Arbeidsnotater

STATISTISK SENTRALBYRÅ

Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20, 41 36 60

IO 73/6

14. februar 1973

METODEHEFTE NR. 2

Notater om hvor detaljerte resultater en kan publisere fra utvalgsundersøkelser

Innhold

Forord.

Metodehefter i serien Arbeidsnotater

I tilknytning til mange prosjekter i Statistisk Sentralbyrå utarbeides det mindre, upretensiøse notater for avklaring av spørsmål av metodisk interesse. Det kan dreie seg om utvalgsteknikk, alternative spørsmålsformuleringer, presentasjonsmetoder, begrepsavklaringer, diskusjon av "funn" i data, system-idéer, eller andre temaer. Selv om mange slike notater bare har begrenset interesse i ettertid, vil det blant dem være noen som kunne fortjene å bli mer alminnelig tilgjengelig enn de har vært hittil. Det kan også være nyttig å ha dem registrert sentralt slik at det blir lettere å få oversikt over det stoffet som foreligger, og å referere tilbake til det.

Byrået har innført en publiseringsordning for stoff av dette slaget. Etter forbilde av serien Technical Notes fra U.S. Bureau of the Census publiserer en leilighetsvis et passende antall slike notater samlet i metodehefter i serien Arbeidsnotater. Inneværende hefte er det andre av denne typen.

Forsker Jan M. Hoem er oppnevnt som redaktør av metodeheftene. Medarbeidere i Byrået som lager stoff som kan være aktuelt, bes sende dette til redaksjonen etter hvert som det blir ferdig.

Kontorlederne bes holde øynene åpne for denne nye publiseringsmuligheten.

Konsulent Bjørn L. Tønnesen er redaksjonssekretær.

Innledning til dette hefte.

Statistisk Sentralbyrå publiserer regelmessig resultater fra utvalgsundersøkelser for at de skal være tilgjengelige for alminnelige statistikkformål. Ved slik publisering må man ta samplingusikkerheten med i betraktningen. Enkelte numeriske resultater fra en utvalgsundersøkelse kan være så beheftet med utvalgsfeil at en lett blir villedet ved tolkningen av dem. Slike resultater bør naturligvis ikke offentliggjøres.

Ved rutinemessig publisering av utvalgsresultater trenger en enkle regler for bestemmelse av hvilke tall som er publiseringsverdige, og hvilke som ikke er det. Med det omfangsrike materialet som vanligvis samles inn i en undersøkelse, er det uoverkommelig å gi hvert enkelt delresultat en grundig skjønnsmessig vurdering. En trenger kriterier som lar seg gjennomføre nokså mekanisk. Disse kriteriene skal gardere mest mulig mot feiltolkning av de tall som presenteres, samtidig som det er viktig å unngå å holde tilbake publiseringsverdige resultater.

Det er ikke lett å utvikle tilfredsstillende regler av denne type, for det er mange stridende hensyn som skal ivaretas. Det man ber om, er i realiteten et enkelt og presist svar på et komplisert og diffust spørsmål.

Saken er at tallene jo kan anvendes for så mange formål og det er umulig å forutse dem alle. En publiseringsregel som kan virke som en god gardering mot én type feiltolking, trenger ikke gi noen beskyttelse mot en annen. Likeledes har brukerne av resultatene sterkt varierende kompetansenivå, slik at tall som uten betenkelighet kan overlates enkelte brukere, kan føre andre brukere helt på avveie. Det er derfor nødvendig å uteksperimentere regler som fremkommer som kompromissløsninger.

I en akutt situasjon der en rakst måtte ha en brukbar tommelfingerregel uten å kunne vente på langvarige utredninger og beregninger, skrev Brenna og Hoem et par notater som senere er blitt brukt en hel del i Byrået, og som har gitt støtet til begynnende forsøk på en utredning av disse spørsmål. Notatene gjengis først i dette metodeheftet, sammen med noen utfyllende beregninger ved Hoem og Ljones. Det følger så fire notater som viser hvordan idéene er blitt brukt i tilknytning til dataene fra arbeidskraftundersøkelsene. I det siste notatet i dette metodeheftet skisserer Thomsen så noen av de videre, spede forsøk som har vært gjort for å finne enkel publiseringsregler, og han oppsummerer dagens situasjon. En forventer at denne saken vil utvikles videre i tiden fremover, både ved at en bruker bedre variansmål, utnytter bedre den informasjon som finnes i dataene, og eksperimenterer mer med publiseringsregler.

ARBEIDSKRAFTSUNDERSØKELSENE BØR IKKE OFFENTLIGGJØRE OPPBLÅSTE TALL UNDER CA. 10 000

av

Svein Brenna og Jan M. Hoem

1. I de tabellene som planlegges offentliggjort på grunnlag av utvalgsundersøkelsene, gir man anslag for det antall personer i befolkningen som har gitte kjennetegn. Disse tallene er beregnet ved at tilsvarende tall i utvalget multipliseres med en passende "oppblåsningsfaktor".

Når man skal avgjøre hvilke resultater som kan publiseres, må man naturligvis ta samplingusikkerheten med i betraktningen. Samtidig er det ønskelig at man utvikler en enkel regel for bestemmelse av hvilke tall som er publiseringsverdige, og hvilke som ikke er det.

Vi vil da anbefale at man for arbeidskraftundersøkelsene ikke offentliggjør tall under ca. 10 000. I tabeller der så mange av de oppblåste tallene overstiger ca. 10 000 at tabellene bør offentliggjøres, bør tall under ca. 10 000 erstattes med et eget tegn, f.eks. kolon.

Dette er en enkel regel, og den er grunngitt i visse overlegninger over samplingfeil som vi presenterer nedenfor.

2. La N være antall personer i befolkningen, og la M av dem ha et gitt kjennetegn. La de tilsvarende antall i utvalget være n og X. Vår estimator for M er

$$(1) \qquad \hat{M} = \frac{N}{n} X,$$

der N/n er "oppblåsningsfaktoren". Et tilnærmet $(1-\varepsilon)$-konfidensintervall for M har grensene

$$(2) \qquad \hat{M} \pm N z \sqrt{\frac{1.5}{n} \frac{X}{n}\left(1 - \frac{X}{n}\right)},$$

der z er $(1-\frac{\varepsilon}{2})$-fraktilen i den normaliserte normalfordelingen.

I tabell 1 gjengir vi noen av tallene fra resultatene av arbeidskraftundersøkelsen i 4. kvartal* 1972 med $\pm$-leddet ovenfor (konfidensavviket) når $\varepsilon = 0.05$. Bortsett fra for sumtallene og tallet på menn som kontaktet arbeids-/sjømannskontoret utgjør konfidensavviket over 50 prosent av estimatet selv. For de minste estimatene

---

*) Det opprinnelige notatet inneholdt tall fra 3. kvartal 1971. Disse skal ikke publiseres, og de er her erstattet med tall fra 4. kvartal 1972.

utgjør konfidensavviket til og med mer enn estimatene selv.

Det har liten hensikt å publisere tall som er så dominert av samplingfeil.

Tabell 1. Estimert antall personer som søkte sysselsetting etter kjønn og måte å søke på. Tall fra arbeidskraftundersøkelsen 4. kvartal 1972*

Måte å søke på	Kjønn	Estimert antall	Konfidensavvik i prosent av estimat
Kontaktet arbeids-/sjømannskontor	M	11 818 ± 4 036	34,2
	K	4 178 ± 2 544	60,9
Svarte på annonse/annonserte selv	M	1 514 ± 1 533	101,3
	K	3 684 ± 2 391	64,9
Kontaktet arbeidsgiver	M	2 733 ± 2 060	75,4
	K	2 383 ± 1 923	80,7
Annen måte	M	657 ± 1 009	153,6
	K	591 ± 958	162,1
I alt, medregnet uoppgitt	M	18 245 ± 5 306	29,1
	K	12 952 ± 4 473	34,5
I ALT, MEDREGNET UOPPGITT		31 197 ± 6 921	22,2

Konfidenskoeffisient: 0,95.

3. La $\hat{p} = \hat{M}/N = X/n$. Den andel som konfidensavviket utgjør av estimatet $\hat{M}$ er

$$f(\hat{p}) = z\sqrt{1{,}5\,(\frac{1}{\hat{p}} - 1)/n}.$$

Hvis vi forlanger at denne høyst skal ha en verdi θ, må

$$p \geq \frac{1}{1+n\theta^2/(1{,}5\,z^2)} \approx \frac{1}{n}\,\frac{1{,}5\,z^2}{\theta^2},$$

Dette tilsvarer at

$$(3) \qquad M \geq \frac{N}{n} \cdot \frac{1{,}5\,z^2}{\theta^2}$$

x) Det opprinnelige notatet inneholdt tall fra 3. kvartal 1971. Disse skal ikke publiseres, og de er her erstattet med tall fra 4. kvartal 1972.

I arbeidskraftsundersøkelsen fra 4. kvartal[x] 1972 kan vi ta

$$\frac{N}{n} \approx \frac{2\ 732\ 000}{10\ 200} = 267.$$

Med $\varepsilon = 0.05$ (konfidensgrad 0.95) blir da (3)

$$\hat{M} \geq \frac{1\ 539}{\theta^2}.$$

I tabell 2 har vi gjengitt noen verdier av $1\ 539/\theta^2$. Vi ser at en θ på ca. 0.38 tilsvarer et minimalt estimat på ca. 10 000. For vår del er vi villige til å akseptere et konfidensavvik som utgjør ca. 38 % av estimatet, og vi fremsetter derfor den anbefaling som vi ga i punkt 1 ovenfor.

Tabell 2

θ	$1\ 440/\theta^2$
1/2	6 160
0,384	10 440
1/3	13 850

4. Konfidensintervallet i (2) er beregnet etter en empirisk formel som etter erfaringer i andre land har vist seg å gi rimelig god tilnærmelse for mange kjennetegn. Det er denne formel som Intervjukontoret har benyttet seg av for å antyde størrelsesordenen av samplingfeil ved de undersøkelser som er utført av kontoret.

