ_i-intervallet, samt at Ψ_i^* er et indre punkt i dette intervallet.

---

1) Tar vi ikke hensyn til forutsetningen om at Ψ_i^* skal være et indre punkt i Ψ_i-intervallene, vil selvsagt en av endepunktsverdiene være minimum av Δ_i (jfr. skjema 2).

Skjema 3: Δ_i's fortegn i endepunktene på Ψ_i-intervallet

Δ_i- verdi	$\Delta_i(-\gamma_{i.})$	$\Delta_i\left[-\frac{1-\mu}{\mu}(1-\gamma_{i.})\right]$	$\Delta_i(1-\gamma_{i.})$	$\Delta_i(\frac{1-\mu}{\mu}\gamma_{i.})$
Positiv når:	$n > \frac{1-\gamma_{i.}(1+\mu)}{\mu\gamma_{i.}}$ Høyresiden er positiv når: $\gamma_{i.} < \frac{1}{1+\mu}$	$n > \frac{1-\frac{\gamma_{i.}[1-\gamma_{i.}(1-\mu)]}{(1-\gamma_{i.})^2}}{1-\mu}$ Høyresiden er positiv når: $\gamma_{i.} < \frac{3-\sqrt{1+4\mu}}{2(2-\mu)}$ og $\gamma_{i.} \frac{3+\sqrt{1+4\mu}}{2(2-\mu)}$	$n > \frac{3\gamma_{i.} - 1 - (1+\mu)\gamma_{i.}^2}{\mu(1-\gamma_{i.})^2}$ Høyresiden er positiv når: $\frac{3-\sqrt{5-4\mu}}{2(1+\mu)} < \gamma_{i.} <$ $< \frac{3+\sqrt{5-4\mu}}{2(1+\mu)}$	$n > \frac{(2+\mu)\gamma_{i.} - 1}{(1-\mu)\gamma_{i.}}$ Høyresiden er positiv når: $\gamma_{i.} > \frac{1}{2+\mu}$
Negativ når:	$\gamma_{i.} < \frac{1}{1+\mu}$ og $n < \frac{1-\gamma_{i.}(1+\mu)}{\mu\gamma_{i.}}$	$\gamma_{i.} < \frac{3-\sqrt{1+4\mu}}{2(2-\mu)}$ eller $\gamma_{i.} > \frac{3+\sqrt{1+4\mu}}{2(2-\mu)}$ og $n < \frac{1-\frac{\gamma_{i.}[1-\gamma_{i.}(1-\mu)]}{(1-\gamma_{i.})^2}}{1-\mu}$	$\frac{3-\sqrt{5-4\mu}}{2(1+\mu)} < \gamma_{i.} <$ $\frac{3+\sqrt{5-4\mu}}{2(1+\mu)}$ og $n < \frac{3\gamma_{i.} - 1 - (1+\mu)\gamma_{i.}^2}{\mu(1-\gamma_{i.})^2}$	$\gamma_{i.} < \frac{1}{2+\mu}$ og $n < \frac{(2+\mu)\gamma_{i.} - 1}{(1-\mu)\gamma_{i.}}$
Δ_i-verdien er aktuell når: (iflg. Skjema 1)	Enten: $\mu \le \min\{\gamma_{i.}, 1-\gamma_{i.}\}$ eller: $\gamma_{i.} \le \frac{1}{2}$ og $\gamma_{i.} \le \mu \le 1-\gamma_{i.}$	Enten: $\gamma_{i.} \ge \frac{1}{2}$ og $1-\gamma_{i.} \le \mu \le \gamma_{i.}$ eller: $\mu \ge \max\{\gamma_{i.}, 1-\gamma_{i.}\}$	Enten: $\mu \le \min\{\gamma_{i.}, 1-\gamma_{i.}\}$ Eller: $\gamma_{i.} \ge \frac{1}{2}$ og $1-\gamma_{i.} \le \mu \le \gamma_{i.}$	Enten: $\gamma_{i.} \le \frac{1}{2}$ og $\gamma_{i.} \le \mu \le 1-\gamma_{i.}$ Eller: $\mu \ge \max\{\gamma_{i.}, 1-\gamma_{i.}\}$

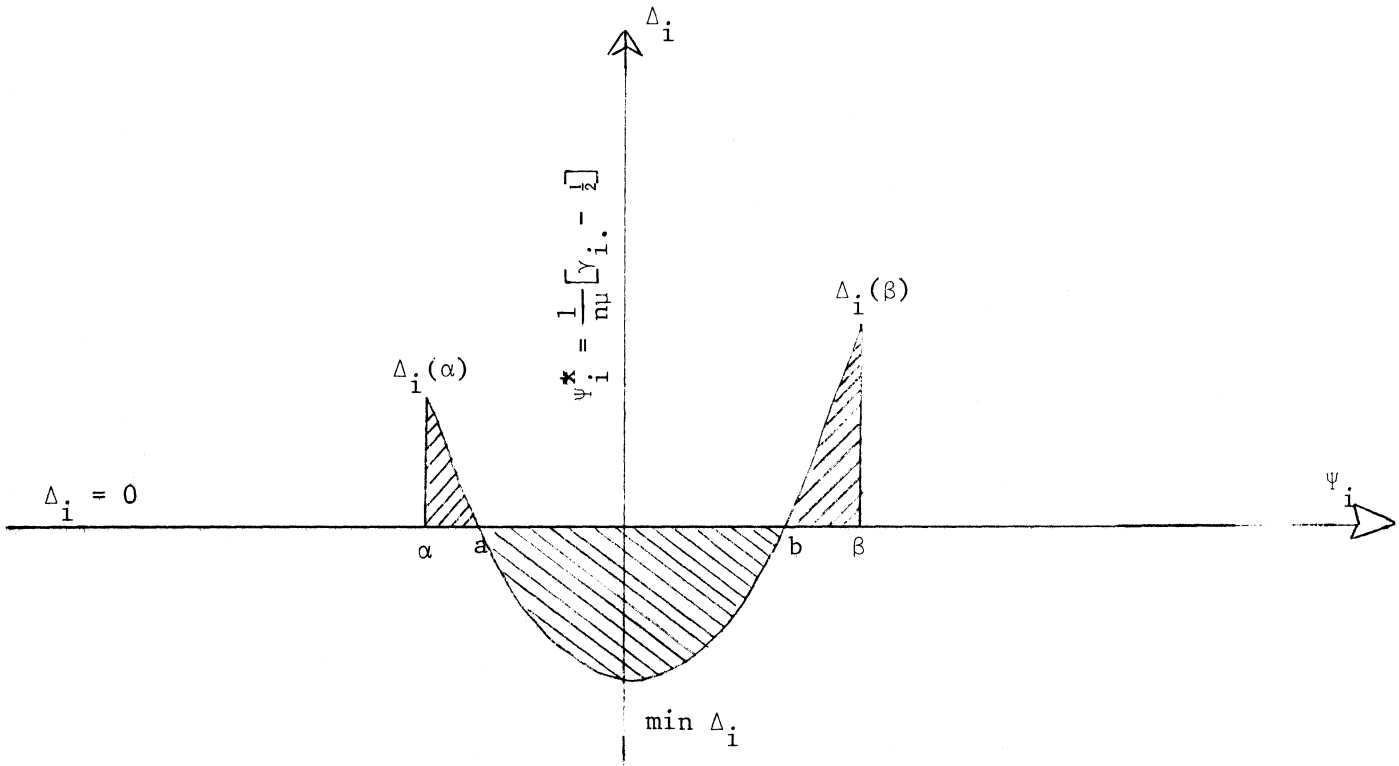

Diagram 1. Δ_i framstilt grafisk som funksjon av Ψ_i

Av diagrammet avleses følgende:

(1) $\tilde{\gamma}_{i.}$ er minst like god som $\bar{\gamma}_{i.}$ når

$$\alpha \leq \Psi_i \leq a \quad \text{og} \quad \text{når} \quad b \leq \Psi_i \leq \beta \tag{4.11}$$

(2) $\bar{\gamma}_{i.}$ er minst like god som $\tilde{\gamma}_{i.}$ når

$$a \leq \Psi_i \leq b \tag{4.12}$$

Lignende diagrammer kan tegnes for de øvrige mulige tilfellene som kan forekomme. Vi skal imidlertid ikke gå nærmere inn på disse da situasjonen som er illustrert i diagram 1, vil være den mest vanlige for relativt store n. De øvrige tilfellene lar seg forøvrig drøfte på grunnlag av skjema 1, 2 og 3 uten nevneverdige vansker.

Siden Ψ_i avhenger av bl.a. den ukjente parameteren $\gamma_{i.}$, kan ikke (1) og (2) ovenfor anvendes umiddelbart. Vi kan imidlertid estimere Ψ_i med

$$\hat{\Psi}_i = \bar{\gamma}_{i.} - \tilde{\gamma}_{i.} \tag{4.13}$$

som er (tilnærmet) forventningsrett. Videre kan a og b estimeres med h.h.v.

$$\begin{aligned} \hat{a} &= \frac{1}{n\mu}\{\tilde{\gamma}_{i.} - \tfrac{1}{2} - \sqrt{n\mu(1-\mu)\tilde{\gamma}_{i.}^2 + (\tilde{\gamma}_{i.} - \tfrac{1}{2})^2}\} \\ \hat{b} &= \frac{1}{n\mu}\{\tilde{\gamma}_{i.} - \tfrac{1}{2} + \sqrt{n\mu(1-\mu)\tilde{\gamma}_{i.}^2 + (\tilde{\gamma}_{i.} - \tfrac{1}{2})^2}\} \end{aligned} \tag{4.14}$$

Kriteriet vi nå kan bruke for å avgjøre hvilken estimator som er (synes å være) best, er:

Velg $\bar{\gamma}_{i.}$ som estimator hvis $\hat{a} < \hat{\Psi}_i < \hat{b}$. I motsatt fall er $\tilde{\gamma}_{i.}$ å foretrekke.

(4.15)

Dette kriteriet er selvsagt beheftet med en viss usikkerhet p.g.a. estimeringen.