Den utvalgsplan som Intervjukontoret bruker og som følgelig også er blitt nyttet for arbeidskraftundersøkelsene, bygger på trekning av utvalget i flere trinn. Den eksakte formel for beregning av samplingvariasjonen inneholder komponenter som gir uttrykk for det bidrag som hvert trinn gir til totalvariasjonen for estimatene. F.eks. for kjennetegn som er sterkt lokalt preget, kan dette bety at variasjonen for første trinn (variasjonen mellom primære utvalgsenheter) bidrar vesentlig til den totale samplingvariasjon. Under slike forhold kan man ikke vente at den empiriske formel som brukes av Intervjukontoret, gir særlig god tilnærmelse til den faktiske samplingvariasjon. Det er nok uforsvarlig i det lange løp utelukkende å bygge på denne tilnærmelsen. For å gi et fullgodt grunnlag for vurdering av resultatene

x) Se fotnote forrige side.

fra arbeidskraftsundersøkelsene for analyseformål bør Byrået etter vår oppfatning satse på mer omfattende beregninger av samplingvariasjonen.

Det må være et mål ved samplingfeilberegningnen å få anslått komponentene fra de enkelte trinn i utvalgsplanen. Dette vil dessuten gi grunnlag for vurdering av utvalgsplanen med sikte på forbedringer ved fremtidige revisjoner.

JMH/GH, 7/12-71.

Utdrag fra:

HVOR SMÅ BEFOLKNINGSGRUPPER KAN VI GI RIMELIG SIKRE TALL FOR I TABELLENE FRA ARBEIDSKRAFTSUNDERSØKELSENE?

KAN VI OVERHODET IKKE GI REGIONALE TALL?

av

Jan M. Hoem

1. Innledning

I notatet "Arbeidskraftsundersøkelser på utvalgsbasis" (SH/GH, 30/11-71) sies det på side 5:

> "En regner med å få brukbare anslag på nivå- og endringstall for grove nærings- og yrkesgrupperinger på landsbasis. Regionale tall vil ikke kunne presenteres på grunnlag av det undersøkte materialet."

Dette er vel i noenlunde overensstemmelse med de intuitive overlegninger vi har gjort i tidligere diskusjoner. Det representerer også et enkelt standpunkt som gir klare linjer, og som det er lett å gjennomføre i praksis hvis vi får Arbeidsdirektoratet til å akseptere det.

Bakgrunnen for at vi ikke vil gi talloppgaver for befolkningsgrupper under en viss størrelse, er naturligvis at samplingfeilen for en liten gruppe blir meget stor i forhold til antall personer i gruppen. Vi har imidlertid ikke tidligere (så vidt jeg kan huske) gjennomført beregninger over hvor små befolkningsgrupper vi synes vi kan tillate oss å gi tall for. Inneværende notat tar sikte på å gi et første sett av slike beregninger.

De anslag jeg her gir, må betraktes som sterkt approksimative. De bygger på to tilnærmelser, én tilnærmelse av en variansformel, og én tilnærmelse til normalfordelingen. Det er viktig at vi utfører nøyaktigere beregninger senere.

Beregningene tar utgangspunkt i én undersøkelse og utnytter ikke at vi egentlig opererer med en plan med roterende utvalg. Dette må også trekkes inn senere.

Med disse forbehold tyder beregningene allikevel på at det siterte standpunkt til muligheten av å gi anslag for visse regionale fordelinger, er litt for enkelt, men ikke mye. Det kan ikke være mer betenkelig å gi en regional fordeling av en viss befolkningsgruppe enn det er f.eks. å gi en alders-/kjønnsfordeling av samme befolkningsgruppe. De viktige spørsmål er i begge tilfeller

(i) hvor stor befolkningsgruppen er,

(ii) hvor fin fordelingen skal være, og

(iii) hvor stor samplingfeil en kan akseptere.

Istedenfor helt å avvise muligheten av regional fordeling av de ulike tallene, bør vi uteske Arbeidsdirektoratets (og andre brukeres) syn på spørsmålene (ii) og (iii) ovenfor. Nedenstående beregninger kan kanskje hjelpe oss både til å få et bedre begrep om det som er involvert, og til å få en kommunikasjonsprosess i gang.

(Dette notatet kan sees som et sidestykke til Notat JMH/SBr/GH, 7/12-71 fra Svein Brenna og Jan M. Hoem: "Arbeidskraftsundersøkelsene bør ikke offentliggjøre oppblåste tall under ca. 10 000".)

2. Litt elementær teori

La et utvalg bestå av n personer, trukket etter Intervjukontorets utvalgsplan fra en samlet bestand på N personer. La M personer i bestanden ha en viss egenskap - la oss si at de er sysselsatte kvinner - og la X av disse komme med i utvalget. La $p = M/N$. Vi vil anvende en tilnærmelse som er brukt en del, og vil regne som om X er normalfordelt med forventning np og varians $1.5\, np(1-p)$. [Dette innebærer i realiteten en tilnærmelse både av variansen og av sannsynlighetsfordelingen til X.] Hvis vi lar z være $(1-\frac{\varepsilon}{2})$-fraktilen i den standardiserte normalfordelingen, vil da X med en sannsynlighet på $1-\varepsilon$ ligge mellom grensene

$$np \pm z\sqrt{1.5\, np(1-p)}.$$

La vår estimator for M være

$$\hat{M} = \frac{N}{n} X.$$

Da vil, med samme tilnærmelse, $\hat{M}$ være normalfordelt med forventning M og varians

$$1.5\, N^2\, p(1-p)/n = 1.5\, \frac{N}{n}\,(1-p)\, M.$$

Med sannsynlighet $1-\varepsilon$ vil $\hat{M}$ ligge mellom grensene

$$(1) \qquad M \pm z\sqrt{1.5\, \frac{N}{n}(1-p)\, M}.$$

La

$$f(p) = \frac{z}{M}\sqrt{1.5\,\frac{N}{n}\,(1-p)M}$$

være den andel som avviket i grensene ovenfor utgjør av M. Enkel omforming gir

$$f(p) = z\sqrt{\frac{1.5}{n}\left(\frac{1}{p}-1\right)}.$$

Vi vil bruke $f(p) = f(\frac{M}{N})$ som mål på nøyaktigheten av anslaget $\hat{M}$ for M. Jo mindre f(p) er, desto bedre er nøyaktigheten. For fast n er f(p) en avtakende funksjon av p, slik at nøyaktighetsgraden er større jo større M er.

Hvis vi krever en nøyaktighetsgrad på minst 100 θ %, må M være så stor at

$$f(\tfrac{M}{N}) \leqq \theta.$$

Det tilsvarer kravet

$$(2) \qquad M \geqq \frac{N}{1+\theta^2\,\frac{n}{1.5\,z^2}} \approx \frac{N}{n}\,\frac{1.5\,z^2}{\theta^2}$$

Omvendt vil naturligvis en gitt M gi en nøyaktighet på

$$(3) \qquad \theta = z\sqrt{\frac{1.5}{n}\left(\frac{N}{M}-1\right)} \approx \frac{1}{\sqrt{M}}\sqrt{\frac{N}{n}\,1.5\,z^2}.$$

3. Talleksempel*

I arbeidskraftundersøkelsen i 4. kvartal 1972 er det omtrent N = 2 732 000 personer i bestanden (= personer mellom 16 og 74 år), og det er omtrent n = 10 150 personer i utvalget. Det gir "oppblåsningsfaktoren"

$$\frac{2\,732\,000}{10\,150} = 269.$$

(Eksakte tall: 2 732 345/10 153 = 269.1.) Med en ε = 0.05 blir høyre side i (2) da omtrent lik

$$1\,550 \;/\; \theta^2.$$

---

*) Det opprinnelige notatet inneholdt tall for 3. kvartal 1971. Denne undersøkelsen skal ikke publiseres, og tallene er derfor byttet ut med tall for 4. kvartal 1972.

(Eksakt: 1 550.66 / θ^2.) Avviket i (1) er $M\theta$, som altså blir minst lik 1 550 / θ. Vi gir noen verdier i tabell 1. I tabell 2 gir vi omvendt noen verdier for θ og 1 550 / θ tilsvarende gitte verdier av M.

Tabell 1

θ	1 550 / θ	1 550 / θ
1/2	6 200	3 100
1/3	13 950	4 650
1/4	24 800	6 200
1/5	38 750	7 750
1/10	155 000	15 500

Tabell 2

M	$\theta = f(\frac{M}{N})$	1 550 / θ
5 000	0.5568	2 784
8 000	0.4401	3 522
10 000	0.3938	3 938
20 000	0.2784	5 568
50 000	0.1761	8 802
100 000	0.1246	12 460
155 000	0.1000	15 500

JMH/OLj/WTD, 14/12-71

[Arbeidskraftundersøkelsene:]

1. Noen følger av å utvide utvalget med 50%.
2. Nøyaktighetsgraden når en benytter andre konfidensgrader.

av

Jan M. Hoem og Olav Ljones

1. Noen følger av å utvide utvalget med 50%.

Det kan være verd å gjøre seg opp en mening om hvilke konsekvenser det vil ha for samplingfeilen om man øker utvalget i en arbeidskraftsundersøkelse med 50%. Med utgangspunkt i tilnærmelsene brukt i to tidligere notater (JMH/GH, 7/12-71 og JMH/SBr/GH, 7/12-71) er det lett å få et grovt inntrykk av dette. I tabell 1 og 2 nedenfor gir vi noen talleksempler.

I tabell 1 viser vi hvor stor en befolkningsgruppe minst må være (min. M) for at nøyaktighetsgraden av vårt estimat $\hat{M}$ for M minst skal ha en verdi θ. (For definisjon av nøyaktighetsgraden, se side 3 i notatet JMH/GH, 7/12-71.)

I tabell 2 viser vi omvendt hvilken nøyaktighetsgrad θ en får når M har gitte verdier.

I begge tabeller har vi også gjengitt størrelsen av nøyaktighetsavviket $M\,\theta$.

Tallene i tabellene i notatet JMH/GH, 7/12-71 ble opprinnelig gitt med et antall sifre som er større enn det nøyaktighetsgraden av beregningene egentlig tilsier. I inneværende notat gjengis derfor avrundede tall.

En hovedkonklusjon i notatet JMH/GH, 7/12-71 var at vi bør kunne spesifisere sysselsatte og sysselsatte i arbeid på noe slikt som 30 klasser med den størrelsen utvalget nå har. Hvis utvalgsstørrelsen økes med 50%, bør vi tilsvarende kunne øke til noe rundt 50 klasser, forutsatt at antall personer ikke er for ujevnt fordelt over klassene.

Tabell 1.

	n = 10 800 (nåværende utvalgsstørrelse)		n = 16 200 (50% større utvalg)	
θ	min. M	θ·min M	min M	θ·min M
1/2	6 000	3 000	4 000	2 000
0.38	10 000	4 000	7 000	2 500
1/3	13 000	4 500	9 000	3 000
1/4	23 000	6 000	15 000	4 000
1/5	36 000	7 000	24 000	5 000
1/10	140 000	14 000	100 000	10 000

Konfidensgrad 0,95.