Eksempel. (Konstruerte tall)

Anta at vi i en undersøkelse om folks holdning til spørsmål om norsk medlemskap i EF har fire delpopulasjoner i den maksimale frafallshomogene partisjonen: $L = \{L_1, L_2, L_3, L_4\}$. Bruttoutvalgets størrelse er $n = 200$. Forøvrig har vi følgende tall-oppgaver:

Delpopulasjon	Antall "JA"	Antall "NEI"	Frafallsannsynlighet
L_1	5	6	$\lambda_1 = 0{,}48$
L_2	15	30	$\lambda_2 = 0{,}21$
L_3	10	4	$\lambda_3 = 0{,}26$
L_4	40	50	$\lambda_4 = 0{,}13$
Hele utvalget	70	90	$\mu = 0{,}80$

Vi skal estimere hvor stor andel av populasjonen ($\gamma_{i.}$) som mener "JA". Ved utregning finner vi:

$$\bar{\gamma}_{i.} = 0{,}438 \quad \text{og} \quad \tilde{\gamma}_{i.} = 0{,}441$$

Herav fås at $\hat{\Psi}_i = 0{,}003$.

Siden a alltid er mindre (eller lik) 0, er det tilstrekkelig å undersøke hvorvidt $\hat{\Psi}_i$ er større eller mindre enn $\hat{b}$, som finnes å være: $\hat{b} = 0{,}015$. Altså er $\hat{\Psi}_i < b$, og dermed vil det være rimelig å foretrekke estimatoren $\bar{\gamma}_{i.}$.

Anta nå at n varierer, men at alle øvrige størrelser - unntatt $\hat{b}$ - holdes fast. Det kan da være interessant å se hvor stor n må velges for at $\tilde{\gamma}_{i.}$ skal være den beste estimatoren. n må da tilfredsstille ulikheten $\hat{b} < \hat{\Psi}_i = 0{,}003$. Vi finner ved innsetting i (4.16) at i så fall må $n > 5000$ (omtrent)!

På den annen side finner vi at for $n = 200$ er (est)bvar $\bar{\gamma}_{i.} = 0{,}0015$. Skal var $\tilde{\gamma}_{i.}$ ha denne verdien, må n velges lik 250 (ca.). (Her har vi altså to utvalg).

Eksemplet leder oss til å studere nærmere hvordan Δ_i avhenger av utvalgsstørrelsen n.

Av (4.3) følger at

$$\lim_{n\to\infty} \Delta_i = \Psi_i^2 \geq 0$$

dvs. - under de forutsetninger analysen er basert på (N "veldig" stor) - at for tilstrekkelig store utvalg vil $\tilde{\gamma}_{i.}$ være den generelt beste estimatoren. Eksemplet illustrerer imidlertid at utvalgsstørrelsen, n, kan måtte velges ekstremt stor for å forsvare en ensidig anvendelse av $\tilde{\gamma}_{i.}$ ($n > 5000$ i eksemplet). Dersom skjevheten i eksemplet hadde vært større, ville åpenbart $\tilde{\gamma}_{i.}$ bli minst like god som $\tilde{\gamma}_{i.}$ for en mindre n. La oss derfor betrakte skjevheten, Ψ_i, som gitt, og studere hvorden Δ_i varierer med n. Δ_i kan skrives:

$$\Delta_i = \Psi_i^2 - \frac{1}{n}\,\frac{1-\mu}{\mu}\left[(\gamma_{i.} + \frac{\Psi_i}{1-\mu})^2 - \frac{\Psi_i}{1-\mu}(1 + \frac{\Psi_i}{1-\mu})\right]$$

Betrakt uttrykket i hakeparentesen:

$$\phi_i = (\gamma_{i.} + \frac{\Psi_i}{1-\mu})^2 - \frac{\Psi_i}{1-\mu}(1 + \frac{\Psi_i}{1-\mu}) \qquad (4.16)$$

Vi finner at:

(i) $\phi_i \geq 0$ når $\Psi_i > 0$ og $\gamma_{i.} \geq \sqrt{\frac{\Psi_i}{1-\mu}(1 + \frac{\Psi_i}{1-\mu})} - \frac{\Psi_i}{1-\mu}$

(ii) $\phi_i \leq 0$ når $\Psi_i > 0$ og $\gamma_{i.} \leq \sqrt{\frac{\Psi_i}{1-\mu}(1 + \frac{\Psi_i}{1-\mu})} - \frac{\Psi_i}{1-\mu}$

(iii) $\phi_i \geq 0$ når $-(1-\mu) \leq \Psi_i \leq 0$ (alle $\gamma_{i.}$)

(iv) $\phi_i \leq 0$ når $\Psi_i < -(1-\mu)$ og

$$\gamma_{i.} \geq -\frac{\Psi_i}{1-\mu} - \sqrt{\frac{\Psi_i}{1-\mu}(1 + \frac{\Psi_i}{1-\mu})}$$

(v) $\phi_i > 0$ når $\Psi_i < -(1-\mu)$ og

$$\gamma_{i.} < -\frac{\Psi_i}{1-\mu} - \sqrt{\frac{\Psi_i}{1-\mu}(1 + \frac{\Psi_i}{1-\mu})}$$

Tilfellene (iv) og (v) må anses som bare av teoretisk interesse idet $\Psi_i < -(1-\mu)$ må karakteriseres som ekstremt store skjevheter (vanligvis).

I diagram 2 er Δ_i avbildet som funksjon av n[1] for de to tilfellene $\phi_i > 0$ og $\phi_i < 0$ (for $\phi_i = 0$ er $\Delta_i = \Psi_i^2$ og uavhengig av n).