Tabell 2.

	n = 10 800 (nåværende utvalgsstørrelse)		n = 16 200 (50% større utvalg)	
M	θ	Mθ	θ	Mθ
5 000	0.54	2 700	0.44	2 200
8 000	0.42	3 400	0.35	2 800
10 000	0.38	3 800	0.31	3 100
20 000	0.27	5 400	0.22	4 400
50 000	0.17	8 500	0.14	7 000
100 000	0.12	12 000	0.10	10 000
150 000	0.10	15 000	0.08	12 000

Konfidensgrad 0,95.

2. Nøyaktighetsgraden når en benytter andre konfidensgrader.

I dette notatet og i andre notater (JMH/GH, 7/12-71 & JMH/SBr/GH, 7/1 7/12-71) som behandler usikkerheten i estimater i Arbeidskraftundersøkelsen har en holdt seg til konfidensgrad 0,95 ($\varepsilon = 0{,}05$). Det er mulig at en vil kunne akseptere en lavere konfidensgrad.

I det følgende gjengis noen beregninger når en benytter andre konfidensgrader.

Tabell 3. n = 10 800 (nåværende utvalgsstørrelse)

θ	Konfidensgrad 0,90 (ε = 0,1) min M	θ·min M	Konfidensgrad 0,80 (ε = 0,2) min M	θ·min M
1/2	4 000	2 000	2 500	1 250
1/3	9 000	3 000	6 000	2 000
1/4	16 000	4 000	10 000	2 500
1/5	25 000	5 000	15 000	3 000
1/10	100 000	10 000	60 000	6 000

Tabell 4. n = 16 200 (50% større utvalg)

	Konfidensgrad 0,90			Konfidensgrad 0,80	
θ	min M	θ min M	θ	min M	θ min M
1/2	3 000	1 500	1/2	1 700	800
1/3	6 000	2 000	1/3	4 000	1 200
1/4	11 000	2 700	1/4	7 000	1 600
1/5	17 000	3 400	1/5	10 000	2 000
1/10	67 000	6 700	1/10	40 000	4 000

Tabell 5. M = 10 000.

	Konfidensgrad 0,95 θ	Konfidensgrad 0,90 θ	Konfidensgrad 0,80 θ
n = 10 800	0,38	0,32	0,25
n = 16 200	0,31	0,26	0,20

ARBEIDSKRAFTUNDERSØKELSENE

Hvor mange klasser kan vi spesifisere i tabellene?

av Stein Østerlund Petersen

I det første settet med prøvetabeller fra arbeidskraftundersøkelsene hadde en for noen av tabellenes vedkommende delt gruppen sysselsatte opp i hele 80 forskjellige klasser. Det viste seg imidlertid at hvis hver klasse skulle inneholde minst 10 000 personer, ville en bare få spesifisert omtrent 30 av klassene. Tallet på sysselsatte var riktignok så stort (ca. 1 670 000) at det teoretisk skulle la seg gjøre å spesifisere godt over 100 klasser, men da størsteparten av de sysselsatte kunne henføres til et fåtall av dem, ville de fleste av klassene komme til å inneholde færre enn 10 000 personer. En kom derfor til den konklusjon at det i publiseringsøyemed ville ha liten hensikt å lage tabeller over de sysselsatte hvor antall klasser oversteg 30.

Senere tabellutkjøringer viser imidlertid at denne konklusjonen kanskje må vurderes på nytt. Blant annet gir en tabell over sysselsatte etter yrke, hvor yrke er spesifisert på 2-siffernivå (yrkesområde), nærmere 50 klasser som hver inneholder over 10 000 personer. En tabell over sysselsatte etter næring, hvor noen av næringene er spesifisert på 3-siffernivå (næringshovedgruppe) viser mellom 40 og 50 klasser med mer enn 10 000 sysselsatte. Det er også laget en tabell over sysselsatte etter kjønn og alder, med ett-årige aldersgrupper, hvor hele 86 klasser har tall større enn 10 000.

Det er derfor sannsynlig at regelen om maksimalt 30 klasser er for streng i enkelte tabeller. Spesielt later det til at antall klasser kan utvides til godt over 30 hvis det er få dimensjoner i tabellen, og bare en av dimensjonene inndeles i mange undergrupper.

BLT/EH, 19/9-72

ALDERSINNDELING I PROGNOSENE FOR TILBUDET AV ARBEIDSKRAFT

av Bjørn L. Tønnesen.

Innledning

Vi skal i dette notatet drøfte hvor fin aldersgruppering vi kan tillate i prognoseberegningene. Det vil bli presentert to metoder som begge viser at 6 000 yrkesaktive personer (oppblåste tall) kan være en passende nedre grense for størrelsen på gruppene, når vi bruker resultatene fra undersøkelsene 71 III og IV og 72 I og II. Vi har da brukt de samme krav til variansen på estimatorene som Hoem og Brenna setter i sitt notat JMH/SBr/GH, 7/12-71, der de konkluderer med at 10 000 bør være en nedre grense for oppblåste tall.

Notatet er kommet til etter samtaler mellom Hoem, Ljones, Tønnesen og Østerlund Petersen.

Utvalgsplan og symboler

Hvis en tenker seg at utvalget hvert kvartal kan deles i 4 puljer, kan utvalgsplanen for de fire undersøkelsene illustreres på følgende måte:

Pulje	71 III	71 IV	72 I	72 II
1	X	X	X	
2	X	X	X	
3	X	X		X
4	X	X		X
5			X	
6			X	X
7				X

Vi setter nå

$\frac{N_i}{n_i}$ = oppblåsingsfaktor for kvartal nr. i (i = 1, 2, 3, 4)

$\hat{Y}_i$ = oppblåst antall yrkesaktive i kvartal nr. i

X_{ij} = antall yrkesaktive i pulje nr. j og kvartal nr. i i utvalget
j = 1, ..., 7. (Mange X_{ij}, f.eks. X_{33}, er lik 0.)

Vi får:

$$Y_i = \frac{N_i}{n_i} X_i$$

Metode A

Hvis vi beregner gjennomsnittlig, oppblåst antall yrkesaktive i de 4 kvartalene, setter vi:

$$\hat{Y} = \frac{1}{4} \sum_{i=1}^{4} \hat{Y}_i = \frac{1}{4} \sum_{i=1}^{4} \frac{N_i}{n_i} X_i = \frac{N}{n} \cdot \frac{1}{4} \sum_{i=1}^{4} \sum_{j=1}^{7} X_{ij}$$

i det vi bruker samme oppblåsingsfaktor for hvert kvartal. La oss beregne var S, der vi setter

$$S = \sum_{i=1}^{4} \sum_{j=1}^{7} X_{ij}$$

og var $X_{ij} = \tau^2$ for alle i og j som er slik at $X_{ij} \neq 0$.

Vi ser at:

$$S = (X_{11} + X_{21} + X_{31}) + (X_{12} + X_{22} + X_{32}) + (X_{13} + X_{23} + X_{43}) + (X_{14} + X_{24} + X_{44}) + (X_{35}) + (X_{36} + X_{46}) + (X_{47})$$

Vi antar nå som en tilnærmelse at de variablene som står i samme parentes her, er like, mens de som står i ulike parenteser er uavhengige.

Vi forutsetter altså at intervjuing av den samme puljen flere ganger, gir nøyaktig samme informasjon hver gang, mens svarene vi får fra to forskjellige puljer er uavhengige.

Vi setter $X_{ij} = X_j$ for alle i og j der $X_{ij} \neq 0$, og vi får:

$$S = 3X_1 + 3X_2 + 3X_3 + 3X_4 + X_5 + 2X_6 + X_7$$

slik at

var $S = 42\tau^2$):

$$\text{var } \hat{Y} = \frac{N^2}{n^2} \cdot \frac{1}{16} 42\tau^2 = \frac{N^2}{n^2} \cdot 2{,}625\tau^2$$

Ser vi på kvartal i alene, får vi

$$\text{var } \hat{Y}_i = \frac{N^2}{n^2} 4\tau^2$$

og dermed blir

$$\frac{\text{var } \hat{Y}}{\text{var } \hat{Y}_i} = \frac{2{,}625}{4} = 0{,}656$$

Vi lar z være $(1 - \frac{0{,}05}{2})$ fraktilen i N(0,1). Hoem og Brenna foreslår at $\hat{Y}_i$ bare skal offentliggjøres derson

$$\theta = \frac{z \sqrt{\text{est var } \hat{Y}_i}}{\hat{Y}_i} > 0{,}38$$

Hvis vi setter samme krav til $\hat{Y}$, får vi

$$\theta = \frac{z \sqrt{\text{est var } \hat{Y}}}{\hat{Y}} > 0{,}38 \iff \theta = \frac{z \sqrt{0{,}66 \text{ est var } \hat{Y}_i}}{\hat{Y}} > 0{,}38$$

Her må vi få en brukbar tilnærmelse om vi erstatter $\hat{Y}$ med $\hat{Y}_i$, dvs.

$$\theta = \frac{z \sqrt{\text{est var } \hat{Y}_i}}{\hat{Y}_i} > \frac{0{,}38}{\sqrt{0{,}66}} \approx 0{,}47$$

En θ på 0,5 gir ifølge Hoem og Brenna, tabell 2, et minimalt oppblåst tall på 5 760. Vi skulle derfor være sikre om vi setter 6 000 som et minimum.

Metode B

Med de samme forutsetningene som er brukt under metode A, kunne vi tenke oss å erstatte S med S', der

$$S' = X_1 + \ldots + X_7$$

Denne estimeringsmetoden ville gå ut på å bruke hver pulje bare én gang, dvs. la de telle like mye uansett i hvor mange kvartal de er med. Hvis vi bryter de enkelte kvartalsundersøkelsene opp på denne måten, vil vi få et nytt stort utvalg. Størrelsen på dette utvalget vil bli:

$$7 \cdot \frac{10\ 800}{4} = 18\ 900$$

i det det er gjennomsnitlig $\frac{10\ 800}{4}$ personer i hver pulje. Med et utvalg på 18 900, får vi ifølge Hoem og Brenna at vi kan tillate å publisere oppblåste tall $\hat{Y}$, der

$$\hat{Y} > \frac{1\ 440}{\theta^2} \cdot \frac{10\ 800}{18\ 900} = 5\ 714 \text{ med } \theta = 0{,}38.$$

Igjen skulle vi ha vist at 6 000 er et brukbart tall, og samtidig har vi vel vist at de to estimeringsmetodene som er beskrevet under A og B er noenlunde like gode, slik at vi godt kan basere oss på metode A, som vi har tenkt å gjøre.