Løses ligningen $\Delta_i = 0$ m.h.p. n ("tillatt" løsning bare for $\phi_i > 0$), fås:

$$n = \frac{1-\mu}{\mu}\frac{\phi_i}{\Psi_i^2} = n_0 \qquad (4.17)$$

Hvis n blir <u>større</u> enn denne verdien, vil $\tilde{\gamma}_{i.}$ alltid være bedre enn $\bar{\gamma}_{i.}$.

---

1) n oppfattes nå som en variabel på de ikke-negative reelle tall.

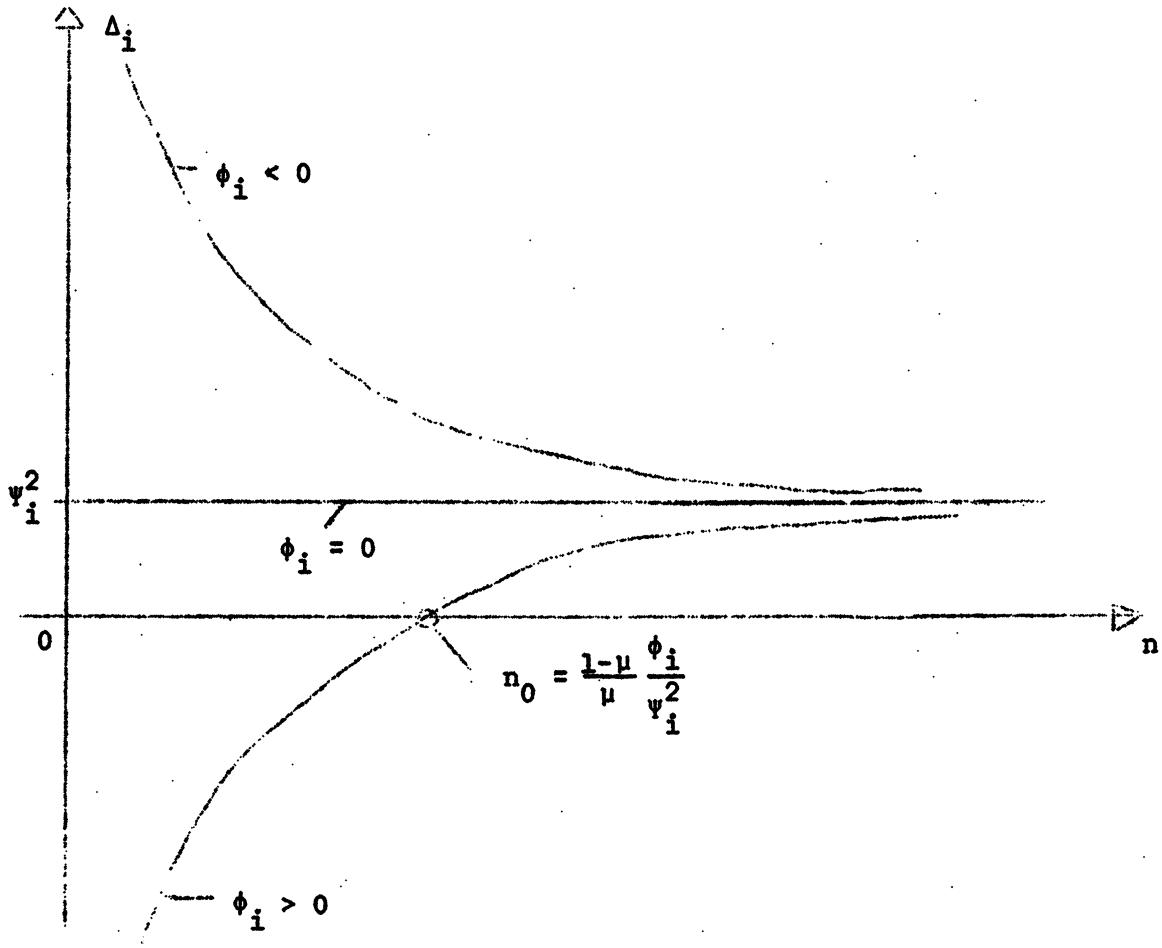

Diagram 2. Δ_i som funksjon av n.

La oss gjøre denne analysen noe mer konkret ved å se på de mest "vanlige" tilfellene som forekommer i praksis. Med "vanlige" tilfeller skal vi mene:

$$\mu = 0{,}8; \quad 0{,}01 \geq \gamma_{i.} \leq 0{,}50 \quad \text{og} \quad -0{,}05 \leq \psi_i \leq 0{,}05$$

Setter vi inn for μ, fås

$$\Delta_i = \psi_i^2 - \frac{1}{n}\frac{1}{4}\phi_i$$

der

$$\phi_i = (\gamma_{i.} + 5\psi_i)^2 - 5\psi_i(1+5\psi_i)$$

Punktene (i) - (iii) blir nå ((iv) og (v) er ikke aktuelle):

(i)' $\phi_i \geq 0$ når $\psi_i > 0$ og $\gamma_{i.} \geq \sqrt{5\psi_i(1+5\psi_i)} - 5\psi_i$

(ii)' $\phi_i \leq 0$ når $\psi_i > 0$ og $\gamma_{i.} \leq \sqrt{5\psi_i(1+5\psi_i)} - 5\psi_i$

(iii)' $\phi_i > 0$ når $\psi_i < 0$

Nedenfor er den minste verdi n kan anta for at $\tilde{\gamma}_{i.}$ med sikkerhet (dvs. for alle verdier av ϕ_i) skal være like god som $\bar{Y}_{i.}$, tabulert for ulike, "vanlige" verdier av $\gamma_{i.}$ og Ψ_i. Denne n-verdien er gitt ved (4.17) som med $\mu = 0{,}8$ blir:

$$n_0 = \frac{(\gamma_{i.}+5\Psi_i)^2 - 5\Psi_i(1+5\Psi_i)}{4\Psi_i^2} ,$$

og er i tabellen kalt "kritisk n-verdi".