ARBEIDSKRAFTUNDERSØKELSENE

Om endringer i tallene fra en undersøkelse til en annen

Av Stein Østerlund Petersen

1. Innledning

I tidligere notater ([2],[3]) er det vist hvordan vi i en enkelt arbeidskraftundersøkelse kan beregne variansen til vårt estimat for antall personer i en bestemt befolkningsgruppe, og derved finne standardavvik og konfidensintervall. Dette notatet gir en tilsvarende forenklet formel for variansen til den estimerte endringen i antall personer i en bestemt befolkningsgruppe fra en undersøkelse til en annen. Notatet bygger på en omtale av endringer i prosenttall i [1].

2. Definisjoner

(Når i er brukt som fotskrift, er det innforstått at i antar verdiene 1 og 2.)

La U_i = undersøkelse foretatt på tidspunkt i

n_i = utvalgets størrelse i U_i

n_{12} = antall personer som er med i både U_1 og U_2

La G være den befolkningsgruppen som vi betrakter. G kan f.eks. være arbeidssøkere uten arbeidsinntekt. La videre

X_1 = antall av de n_1 personene i U_1 som hører til G

X_{11} = antall av de n_{12} personene som er med i både U_1 og U_2 og som hørte til G i U_1

X_{12} = antall av de n_1-n_{12} personene som er med i U_1, men ikke i U_2, og som hørte til G i U_1

X_2 = antall av de n_2 personene i U_2 som hører til G

X_{21} = antall av de n_{12} personene som var med i både U_1 og U_2 og som hørte til G både i U_1 og U_2

X_{22} = antall av de n_{12} personene som var med i både U_1 og U_2 og som hørte til G i U_2, men ikke i U_1

X_{23} = antall av de n_2-n_{12} personene som var med i U_2, men ikke i U_1, og som hørte til G i U_2

Vi har naturligvis at

$$X_1 = X_{11} + X_{12}$$

$$X_2 = X_{21} + X_{22} + X_{23}$$

La så begivenhenten A_i være gitt ved:

A_i = {en vilkårlig valgt person skal høre til G på tidspunkt i}

La $p_i = P\{A_i\}$

$p_{12} = P\{A_2/A_1\}$

$q_{12} = P\{A_2/\bar{A}_1\}$

Anta til slutt at vi har følgende befolkningstotaler:

N_i = antall personer i befolkningen på tidspunkt i

M_i = antall personer som hører til gruppen G på tidspunktet i

3. Estimatorer

De vanlige estimatorene for p_i og M_i, er

$$\hat{p}_i = \frac{X_i}{n_i}$$

$$\hat{M}_i = \frac{X_i}{n_i} \cdot N_i = \hat{p}_i \cdot N_i$$

Endringen i antall personer som hører til gruppen G fra tidspkt.1 til tidspkt vil vi uttrykke ved

$$E_{12} = M_2 - M_1$$

Som estimator for E_{12} vil vi bruke

$$\hat{E}_{12} = \hat{M}_2 - \hat{M}_1$$

Vi vil også angi en naturlig estimator for p_{12}:

$$\hat{p}_{12} = \frac{X_{21}}{X_{11}}$$

4. Varianser og covarianser

Vi er interessert i å finne et uttrykk for variansen til $\hat{E}_{12}$.

Vi har at

$$\begin{aligned} \operatorname{var} \hat{E}_{12} &= \operatorname{var} (\hat{M}_2 - \hat{M}_1) \\ &= \operatorname{var} \hat{M}_1 + \operatorname{var} \hat{M}_2 - 2 \operatorname{cov} (\hat{M}_1, \hat{M}_2) \\ &= N_1^2 \operatorname{var} \hat{p}_1 + N_2^2 \operatorname{var} \hat{p}_2 - 2N_1N_2 \operatorname{cov} (\hat{p}_1, \hat{p}_2) \end{aligned}$$

For var $\hat{p}_1$ og var $\hat{p}_2$ bruker vi den vanlige tilnærmelsesformelen:

$$\text{var } \hat{p}_i \approx 1.5 \frac{p_i(1-p_i)}{n_i}$$

Vårt problem blir da å finne cov $(\hat{p}_1, \hat{p}_2)$. Vi har at

$$\text{cov}(\hat{p}_1, \hat{p}_2) = E\hat{p}_1.\hat{p}_2 - E\hat{p}_1.E\hat{p}_2$$
$$= E\hat{p}_1.\hat{p}_2 - p_1p_2$$

Nå er

$$E\hat{p}_1\hat{p}_2 = \frac{1}{n_1.n_2} \cdot EX_1 \cdot X_2$$

$$= \frac{1}{n_1.n_2} E(X_{11} + X_{12})(X_{21} + X_{22} + X_{23})$$

$$= \frac{1}{n_1.n_2} E(X_{11}.X_{21}+X_{11}X_{22}+X_{11}X_{23}+X_{12}X_{21}+X_{12}X_{22}+X_{12}X_{23})$$

Her er X_{11} og X_{23} uavhengige fordi de er trukket fra to uavhengige utvalg. Tilsvarende er også X_{12} og X_{21}, X_{12} og X_{22} og X_{12} og X_{23} uavhengige. X_{11} og X_{21} og X_{11} og X_{22} er derimot avhengige.

Vi får da (under forutsetning av binomiske sannsynligheter):

$$EX_{11} = n_{12}\, p_1$$
$$EX_{11}^2 = \text{var } X_{11} + (EX_{11})^2 = n_{12}p_1(1-p_1) + n_{12}^2 \cdot p_1^2$$

$$EX_{11}.X_{21} = EE(X_{11}.X_{21}/X_{11}) = EX_{11}.p_{12}.X_{11}$$
$$= p_{12}.EX_{11}^2$$

$$EX_{11}.X_{22} = EE(X_{11}.X_{22}/X_{11}) = EX_{11}.(n_{12}-X_{11})q_{12}$$
$$= EX_{11}.n_{12}.q_{12} - EX_{11}^2.q_{12}$$

$$EX_{11}.X_{23} = EX_{11}.EX_{23} = n_{12}.p_1(n_2-n_{12}).p_2$$

$$EX_{12}.X_{21} = EX_{12}.EX_{21} = (n_1-n_{12}).p_1.n_{12}.p_{12}.p_1$$

$$EX_{12}.X_{22} = EX_{12}.EX_{22} = (n_1-n_{12})p_1.n_{12}.q_{12}(1-p_1)$$

$$EX_{12}.X_{23} = EX_{12}.EX_{23} = (n_1-n_{12})p_1.(n_2-n_{12}).p_2$$

Dette gir:

$$\text{est var } \hat{E}_{12} = 1.5 \left\{ \frac{\hat{M}_1(N_1-\hat{M}_1)}{n_1} + \frac{\hat{M}_2(N_2-\hat{M}_2)}{n_2} \right\} - 2\hat{M}_1 \cdot \frac{n_{12}}{n_1 n_2} (N_2\hat{p}_{12}-\hat{M}_2)$$

5. En forenklet formel for variansen

I arbeidskraftundersøkelsene kan vi gå ut fra at befolkningstotalene N_1 og N_2 vil være omtrent like store. Det samme gjelder for utvalgsstørrelsene n_1 og n_2. Vi kan da sette:

$$N_1 = N_2 = N$$
$$n_1 = n_2 = n$$

Antakelig vil det bli mest aktuelt å se på endringer fra kvartal til kvartal og fra år til år. På grunn av egenskapene ved rotasjonsplanen vil vi i begge disse tilfellene ha

$$n_{12} = \tfrac{1}{2}n$$

Lar vi så

$$\frac{\hat{M}_2-\hat{M}_1}{\hat{M}_1} = \alpha$$

får vi

$$\text{est var } \hat{E}_{12} = \frac{\hat{M}_1}{n} \{3N - 2\hat{M}_1(1+\alpha) + \tfrac{3}{2}\alpha\ (N-\alpha\hat{M}_1) - N\hat{p}_{12}\}$$

I arbeidskraftundersøkelsene har vi tilnærmet

$$N = 2\ 720\ 000$$
$$n = \quad 10\ 000$$

Ved å bruke dette, og samtidig sette

$$\frac{\hat{M}_1}{10\ 000} = m_1$$

får vi at

$$\text{est var } \hat{E}_{12} = \hat{M}_1 \{816-272\hat{p}_{12}+408\alpha-m_1(2+\alpha+1.5\alpha^2)\}$$

6. Noen resultater

Som illustrasjon til punktene foran skal vi beregne standardavviket til noen av endringene fra 2. kvartal 1972 til 3. kvartal 1972. Fotskrift 1 betegner altså 2. kvartal 1972 og fotskrift 2 3. kvartal 1972. Vi vil basere oss på den forenklede formelen i pkt. 5.

a) Arbeidssøkere uten arbeidsinntekt

Vi har at

$$\hat{M}_1 = 26\ 104 \text{ og } \hat{M}_2 = 30\ 624$$

som gir

$$\alpha = 0.147$$

Videre er

$$\hat{p}_{12} = 4/58 = 0.069$$

som gir

$$\sqrt{\text{est var } \hat{E}_{12}} = 4\ 700$$

Et 95-prosent konfidensintervall for endringen blir da

$$4\ 500 \pm 9\ 200$$

eller $<-4\ 700, 13\ 700>$

80-prosent konfidensintervall:

$<-1\ 500, 10\ 500>$

b) Sysselsatte

$$\hat{M}_1 = 1\ 661\ 792, \quad \hat{M}_2 = 1\ 643\ 885$$

$$\alpha = -0.011$$

$$\hat{p}_{12} = 2\ 724/3\ 017 = 0.903$$

$$\sqrt{\text{est var } \hat{E}_{12}} = 19\ 700$$

95-prosent konfidensintervall

$$-17\ 900 \pm 38\ 600$$

eller

$<-56\ 500, 20\ 700>$

80-prosent konfidensintervall

$<-43\ 100, 7\ 300>$

c) Sysselsatte i bergverksdrift, industri, kraft- og vannforsyning

$\hat{M}_1 = 433\ 966$, $\hat{M}_2 = 420\ 937$

$\alpha = -0.031$

$\hat{P}_{12} = 679/782 = 0.868$

$\sqrt{\text{est var } \hat{E}_{12}} = 14\ 400$

95-prosent konfidensintervall

$-13\ 000 \pm 28\ 200$

eller

$<-41\ 200,\ 15\ 200>$

80-prosent konfidensintervall

$<-31\ 400,\ 5\ 400>$

Konfidensintervallene er beregnet under forutsetning av normalfordeling.

Som eksemplene ovenfor viser, må vi være ekstra forsiktige når vi skal kommentere endringstall. Vi må hele tiden være klar over at vi på grunnlag av en observert endring i vårt tallmateriale kun kan angi et intervall som med en viss sannsynlighet dekker den reelle endringen. I noen tilfeller vil dette intervallet antyde retningen av denne endringen, men intervallet vil også ofte være slik at vi ikke kan si noe sikkert om vi i virkeligheten har hatt en økning eller en nedgang i antall personer med et gitt kjennetegn.

7. Referanser

[1] "Noen tekniske problemer i forbindelse med arbeidskraftstellingene". Notat IT/Sin, 15/12-70.

[2] "Hvor små befolkningsgrupper kan vi gi rimelig sikre tall for i tabellene fra arbeidskraftsundersøkelsene?" Notat JMH/GH, 7/12-71.

[3] "Arbeidskraftsundersøkelsene bør ikke offentliggjøre oppblåste tall under ca. 10 000". Notat JMH/SBr/GH, 7/12-71.

SØP/IH, 30/11-72

ARBEIDSKRAFTUNDERSØKELSENE

Variansen til gjennomsnittstall

av Stein Østerlund Petersen

1. Innledning

I et tidligere notat ([3]) har vi behandlet usikkerheten ved endringstall, uttrykt ved variansen til den estimerte endringen. På grunnlag av formlene vi utviklet der, kan vi forholdsvis lett også finne variansen til gjennomsnittstall beregnet på grunnlag av 2 eller flere undersøkelser. Da det i forbindelse med en årspublikasjon kan bli aktuelt å publisere årsgjennomsnitt, kan det være gunstig å ha en slik variansformel å støtte seg til når vi skal vurdere hvor små gjennomsnittstall vi skal offentliggjøre.

Variansen til gjennomsnittstall er forøvrig også behandlet i ([2]).

2. Betegnelser

Vi vil bruke de samme betegnelsene som i [3]. Vi skal imidlertid ikke begrense oss til 2 undersøkelser, men anta at gjennomsnittstall beregnes på grunnlag av s undersøkelser. Som fotskrifter vil vi bruke i og j, som begge kan anta verdiene 1, 2, ... , s.

3. En naturlig estimator for gjennomsnittet

Det vi skal estimere er den gjennomsnittlige verdi av en bestemt størrelse i et bestemt tidsrom. La M betegne dette gjennomsnittet. Hvis vi antar at vi i det aktuelle tidsrommet har gjennomført s undersøkelser noenlunde jevnt fordelt over tidsrommet, vil en naturlig estimator for M være

$$\hat{M} = \frac{1}{s} \sum_{i=1}^{s} \hat{M}_i$$

4. Variansen til $\hat{M}$

Ved å bruke regelen for variansen til en sum får vi at

$$\operatorname{var} \hat{M} = \operatorname{var} \frac{1}{s} \sum_{i=1}^{s} \hat{M}_i$$

$$= \frac{1}{s^2} \{ \sum_{i=1}^{s} \text{var } \hat{M}_i + 2 \sum_{i<j} \text{cov } (\hat{M}_i, \hat{M}_j) \}$$

Av avsnitt 4 i [3] følger umiddelbart at

$$\text{cov } (\hat{M}_i, \hat{M}_j) = N_i N_j \frac{n_{ij}}{n_i n_j} p_i \, (p_{ij} - p_j)$$

Når vi da for var $\hat{M}_i$ bruker den vanlige formelen

$$\text{var } \hat{M}_i = 1{,}5 \frac{p_i \, (1-p_i)}{n_i} N_i^2$$

blir

$$\text{var } \hat{M} = \frac{1}{s^2} \{ 1{,}5 \sum_{i=1}^{s} N_1^2 \frac{p_i \, (1-p_i)}{n_i} + 2 \sum_{i<j} N_i N_j \frac{n_{ij}}{n_i n_j} p_i \, (p_{ij} - p_j) \}$$

5. Estimering av variansen

Som estimator for var $\hat{M}$ vil vi bruke uttrykket ovenfor innsatt $\hat{p}_i$, $\hat{p}_j$ og $\hat{p}_{ij}$ for henholdsvis p_i, p_j og p_{ij}. Dette gir

$$\text{est var } \hat{M} = \frac{1}{s^2} \{ 1{,}5 \sum_{i=1}^{s} N_1^2 \frac{\hat{p}_i \, (1-\hat{p}_i)}{n_i} + 2 \sum_{i<j} N_i N_j \frac{n_{ij}}{n_i n_j} \hat{p}_i \, (\hat{p}_{ij} - \hat{p}_j) \}$$

$$= \frac{1}{s^2} \{ 1{,}5 \sum_{i=1}^{s} \frac{\hat{M}_i \, (N_i - \hat{M}_i)}{n_i} + 2 \sum_{i<j} \hat{M}_i \frac{n_{ij}}{n_i n_j} (N_j \cdot \hat{p}_{ij} - \hat{M}_j) \}$$

6. Noen forenklinger

Vi vil anta at

$$N_i = N = 2\,720\,000$$

$$n_i = n = \quad 10\,000$$

For arbeidskraftundersøkelsene er disse forutsetningene ikke urimelige. Vi setter videre

$$\frac{\hat{M}_i}{10\,000} = m_i$$

Etter noe regning kommer vi da fram til følgende uttrykk:

$$\text{est var } \hat{M} = \frac{408}{s} \hat{M} - \frac{1{,}5}{s^2} \sum_{i=1}^{s} m_i \hat{M}_i + \sum_{i<j} m_i n_{ij} \left(\frac{544}{s^2} \hat{p}_{ij} - \frac{2}{s^2} m_i \right) \}$$

Uttrykket kan muligens forenkles ytterligere.

7. Hvor små gjennomsnittstall kan vi publisere?

Hvis vi bruker samme krav til nøyaktighet som i [1], (og som ledet oss til den konklusjon at for en enkelt undersøkelse bør vi ikke publisere tall under 10 000), må $\hat{M}$ være så stor at

$$\frac{1{,}96\sqrt{\text{est var } \hat{M}}}{\hat{M}} < 0{,}38$$

for at vi skal kunne offentliggjøre $\hat{M}$.

Innsetting av uttrykket for est var $\hat{M}$ i ulikheten ovenfor gir:

$$\frac{1{,}96\sqrt{\frac{408}{s}\hat{M} - \frac{1{,}5}{s^2}\sum_{i=1}^{s} m_i\hat{M}_i + \sum_{i<j} m_i n_{ij}\left(\frac{544}{s^2}\hat{p}_{ij} - \frac{2}{s^2}m_i\right)}}{\hat{M}} < 0{,}38$$

For å forenkle regningen vil vi anta at alle $\hat{M}_i$ er like store ($=\hat{M}$). Selv om dette bare unntaksvis vil være oppfylt, vil vi allikevel få et inntrykk av hvor stor $\hat{M}$ må være. Vi vil videre sette $\hat{p}_{ij} = 1$ for alle $i < j$. Dette er den minst gunstige verdi av $\hat{p}_{ij}$.

Etter noe regning (hvor vi utelater ledd som bare i liten grad påvirker størrelsen av uttrykket under rottegnet ovenfor) kommer vi fram til ulikheten

$$\hat{M} > \frac{10\,850}{s} + \frac{1{,}447}{s^2}\sum_{i<j} n_{ij}$$

Det første leddet på høyre side av ulikhetstegnet sier hvor stor $\hat{M}$ måtte ha vært hvis utvalgene i de s undersøkelsene var uavhengige. Det andre leddet sier hvor mye vi må øke denne verdien fordi noen personer er med i mer enn en av undersøkelsene.

8. Noen talleksempler

a) For de fire undersøkelsene i 1972 gjelder:

$s = 4$

$\sum_{i<j} n_{ij} = 17\,500$

og $\hat{M} > 4\,296$

Det virker som om 5 000 vil være en rimelig nedre grense for årsgjennomsnitt for 1972.

b) Til bruk for Langtidsprogrammet 1974-77 ble det laget noen tabeller basert på gjennomsnittstall fra de to arbeidskraftundersøkelsene i 1971 og de to første i 1972. I [2] kom en fram til at 6 000 burde være

nedre grense for hvor små tall en kunne gi i disse tabellene. Vi skal se om ulikheten i pkt. 7 leder til samme konklusjon. Vi har da

$$s = 4$$

$$\sum_{i<j} n_{ij} \approx 35\,000$$

og $\hat{M} > 5\,880$

Også her finner vi at det er naturlig å sette 6 000 som nedre grense.

c) Hvis det blir 6 arbeidskraftundersøkelser i 1973 får vi

$$s = 6$$

$$\sum_{i<j} n_{ij} \approx 25\,000$$

og $\hat{M} > 2\,813$

Vi vil da kunne tillate oss å offentliggjøre årsgjennomsnitt av størrelsesorden 3 000 med samme nøyaktighet som et tall på 10 000 fra en enkelt undersøkelse.

9. Referanser

[1] "Hvor små befolkningsgrupper kan vi gi rimelig sikre tall for i arbeidskraftundersøkelsene?" Notat JMH/GH, 7/12-71.

[2] "Arbeidsrapport om aldersinndeling i prognosene for tilbudet av arbeidskraft." Notat BLT/EH, 19/9-72.

[3] "Om endringer i tallene fra en undersøkelse til en annen." Notat SØP/IH, 23/11-72.

SØP/IH, 14/12-72

ARBEIDSKRAFTUNDERSØKELSENE

Et forsøk på å finne variansen til gjennomsnittlig arbeidstid pr. uke

av Stein Østerlund Petersen.

1. Innledning

I tabellene fra AKU har vi hittil gitt gjennomsnittlig arbeidstid pr. uke i hele timer. Det har imidlertid vært antydet at vi kanskje burde gi disse tallene med én desimal, for derved å få et mer fintfølende instrument til oppfanging av konjunktursvingninger. Problemet som da melder seg, er om tallene våre er så presise at det i det hele tatt har noen hensikt å bruke desimaler.

I dette notatet er det gjort et forsøk på å si noe om usikkerheten som knytter seg til estimatene for gjennomsnittlig arbeidstid, slik at vi bedre skal kunne vurdere spørsmålet om desimaltall.