Tabell 2: Kritisk n-verdi for ulike verdier av $\gamma_{i.}$ og Ψ_i. $\mu = 0{,}8$

Ψ_i \ $\gamma_{i.}$	0,01	0,05	0,10	0,20	0,30	0,40	0,50
- 0,05	*[1]	23	21	19	19	21	25
- 0,04	*	29	27	25	27	32	40
- 0,03	*	39	37	37	42	53	70
- 0,02	*	58	57	63	82	113	157
- 0,01	129	119	125	175	275	425	625
- 0,008	154	151	165	250	415	657	977
- 0,006	206	206	237	404	710	1 154	1 737
- 0,004	309	322	407	814	1 532	2 564	3 907
- 0,002	619	719	1 126	2 876	5 876	10 126	15 626
0	∞	∞	∞	∞	∞	∞	∞
0,002	0[2]	0	126	2 126	5 376	9 876	15 626
0,004	*	0	0	438	1 282	2 439	3 907
0,006	*	0	0	154	543	1 034	1 736
0,008	*	0	0	63	290	595	978
0,01	*	0	0	26	176	376	626
0,02	*	*	0	0	44	101	169
0,03	*	*	*	0	9	37	71
0,04	*	*	*	0	2	19	40
0,05	*	*	*	0	0	12	25

1) "*" betyr at verdikombinasjonen av $\gamma_{i.}$ og Ψ_i ikke er tillatt iflg. skjema 1.

2) "0" betyr at $\phi_i < 0$, og da er iflg. (4.17) n_0 negativ, hvilket ikke er noen "tillatt" verdi. I dette tilfellet er $\tilde{\gamma}_{i.}$ den beste estimatoren uansett størrelse på utvalget, n.

De forholdene som avspeiles i tabell 2, er stort sett gyldige om en også lar μ variere innen en "vanlig" verdiområde (0,70 - 0,90).

Resultatene av denne drøftingen er for oversiktens skyld oppsummert i skjema 4.

Skjema 4. Oversikt over preferanseområdene for estimatorene $\bar{\gamma}_{i.}$ og $\tilde{\gamma}_{i.}$

	$\Psi_i < -(1-\mu)$		$-(1-\mu) \leq \Psi_i \leq 0$	$\Psi_i > 0$	
	$\gamma_{i.} < -\frac{\Psi_i}{1-\mu} - \sqrt{\frac{\Psi_i}{1-\mu}(1+\frac{\Psi_i}{1-\mu})}$	$\gamma_{i.} \geq -\frac{\Psi_i}{1-\mu} - \sqrt{\frac{\Psi_i}{1-\mu}(1+\frac{\Psi_i}{1-\mu})}$	Alle $\gamma_{i.}$	$\gamma_{i.} \leq -\frac{\Psi_i}{1-\mu} + \sqrt{\frac{\Psi_i}{1-\mu}(1+\frac{\Psi_i}{1-\mu})}$	$\gamma_{i.} > -\frac{\Psi_i}{1-\mu} + \sqrt{\frac{\Psi_i}{1-\mu}(1+\frac{\Psi_i}{1-\mu})}$
$n<n_0$	$\bar{\gamma}_{i.}$ er best	Ikke aktuelt idet $n_0 = 0$	$\bar{\gamma}_{i.}$ er best	Ikke aktuelt idet $n_0 = 0$	$\bar{\gamma}_{i.}$ er best
$n \geq n_0$	$\tilde{\gamma}_{i.}$ er best	$\tilde{\gamma}_{i.}$ er best	$\tilde{\gamma}_{i.}$ er best	$\tilde{\gamma}_{i.}$ er best	$\tilde{\gamma}_{i.}$ er best
	(v)	(iv)	(iii)	(ii)	(i)

I praksis vil de tre kolonnene lengst til høyre være mest aktuelle. Det forholdet at $\bar{\gamma}_{i.}$ tenderer å være den beste estimatoren for små n, står kanskje noe i strid med det en vanligvis er tilbøyelig til å tro om frafallets virkninger, nemlig at det er mest påkrevet å ta hensyn til frafallet når utvalgsstørrelsen er liten. At dette synet ikke alltid er like holdbart, ses tydelig av tabell 2 og skjema 4. Forøvrig framgår dette forholdet umiddelbart av (4.1) idet skjevheten, Ψ_i, jo ikke avhenger av n, mens variansen til $\tilde{\gamma}_{i.}$ som regel vil være større enn variansen til $\bar{\gamma}_{i.}$. Forskjellen er dessuten størst for små n.

5. SLUTTORD

Analysen i foregående avsnitt er en illustrasjon på hvordan en - når visse betingelser er oppfylt - konkret kan vurdere om det er nødvendig å korrigere for frafall. Skal teorien utvides til f.eks. to-trinns utvelging, oppstår matematiske problemer som vanskeliggjør en noenlunde oversiktlig analyse.