2. Betegnelser

La N være antall personer i befolkningen i alder 16-74 år, og la n være det tilsvarende antall personer i utvalget. La videre

M = antall personer innen en bestemt gruppe av sysselsatte i inntektsgivende arbeid

M_i = antall av de M personene som arbeidet i timer i undersøkelsesuken; $i = 1, 2, \ldots, 99$

S_M = totalt antall arbeidstimer utført av de M personene i undersøkelsesuken

T_M = gjennomsnittlig arbeidstid i undersøkelsesuken for de M personene

Vi har da følgende sammenhenger:

$$M = \sum_i M_i$$

$$S_M = \sum_i i M_i$$

$$T_M = \frac{S_M}{M}$$

La $\hat{M}$, $\hat{M}_i$, $\hat{S}_M$ og $\hat{T}_M$ være våre estimatorer for henholdsvis M, M_i, S_M og T_M. Her er $\hat{M}$ og $\hat{M}_i$ de vanlige estimatorene, og

$$\hat{S}_M = \sum_i i\hat{M}_i \quad ; \quad \hat{T}_M = \hat{S}_M/\hat{M}$$

3. Et hjelperesultat

Anta at en befolkningsgruppe G splittes i k undergrupper $G_1, \ldots, G_k$; og anta at vi i AKU finner at antall personer i G er $\hat{Y}$, mens det er $\hat{Y}_1, \ldots, \hat{Y}_k$ i undergruppene. La G_i og G_j være to vilkårlige undergrupper. Vi vil da trenge å kjenne kovariansen mellom $\hat{Y}_i$ og $\hat{Y}_j$. Vi har at

$$\text{var}\,(\hat{Y}_i + \hat{Y}_j) = \text{var}\,\hat{Y}_i + \text{var}\,\hat{Y}_j + 2\,\text{cov}\,(\hat{Y}_i, \hat{Y}_j)$$

Nå er

$$\text{var}\,(\hat{Y}_i + \hat{Y}_j) = \frac{1,5}{n}(Y_i + Y_j)\{N - (Y_i + Y_j)\}$$

$$\text{var}\,\hat{Y}_i + \text{var}\,\hat{Y}_j = \frac{1,5}{n}\{Y_i(N - Y_i) + Y_j(N - Y_j)\}$$

slik at

$$\text{cov}\,(\hat{Y}_i, \hat{Y}_j) = -\frac{1,5}{n} Y_i Y_j$$

Y_i og Y_j er estimandene.

4. Variansen til $\hat{T}_M$

Vi vil anta at $\hat{M}$ er tilnærmet normalfordelt og forventningsrett. Da er også $\hat{S}_M$ tilnærmet normalfordelt og forventningsrett.

Hvis var $\hat{M}$ og var $\hat{S}_M$ begge er små, kan (se [1] , s. 143-144) var $\hat{T}_M$ tilnærmet settes lik

$$\text{Var}\,\hat{T}_M \approx \frac{1}{M^2}\,\text{var}\,\hat{S}_M + \frac{S_M^2}{M^4}\,\text{var}\,\hat{M} - \frac{2\,S_M}{M^3}\,\text{cov}\,(\hat{S}_M, \hat{M})$$

Vi vil anta at var $\hat{M}$ og var $\hat{S}_M$ er så små at tilnærmelsen er brukbar. Muligens er dette en tilsnikelse.

Vi må da finne uttrykk for var $\hat{S}_M$ og cov $(\hat{S}_M, \hat{M})$. For var $\hat{M}$ har vi jo direkte

$$\operatorname{var} \hat{M} = \frac{1,5}{n} M(N - M)$$

Vi har videre:

$$\operatorname{var} \hat{S}_M = \operatorname{var} \sum_i i\hat{M}_i$$

$$= \sum_i i^2 \operatorname{var} \hat{M}_i + 2 \sum_{i<j} ij \operatorname{cov} (\hat{M}_i, \hat{M}_j)$$

$$= \frac{1,5}{n} \{N \sum_i i^2 M_i - S_M^2\}$$

$$\operatorname{cov} (\hat{S}_M, \hat{M}) = E\hat{S}_M\hat{M} - S_M M$$

Nå er

$$E\hat{S}_M\hat{M} = E \{(\sum_i i\hat{M}_i)(\sum_i \hat{M}_i)\}$$

$$= \sum_i i\, E\hat{M}_i^2 + \sum_{i \neq j} Ei\hat{M}_i\hat{M}_j$$

$$E\hat{M}_i^2 = \operatorname{var} \hat{M}_i + (E\hat{M}_i)^2$$

$$= \frac{1,5}{n} NM_i + (1 - \frac{1,5}{n})M_i^2$$

$$E\hat{M}_i\hat{M}_j = \operatorname{cov} (\hat{M}_i, \hat{M}_j) + E\hat{M}_i E\hat{M}_j$$

$$= (1 - \frac{1,5}{n})M_i M_j$$

n vil være så stor at $(1 - \frac{1,5}{n}) \approx 1$. Innsetting i uttrykket for $\operatorname{cov} (\hat{S}_M, \hat{M})$ gir da:

$$\operatorname{cov} (\hat{S}_M, \hat{M}) = \frac{1,5}{n} NS_M + \sum_i i M_i^2 + \sum_{i \neq j} i M_i M_j - S_M M$$

$$= \frac{1,5}{n} NS_M$$

Variansen til $\hat{T}_M$ blir da tilnærmet lik

$$\operatorname{var} \hat{T}_M \approx \frac{1,5}{M^2} \frac{N}{n} \sum_i i^2 M_i - T_M^2 \{2 \frac{1,5}{n} + \frac{N}{n} \frac{1,5}{M}\}$$

Det første leddet i parentesen er tilnærmet lik 0, slik at vi til slutt

får

$$\text{var } \hat{T}_M \approx \frac{1,5}{M} \frac{N}{n} \{ \frac{1}{M} \sum_i i^2 M_i - T_M^2 \}$$

Som estimator for var $\hat{T}_M$ vil vi bruke uttrykket for var $\hat{T}_M$ innsatt $\hat{M}$, $\hat{M}_i$ og $\hat{T}_M$ for henholdsvis M, M_i og T_M. Dette gir

$$\text{est var } \hat{T}_M = \frac{1,5}{\hat{M}} \frac{N}{n} \{ \frac{1}{\hat{M}} \sum_i i^2 \hat{M}_i - \hat{T}_M^2 \}$$

5. To eksempler

Eksempel 1: Anta at

$\frac{N}{n} = 250$

$\hat{T}_M = 40$

$\hat{M} = 50\ 000$

$\hat{T}_M$ kan f.eks. tolkes som gjennomsnittlig arbeidstid for menn i alder 65-69 år. Anta for enkelhets skyld at

$\hat{M}_{20} = 4\ 000$

$\hat{M}_{40} = 30\ 000$

$\hat{M}_{45} = 16\ 000$

$\hat{M}_i = 0$ for $i \neq 20, 40, 45$

Da blir

$$\sqrt{\text{est var } T_M} = 3,5$$

Hvis vi antar normalfordeling blir et 95-prosent konfidensintervall for T_M lik

<33,47>

I dette tilfellet ville det ha liten hensikt å angi $\hat{T}_M$ med én desimal.

Eksempel 2: Vi vil igjen anta at $\frac{N}{n} = 250$ og $\hat{T}_M = 40$, men vi vil la $\hat{M} = 400\ 000$

$\hat{T}_M$ kan f.eks. tolkes som gjennomsnittlig arbeidstid i industrien. Vi vil for enkelhets skyld anta at

$$\hat{M}_{30} = 50\ 000$$

$$\hat{M}_{40} = 250\ 000$$

$$\hat{M}_{45} = 100\ 000$$

$$\hat{M}_i = 0 \text{ for } i \neq 30, 40, 45$$

Da blir

$$\sqrt{\text{est var } \hat{T}_M} = 0{,}13$$

og et 95-prosent konfidensintervall for T_M blir

$$\langle 39{,}7,\ 40{,}3 \rangle$$

Det ville her ikke være urimelig å angi $\hat{T}_M$ med én desimal.

6. Konklusjon

Uttrykket for est var $\hat{T}_M$ i pkt. 4 bygger på forutsetninger som muligens ikke er tilfredsstillende oppfylt i den foreliggende situasjon. Beregningene i pkt. 5 må sees i lys av dette. Vi kan kanskje allikevel tillate oss å trekke den konklusjon at vi for de aller største gruppene (totalt antall sysselsatte, sysselsatte menn, sysselsatte i industrien etc.) ikke uten videre bør avvise forslaget om å gi gjennomsnittlig arbeidstid pr. uke med én desimal. Inntil vi har foretatt flere og mer presise beregninger enn det som er gjort her, bør vi vel imidlertid fortsette å gi gjennomsnittlig arbeidstid i hele antall timer, også for de største gruppene.

7. Referanse

[1] Erling Sverdrup: "Lov og tilfeldighet", bind 1. Universitetsforlaget 1964.

HVOR OPPDELT KAN EN OFFENTLIGGJØRE RESULTATENE FRA EN INTERVJUUNDERSØKELSE?

Av

Ib Thomsen.

1. Innledning

Under planleggingen av en undersøkelse bestemmes utvalgsstørrelsen som regel slik at visse høyt prioriterte gjennomsnitt for hele populasjonen blir estimert med en ønsket nøyaktighet. Ved presentasjonen av resultatene fra en undersøkelse ønsker en naturlig nok ofte å spalte opp utvalget etter visse kjennetegn, og gi gjennomsnitt for så små befolkningsgrupper som mulig. På den andre siden ønsker en ikke å spalte opp materialet i en slik grad at de publiserte tallene er beheftet med for stor usikkerhet. Da kravet til nøyaktigheten kan variere fra leser til leser, ønsker en kanskje å publisere tall, som en vet har mindre god kvalitet, men en vil da i tabellverket markere at disse tall må "tas med en klype salt". (Se f.eks. publisering av resultatene fra EF-undersøkelsen i Statistisk ukehefte nr. 44/72.) Hensikten med dette notat er å vise hvor vanskelig det er å finne fram til en standard for å avgjøre hvor små befolkningsgrupper en kan publisere tall for. Den vesentligste årsaken til dette er at vi vet lite om hvorledes tallene blir brukt, og dermed lite om den nøyaktighetsgrad som skal til for å unngå mistolking av resultatene.

2. Valg av kvalitetsmål

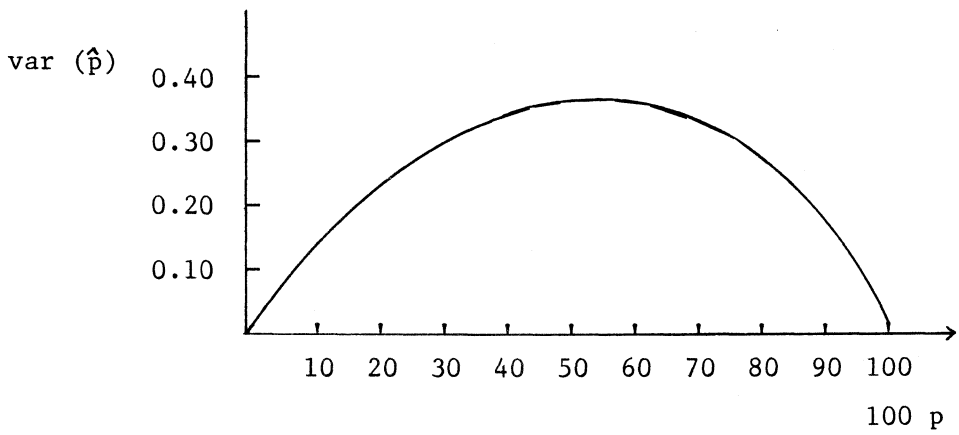
For å avgjøre om en estimator har en ønsket kvalitet må vi først velge et kvalitetsmål. Jeg foreslår å bruke samplingvariansen. Dette kan kritiseres fordi systematiske feil også bør komme i betraktning. Årsakene til at jeg likevel velger samplingvariansen, er at vi vet for lite om de systematiske feil, og dessuten er det rimelig å anta at de systematiske feil utgjør en relativ liten andel av totalvariansen for et gjennomsnitt innen en "liten" befolkningsgruppe.

3. Bruk av variansen

Anta at vi ønsker å estimere hyppigheter. La den ukjente hyppighet vi ønsker å estimere, være p. For Byråets utvalgsplan vil det da alltid i prinsippet være mulig å anslå variansen på estimatoren ut fra utvalget. Da vi imidlertid ennå ikke har fått et standardprogram for beregning av variansen, bruker vi følgende tilnærmelse:

(1) $\quad \text{var}\ (\hat{p}) = 1.5\ \frac{p(1-p)}{n}$

hvor n er utvalgsstørrelsen. Nedenfor er variansen gitt som funksjon av p for n=1.



Figur 1. var (p).

En ser det kanskje noe overraskende at variansen er størst når $p = \frac{1}{2}$.

Det følger nå av (1) at en for hver verdi av p kan bestemme den minste utvalgsstørrelse n_0 som er slik at var $(\hat{p})$ er mindre enn en vilkårlig gitt verdi k når $n \geq n_0$.

I Fig. 3 er n_0 tegnet inn som funksjon av p for forskjellige verdie av variansen. (Kurvene er symmetriske om 0.50 og er derfor bare gitt for $0 \leq p \leq 0.50$.) Vi skal først gi et eksempel på hvorledes en kan bruke resultatene i Fig. 3.

Hvis en f.eks. har n = 150 og p = 0.20, kan en avlese variansen til å være 0.0016. Herav følger at standardavviket er 0.04, hvilket vil si at et 0.95 konfidensintervall blir av lengde 0.16.

Av større interesse i denne sammenheng er det at selv om en vet hvilken nøyaktighet en ønsker, vil n_0 være avhengig av p. Hvis en ønsker en varians på 0.0016, trenger en 53 observasjoner hvis p er 0.06, men 234 observasjoner hvis p er 0.50. Det enkle spørsmål stilt i tittelen har altså ikke noe enkelt svar selv når en vet hvilken nøyaktighet en ønsker. Hvis en i tillegg ikke vet hvilken nøyaktighet en ønsker, viser kurvene i Fig. 3 at det er umulig å si hvor små befolkningsgrupper en kan gi tall for idet nødvendig utvalgsstørrelse varierer med en faktor på nesten 10 når ønsket varians går fra 0.0025 til 0.000225.

Når en er opptatt av et bestemt problem i en undersøkelse, kan en ofte si noe om den nøyaktighet som ønskes for det bestemte formål. Når det derimot er tale om å finne en standard for publisering av en hel undersøkelse, er det vanskelig å si noe om den usikkerhet en bør tillate på hvert av de publiserte tallene, hvilket gjør det nærmest umulig å avgjøre generelt hvor små befolkningsgrupper en bør publisere tall for.

4. Eksempel på en enkel regel

Jeg skal likevel gi et eksempel på en enkel regel: Hvis $0.10 \leq p \leq 0.90$ forlanger jeg et utvalg av størrelse 150 eller flere. Hvis $p \leq 0.10$ forlanger jeg $n \geq 1500\ p$. Ved å følge den stiplede kurven i Fig. 3 kan en avlese konsekvensene av denne enkle regel: Hvis $0 \leq p \leq 0.20$ eller $0.80 \leq p \leq 1.00$, ligger variansen mellom 0.0016 og 0.0004. Dvs. standardavviket ligger mellom 0.02 og 0.04. Hvis $0.20 \leq p \leq 0.80$, ligger standardavviket mellom 0.04 og 0.05. For visse formål ville en slik regel kanskje være rimelig.

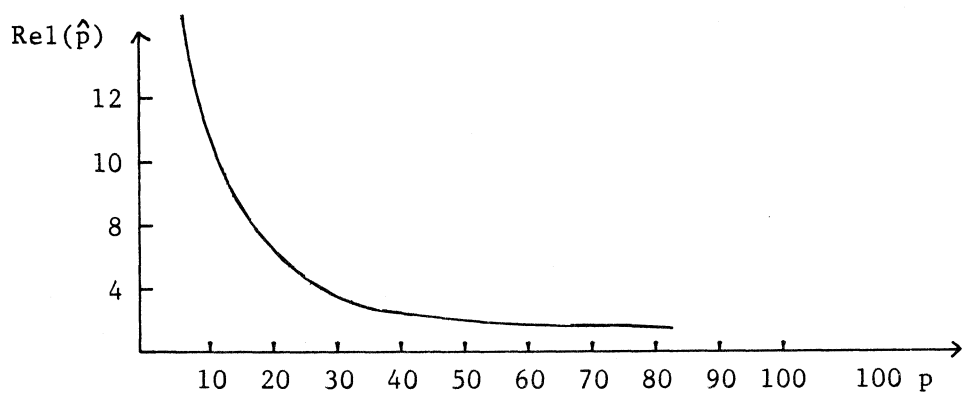
5. Bruk av relativ varians

Den enkle regel under pkt. 4 gir forskjellig nøyaktighet for forskjellige verdier av p. Dette er da også ofte rimelig, da f.eks. et standardavvik på 0.02 er lite hvis p er 0.50, men stort hvis p er 0.02.

En ser derfor ofte brukt relativ varians i stedet for varians. [1], [2], [3], [4]. Vi skal i det følgende vise at dette ikke løser de problemer en hadde under pkt. 3, hvor variansen ble brukt. Den relative varians, Rel $(\hat{p})$, er definert som forholdet mellom varians og forventningen kvadrert,

$$\text{Rel}\ (\hat{p}) = \frac{\text{var}\ (\hat{p})}{p^2}$$

Som en ser av figur 2, har Rel $(\hat{p})$ et noe annet utseende enn var $(\hat{p})$ som funksjon av p. Rel $(\hat{p})$ er en monoton avtakende funksjon av p.



Figur 2. Rel $(\hat{p})$.

Som for variansen kan en nå bestemme den minste utvalgsstørrelse en må ha for å få Rel $(\hat{p})$ under en gitt verdi. I Fig. 4 er minste tillatte n gitt som funksjon av p for forskjellige valg av Rel $(\hat{p})$.