Imidlertid er det rimelig å tro at resultatene av en slik analyse ikke vil avvike noe vesentlig fra dem vi har funnet for enkel tilfeldig utvelging. Denne antakelsen er rimelig fordi det er ingenting som taler for at to-trinns utvelging i seg selv endrer det innbyrdes forholdet mellom

estimatorene $\bar{\gamma}_{i.}$ og $\tilde{\gamma}_{i.}$ (dvs. deres "maker" i to-trinns utvelging) i noen særlig grad.

Når det gjelder stratifiserte utvalgsplaner, kan visse avvik forekomme i analysen, særlig hvis stratifikasjons-variabelen i stor grad "forklarer" forskjeller i frafalls-hyppigheter. I så fall vil "$\bar{\gamma}_{i.}$" - hvis den er en lineær funksjon av stratumgjennomsnittene - forbedres som estimator idet skjevheten reduseres.

Dersom stratifikasjons-variabelen ikke "forklarer" variasjoner i frafalls-hyppigheter, vil en formodentlig finne omtrent de samme resultater som her er beskrevet.

I avsnitt 3 og 4 er λ_j-ene behandlet som faste og kjente parametre. I praksis vil de naturligvis være ukjente og må derfor estimeres for at $\tilde{\gamma}_{i.}$ skal kunne anvendes. La

$$\tilde{\gamma}'_{i.} = \frac{1}{n} \sum_{j=1}^{H} \frac{\tilde{n}_{ij}}{1-\hat{\lambda}_j} \qquad (5.1)$$

der $\hat{\lambda}_j$ er en estimator for λ_j. Det antas at $\hat{\lambda}_j$ er forventningsrett og at $P[0 \leq \hat{\lambda}_j < 1] = 1$.

$\tilde{\gamma}'_{i.}$ vil ikke ha de samme egenskaper som $\tilde{\gamma}_{i.}$. Det er f.eks. grunn til å tro at $\tilde{\gamma}'_{i.}$ har større varians enn den opprinnelige estimatoren, - en formodning vi skal underbygge litt nærmere. Da det her vil føre for langt å finne forventning og varians til (5.1) for generelle estimatorer $\hat{\lambda}_j$ (som tilfredsstiller de to forutsetningene nevnt ovenfor), skal vi nøye oss med noe mindre stingente resonnementer (eller "begrunnede gjetninger").

La oss for enkelthets skyld sette

$$X = \frac{\tilde{n}_{ij}}{1-\lambda_j} \qquad \text{og} \qquad Y = \frac{1-\lambda_j}{1-\hat{\lambda}_j} = (1-\lambda_j) \sum_{x=0}^{\infty} \hat{\lambda}_j^x$$

slik at

$$\frac{\tilde{n}_{ij}}{1-\hat{\lambda}_j} = XY \qquad (5.2)$$

Forventningen til $\tilde{\gamma}'_{i.}$:

Generelt har vi dette uttrykket for forventningen til produktet av to stokastiske variable:

$$E_M(XY) = E_M X \cdot E_M + \text{cov}(X,Y) \qquad (5.3)$$

Dersom $\tilde{n}_{ij}$ og $\hat{\lambda}_j$ er basert på observasjoner fra samme utvalg, vil de vanligvis være negativt korrelerte. Dette innebærer at forventningen til (5.2) blir mindre når X og Y er negativt korrelerte enn når de er positivt korrelerte eller ukorrelerte. I sistnevnte tilfelle har vi:

$$E_M X = E_M \frac{\tilde{n}_{ij}}{1-\lambda_j} = n\gamma_{ij}$$

mens

$$E_M Y = (1-\lambda_j) E_M \sum_{x=0}^{\infty} \hat{\lambda}_j^x = 1 \quad \text{idet}$$

det av Hölders ulikhet ([5] s. 95) følger at

$$E_M \hat{\lambda}_j^x \geq [E_M \hat{\lambda}_j]^x = \lambda_j^x \quad \text{og}$$

dessuten $\frac{1}{1-\lambda_j} = \sum_{x=0}^{\infty} \lambda_j^x$.

Siden forventningen til en sum av variable er lik summen av forventningene, kan vi på grunnlag av ovenstående konkludere med følgende ulikheter (parentesene sikter til typen av "sammenheng" mellom $\tilde{n}_{ij}$ og $\hat{\lambda}_j$; $j = 1,2,\ldots,H$):

$$E_M \tilde{\gamma}'_{i.}(\text{pos.korr.}) \geq E_M \tilde{\gamma}'_{i.}(\text{ukorr.}) \geq E_M \tilde{\gamma}'_{i.}(\text{neg.korr.})$$

og (5.4)

$$E_M \tilde{\gamma}'_{i.}(\text{ukorr.}) \geq E_M \tilde{\gamma}_{i.}$$

Vi ser altså at ulikhetsrelasjon mellom $E_M \tilde{\gamma}_{i.}$ og $E_M \tilde{\gamma}'_{i.}$(neg.korr.) mangler, hvilket betyr at vi i tilfellet negativ korrelasjon mellom $\tilde{n}_{ij}$ og $\hat{\lambda}_j$; $j = 1,2,\ldots,H$, ikke (her) kan avgjøre hvorvidt $\tilde{\gamma}'_j$ over- eller underestimerer $\gamma_{i.}$.

Variansen til $\tilde{\gamma}_{i.}$:

La oss så se på variansen til $\tilde{\gamma}'_{i.}$:

$$\text{var}\, \tilde{\gamma}'_{i.} = \frac{1}{n^2} \sum_{j=1}^{H} \text{var} \frac{\tilde{n}_{ij}}{1-\hat{\lambda}_j} + \frac{2}{n^2} \sum_{k<j} \text{cov}\left(\frac{\tilde{n}_{ij}}{1-\hat{\lambda}_j}, \frac{\tilde{n}_{ik}}{1-\hat{\lambda}_k}\right) \qquad (5.5)$$

Vi skal her tillate oss noe mindre strenge krav til stringens og delvis basere oss på intuisjon for å underbygge følgende påstand (eller "begrunnede gjetning"):

$$\operatorname{var} \tilde{\gamma}'_{i.} \geq \operatorname{var} \tilde{\gamma}_{i.} \qquad (5.6)$$

Argumentasjonen for (5.6) er som følger:

(i) Det virker intuitivt rimelig at

$$\left|\operatorname{cov}\left(\frac{\tilde{n}_{ij}}{1-\hat{\lambda}_j}, \frac{\tilde{n}_{ik}}{1-\hat{\lambda}_k}\right)\right| \leq \left|\operatorname{cov}\left(\frac{\tilde{n}_{ij}}{1-\lambda_j}, \frac{\tilde{n}_{ik}}{1-\lambda_k}\right)\right|$$

"Rimeligheten" ligger i at leddene på venstre side inneholder flere stokastiske komponenter enn hva vi har på høyre side. Dette skulle medføre en noe svakere samvariasjon mellom leddene der $\hat{\lambda}_j$-ene inngår, enn mellom leddene der bare parametrene λ_j ($j = 1,2,\ldots,H$) forekommer.

I avsnitt 3 (s. 7) fant vi dessuten at

$$\operatorname{cov}\left(\frac{\tilde{n}_{ij}}{1-\lambda_j}, \frac{\tilde{n}_{ik}}{1-\lambda_k}\right) < 0$$

Kombineres dette med "intuisjonen" ovenfor, har vi dermed:

$$\operatorname{cov}\left(\frac{\tilde{n}_{ij}}{1-\hat{\lambda}_j}, \frac{\tilde{n}_{ik}}{1-\hat{\lambda}_k}\right) \geq \operatorname{cov}\left(\frac{\tilde{n}_{ij}}{1-\lambda_j}, \frac{\tilde{n}_{ik}}{1-\lambda_k}\right) \qquad (5.7)$$

(ii) For å drøfte variansleddene i (5.5), vil vi skrive $\tilde{n}_{ij}(1-\hat{\lambda}_j)^{-1}$ på formen (5.2). Variansen kan da uttrykkes slik:

$$\operatorname{var}(XY) = E_M\left[Y^2 \operatorname{var}(X|Y)\right] + \operatorname{var}\left[Y E_M(X|Y)\right] \qquad (5.8)$$

Vi vil nå anta at (X,Y) er binormalt fordelt[1)] med korrelasjon ρ. Da blir:

$$\operatorname{var}\left[X|Y\right] = \sigma_x^2(1-\rho^2) \quad \text{der} \quad \sigma_x^2 = \operatorname{var} X$$

og

$$E_M\left[X|Y\right] = E_M X + \frac{\sigma_x}{\sigma_y}\rho(Y-E_M Y) \quad \text{der} \quad \sigma_y^2 = \operatorname{var} Y .$$

---

1) Dette kan oppnås tilnærmet ved f.eks. å erstatte X med Xn^{-1} og Y med $Y\gamma m^{-1}$ der m er øvre summasjonsindeks i summen $Y \approx \sum_{x=0}^{m} \hat{\lambda}_j^x$ (m kan velges slik at tilnærmelsen får en ønsket nøyaktighetsgrad).

Settes disse uttrykkene inn i (5.8), fås etter noen enkle omforminger:

$$\text{var}(XY) = \sigma_x^2 E_M Y^2 + \sigma_y^2 (E_M X)^2 - \sigma_x^2 \sigma_y^2 \rho^2 \left[1 - \frac{\text{var} Y^2}{\sigma_y^4}\right] \qquad (5.9)$$

var Y^2 kan vi finne ved å utnytte det forhold at $(Y-E_M Y)^2 \sigma_y^{-2}$ er χ^2 - fordelt med 1 frihetsgrad. Da er

$$\text{var}\left(\frac{Y-E_M Y}{\sigma_y}\right)^2 = \frac{1}{\sigma_y^4} \text{var}\left[Y^2 - 2YE_M Y\right] = 2$$

hvorav fås

$$\text{var}\left[Y^2 - 2YE_M Y\right] = 2\sigma_y^4$$

Venstresiden i denne ligningen er:

$$\text{var}\left[Y^2 - 2YE_M Y\right] = \text{var}\, Y^2 + 4(E_M Y)^2 \sigma_y^2 - 2E_M Y \text{cov}(Y^2, Y)$$

Altså blir:

$$\text{var}\, Y^2 = 2\left[\sigma_y^4 - 2(E_M Y)^2 \sigma^2 + E_M Y \cdot \text{cov}(Y^2, Y)\right]$$