FIG. 3. n_0 SOM FUNKSJON AV p FOR FIRE FORSKJELLIGE VERDIER AV VARIANSE

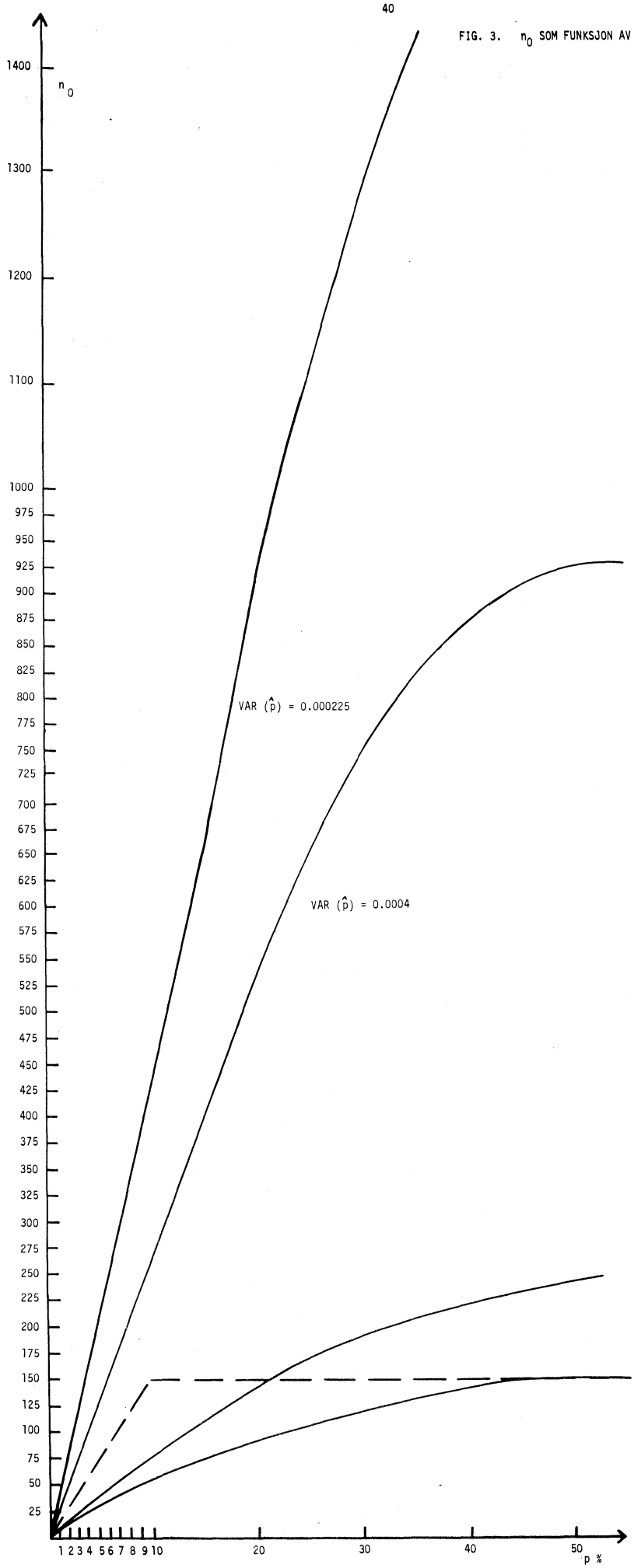
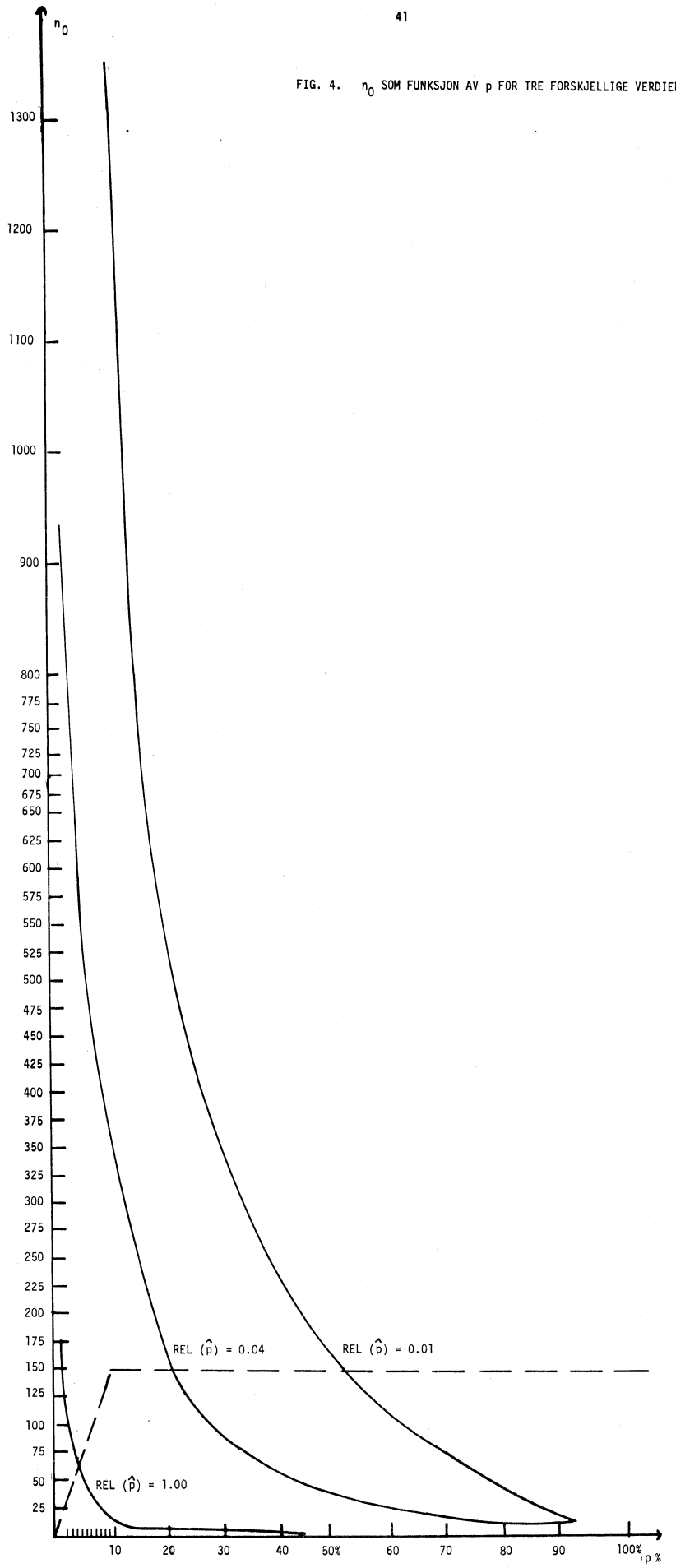


FIG. 4. n_0 SOM FUNKSJON AV p FOR TRE FORSKJELLIGE VERDIER AV RELATIV VARIANS.



På tross av at funksjonsformen er annerledes her enn for variansen, har en de samme vanskeligheter med å gi et kort svar på spørsmålet om hvor små befolkningsgrupper en kan publisere tall for. Det avhenger igjen av to ting

1. Størrelsen på p
2. Den ønskede nøyaktighet.

6. Andre kriterier

Til nå har vi vært opptatt av å sikre kvaliteten på hvert enkelt tall. Det er imidlertid ikke tilstrekkelig å vasere seg på dette når en skal velge publiseringsstandard. Svært ofte vil det nemlig for den som har behov for statistikk være relevant å sammenlikne tall, og vi må derfor også ta hensyn til dette når kravene skal settes opp. En kan derfor tenke seg å lage kriterier som går ut på at utvalget skal være så stort at forskjeller i publiserte gjennomsnitt skal være signifikante på et visst nivå. La oss gi et eksempel på dette også.

Følgende tabell er hentet fra Statistisk ukehefte 25/10 1972.

Tabell 1. Folkeavstemningen om EF.
Personer i forskjellige yrkesgrupper, etter stemmegiing. Prosent.
Del 1, intervjuperiode 15.- 22. september

Yrke	Vil stemme				Vil ikke stemme	Vet ikke om de vil stemme	I alt	Tallet på personer som svarte		
	JA	NEI	Vet ikke hvordan	Nekter å svare hvordan				I alt	Prosent	
Selvstendige i jordbruk, skogbruk og fiske	(12)	75	(2)	(0)	(5)	(5)	100	58	5	
	(14)	86	.	.	.	.	100	51		5
Andre selvstendige	46	38	5	(0)	(3)	(8)	100	63	5	
	55*	45*	.	.	.	.	100	53		6
Ansatte i industri, bygg, anlegg og gruvedrift	37	49	(4)	(0)	(4)	(6)	100	187	16	
	43	57	.	.	.	.	100	161		17
Andre ansatte	49	39	(3)	(1)	(3)	5	100	335	28	
	55	45	.	.	.	.	100	294		31
Skoleelever, studenter	39	39	(3)	(0)	(16)	(3)	100	31	3	
	50*	50*	.	.	.	.	100	24		2
Pensjonister, trygdede	37	40	(6)	(2)	(8)	(7)	100	183	16	
	48*	52*	.	.	.	.	100	142		15
Husarbeid hjemme	32	40	7	(0)	10	11	100	307	26	
	45	55	.	.	.	.	100	222		23
Andre	:	:	:	:	:	:	100	8	(1)	
	:	:	.	.	.	.	100	7		(1)
Alle	38	43	5	(1)	6	7	100	1 172	100	
	47	53	.	.	.	.	100	954		100

Ved publisering av resultatene valgte en å sette parantes rundt tall med stor relativ usikkerhet. Dessuten skal vi sette * ved de tall som ikke ville være blitt publisert hvis en hadde brukt regelen nedenfor. I dette eksemplet skal vi sette strek under noen tall for å markere at observerte forskjeller bør "tas med en klype salt".

En kunne overveie å publisere tallene for de yrkesgrupper der "ja"-prosenten eller "nei"-prosenten er signifikant større enn (eventuelt mindre enn) 50% ved et visst signifikans nivå. En slik bestemmelse fører ikke til at en kan gi en enkel regel for hvor små delpopulasjoner en kan gi tall for. Nødvendig utvalgsstørrelse for at en observert forskjell fra 50% skal være signifikant, avhenger nemlig av det observerte avvik fra 50%. I tabell 2 nedenfor har vi gitt nødvendig utvalgsstørrelse for å forkaste hypotesen $p > 0.5$ for to forskjellige signifikansnivåer og forskjellige verdier for observert prosentandel, $\hat{p}$. Jeg skal likevel sette opp følgende regel:

1. Hvis observert "ja"-prosent eller "nei"-prosent er signifikant større enn 50% med signifikansnivå 0.05 publiseres tallet uten strek under.
2. Hvis observert "ja"-prosent eller "nei"-prosent er signifikant større enn 50% med signifikansnivå 0.20 publiseres tallet med strek under.
3. Hvis observert "ja"-prosent eller "nei"-prosent er signifikant større enn 50% bare når signifikansnivået er mindre enn 0.20 publiseres tallet ikke.

I tabell 1 har jeg satt * ved tall som ikke ville vært publisert hvis en hadde brukt denne regel.

Tabell 2. Minste utvalgsstørrelse n_0 som kreves for å forkaste hypotesen $p < 0.5$ mot $p > 0.5$ for forskjellige verdier av $\hat{p}$ og forskjellige nivåer på testen.

$\hat{p}$	Nivå 5% n_0	Nivå 20% n_0
0.51	10 147	2 650
0.52	2 537	663
0.53	1 128	295
0.54	635	166
0.55	406	106
0.56	282	74
0.57	208	55
0.58	159	42
0.59	126	33
0.60	102	27
0.70	26	7
0.80	12	3
0.90	7	2

7. Sluttmerknader

Når en arbeider med publiseringsmetoder som tar sikte på å unngå misbruk av statistikken, må en ikke glemme at dette meget lett kan føre til at en holder tilbake mange tabeller som viser viktige og ukjente tendenser i befolkningen. Slike tabeller kan være viktige for oppdragsgiveren fordi de kan danne grunnlag for hypoteser som kan testes gjennom framtidige undersøkelser. Dessuten kan oppdragsgiveren sitte inne med kunnskaper som kan være av stor verdi når de kombineres med resultatene fra en utvalgsundersøkelse.

Når slike argumenter fører til at en velger en liberal publiseringspraksis, er det imidlertid viktig at dette kommer fram i innledning til rapporten for utvalgsundersøkelsen, og at det gjøres mulig for leseren å beregne (i det minste tilnermet) variansene på for samtlige tall i rapporten. Problemet med å publisere varianser har opptatt mange, og forskjellige løsninger er foreslått. Her i Byrået har vi arbeidet lite med problemet, men det er naturlig å lede det videre arbeidet med publiseringsstandarder i den retning, [5, side 475].

8. Referanser

[1] Noen tekniske problemer i forbindelse med arbeidskrafttellingene. Notat IT/SiN, 15/12-70.

[2] Hvor små befolkningsgrupper kan vi gi rimelig sikre tall for i tabellene fra arbeidskraftundersøkelsene? Notat JMH/GH, 7/12-71.

[3] Arbeidskraftundersøkelsene bør ikke offentliggjøre oppblåste tall under ca. 10 000. Notat JMH/SBr/GH, 7/12-71.

[4] Sammenhengen mellom tallet på spurte, estimert prosentsats og nøyaktighetsgrad. Notat SØP/WA.

[5] Moser, C.A. and Kalton (1971): Survey Methods in Social Investigations. Heinemann Educational Book Ltd., London.