I dette uttrykket er:

$$\text{cov}(Y^2, Y) = E_M\left[(Y^2 - E_M Y^2)(Y - E_M Y)\right] =$$

$$= E_M\left[Y^2 - (E_M Y)^2 - \sigma_y^2\right]\left[Y - E_M Y\right] = E_M\left[Y^2 (Y - E_M Y)\right] =$$

$$= E_M (Y - E_M Y)^3 + 2E_M Y \cdot \sigma_y^2 = 2E_M Y \cdot \sigma_y^2$$

Dermed fås:

$$\text{var}\, Y^2 = 2\sigma_y^4$$

hvorav vi finner ved innsetting i (5.9):

$$\text{var}\,(XY) = \sigma_x^2 E_M Y^2 + \sigma_y^2 (E_M X)^2 + \sigma_x^2 \sigma_y^2 \rho^2$$

De to første leddene på høyre side uttrykker variansen til (X) når X og Y er ukorrelerte. Siden tillegget ved korrelasjon alltid blir positivt, innebærer avhengighet mellom X og Y at var (XY) blir større enn når de er uavhengige.

I det sistnevnte tilfellet finner vi at var(XY) blir (uten forutsetning om binormal fordeling):

$$\operatorname{var}\frac{\tilde{n}_{ij}}{1-\hat{\lambda}_j} = \sigma_x^2 E_M Y^2 + (E_M X)^2 \sigma_y^2 \geq \sigma_x^2 E_M Y^2 =$$

$$= E_M\left(\frac{1-\lambda_j}{1-\hat{\lambda}_j}\right)^2 \operatorname{var}\frac{\tilde{n}_{ij}}{1-\lambda_j} \geq \operatorname{var}\frac{\tilde{n}_{ij}}{1-\lambda_j} \qquad (5.10)$$

idet iflg. Hölders ulikhet:

$$E_M\left(\frac{1-\lambda_j}{1-\hat{\lambda}_j}\right)^2 \geq \left[E_M \frac{1-\lambda_j}{1-\hat{\lambda}_j}\right]^2 \geq (1-\lambda_j)^2 \left[\sum_{x=0}^{\infty} E_M \hat{\lambda}_j^x\right]^2 \geq$$

$$\geq (1-\lambda_j)^2 \left[\sum_{x=0}^{\infty} \lambda_j^x\right]^2 = 1 \ .$$

(iii) Av (5.7) og (5.10) følger ved innsetting i (5.5):

$$\operatorname{var} \tilde{\bar{Y}}'_{i.} \geq \frac{1}{n^2} \sum_{j=1}^{H} \operatorname{var} \frac{\tilde{n}_{ij}}{1-\lambda_j} + \frac{2}{n^2} \sum_{k<j} \operatorname{cov}\left(\frac{\tilde{n}_{ij}}{1-\lambda_j} , \frac{\tilde{n}_{ik}}{1-\lambda_k}\right) = \operatorname{var} \tilde{\bar{Y}}_{i.}$$

som er identisk med påstanden (5.6).

Den nye estimatorens ($\tilde{\bar{Y}}'_{i.}$'s) egenskaper, - så langt vi har skissert dem her, medfører at konklusjonene i avsnitt 4 må modifiseres en del i favør av estimatoren $\bar{Y}_{i.}$ for å bli gyldige i en sammenligning mellom $\bar{Y}_{i.}$ og $\tilde{\bar{Y}}'_{i.}$.

Estimeringen av λ_j-ene reiser nye problemer som krever inngående behandling. Som antydet i avsnitt 2 må en bl.a. legge stor vekt på valg av modell for hvordan λ avhenger av ulike variable. Et annet problem er spørsmålet om hvor mange frafallshomogene delpopulasjoner en skal velge (dvs. hvor stor H skal velges). Behandling av disse problemene vil imidlertid ikke bli gjort her.

LITTERATURLISTE

[1] Cochran, W.G.: "Sampling Techniques", J. Wiley & Sons, Inc., New York 1953.

[2] Dahl, G.: "Om frafallsproblemet ved statistiske undersøkelser. En drøfting av problemet spesielt i forbindelse med forbruksundersøkelsen". SSB, Kontoret for intervjuundersøkelser, 4/5-72, stensil.

[3] Elofsson, G.: "Teoretisk jämförelse mellan tre metoder att korrigera för bortfall". SCB, Sverige, Metodinformation, 15.9.72, stensil. (Senere publisert i Statistisk Tidskrift nr. 2 1973).

[4] Hoem, J.M.: "Variansberegninger ved intervjuundersøkelser. VI: Estimering av prosentfordelinger". Sosiodemografisk forskningsgruppe, SSB, J.M. Hoem/JD/GH, 1/9-72, stensil.

[5] Royden, H.L.: "Real Analysis", Macmillan, New York, 1963.

[6] SCB, Sverige: "Om bortfallsproblemet i utredningsinstitutets undersökningar". SCB, Utredningsinstitutet, Stockholm 1971, stensil.

[7] Vestbye, P.: "Metodisk vurdering av Forbruksundersøkelsen 1967". Arbeidsnotat IO 70/6. SSB, Oslo.

[8] Wahlström, S.: "Några synpunkter på svarsbortfallsproblemet vid stickprovsundersökningar av ändliga populationer". Lund Universitet, Sverige, 1968. Stensil.