Bortfaldsproblemet i stikprøveundersøgelser

Af Carsten Stig Poulsen*)

Indledning

I en artikel i dette tidsskrift[1]) rejser Flemming Hansen spørgsmålet om den eksisterende statistiske teoris praktiske anvendelighed på stikprøveundersøgelser af populationer, der består af personer, grupper af personer m.v. Udover den usikkerhed, som følger af, at det kun er en del af populationen, der undersøges, optræder der her tillige kilder til fejl i form af målefejl og bortfald (non-response), som stort set ikke indgår i de teoretisk-statistiske modeller.

Det er hensigten i det følgende at opstille en formel model, som kan belyse nogle sider af bortfaldsproblemets betydning for statistisk inferens. De spørgsmål, som primært behandles er: Hvilken indflydelse har bortfaldet på egenskaberne hos de estimatorer, som traditionelt anvendes i stikprøveundersøgelser, og hvorledes kan undersøgelserne tilrettelægges under hensyntagen til bortfaldet.

Fremstillingen prætenderer ikke at være fuldt dækkende, men repræsenterer nogle indledende overvejelser, som måske kan danne grundlag for videre forskning inden for området.

*) Cand. polit., adjunkt, AUC.

[1]) Nr. 4/1977.

Bortfaldsproblemet i stikprøveundersøgelser

Af Carsten Stig Poulsen*)

Indledning

I en artikel i dette tidsskrift[1]) rejser Flemming Hansen spørgsmålet om den eksisterende statistiske teoris praktiske anvendelighed på stikprøveundersøgelser af populationer, der består af personer, grupper af personer m.v. Udover den usikkerhed, som følger af, at det kun er en del af populationen, der undersøges, optræder der her tillige kilder til fejl i form af målefejl og bortfald (non-response), som stort set ikke indgår i de teoretisk-statistiske modeller.

Det er hensigten i det følgende at opstille en formel model, som kan belyse nogle sider af bortfaldsproblemets betydning for statistisk inferens. De spørgsmål, som primært behandles er: Hvilken indflydelse har bortfaldet på egenskaberne hos de estimatorer, som traditionelt anvendes i stikprøveundersøgelser, og hvorledes kan undersøgelserne tilrettelægges under hensyntagen til bortfaldet.

Fremstillingen prætenderer ikke at være fuldt dækkende, men repræsenterer nogle indledende overvejelser, som måske kan danne grundlag for videre forskning inden for området.

*) Cand. polit., adjunkt, AUC.

[1]) Nr. 4/1977.

Bortfaldet som problem

Ved *bortfald* vil vi i det følgende forstå manglende information om nogle af de analyseenheder, som var planlagt at skulle indgå i undersøgelsen. Denne fejlkilde knytter sig ikke specielt til stikprøveundersøgelser, men kan lige såvel optræde ved totaltællinger.

Bortfaldet som problem forstærkes i det omfang de bortfaldne analyseenheder adskiller sig væsentligt fra resten af stikprøven med hensyn til det undersøgte kendetegn. Da vil de skøn over forholdene i populationen, som bygger på den gennemførte stikprøve, nødvendigvis blive fejlagtige. I forbindelse med interviewundersøgelser kan der sondres mellem to hovedtyper af bortfald: respondenter, som ikke træffes og nægtere. De to typer af bortfald adskiller sig på en række punkter. Man vil således ved genbesøg være i stand til at nedbringe antallet af ikke-trufne respondenter, mens det er mere krævende at overtale folk med nægter-indstilling til at deltage i undersøgelsen. Forudsætningen om, at bortfaldet ikke afviger væsentligt fra resten af populationen kan også være mere hasaderet i forbindelse med nægtergruppen end for gruppen af ikke-trufne. Endelig skal man være opmærksom på en mulig sammenhæng mellem formålet med undersøgelsen og tilbøjeligheden til bortfald.

Hvis man f.eks. udtager en stikprøve med henblik på en vurdering af antallet af sortseere, er det ikke utænkeligt, at nægterprocenten blandt sortseerne er større end blandt licensbetalere. (En anden mulig fejlkilde er her målefejl i form af urigtige oplysninger).

Vi skal overalt i det følgende antage, at bortfaldet er uden sammenhæng med det undersøgte kendetegn.

En statistisk model med bortfald

Problemstillingen i en model med bortfald sammenlignet med traditionel statistisk stikprøveteori er søgt anskueliggjort i figur 1.

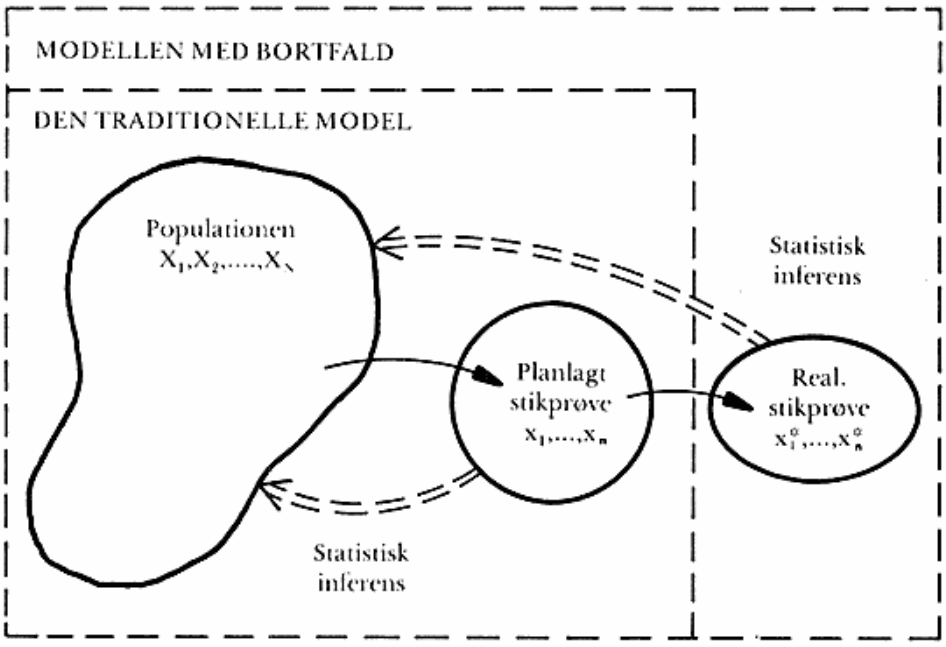

Fig. 1.

Grundlaget i den traditionelle model er en stikprøve, som er udtaget *tilfældigt,* d.v.s. sandsynligheden for, at en vilkårlig enhed i populationen indgår i stikprøven er *kendt* – men ikke nødvendigvis den samme for alle enheder.

Derved etableres en sandsynlighedsteoretisk sammenhæng mellem kendetegn i populationen og observerede kendetegn i stikprøven, som danner grundlag for *statistisk inferens* med angivelse af den dertil knyttede usikkerhed. Udarbejdelsen af en stikprøveplan bygger på et *effektivitetskriterium:* For et givet økonomisk budget søges stikprøveusikkerheden minimeret eller – omvendt – en given usikkerhed ønskes opnået ved de lavest mulige omkostninger.

I den mere generelle model med bortfald reduceres den planlagte stikprøve, som følge af non-response hos nogle af de udvalgte enheder. Stikprøven omfatter kun n^* enheder i stedet for de planlagte n, et bortfald på $n-n^*$. Problemet er nu, i hvilket omfang inferens reglerne fra den traditionelle model kan anvendes i dette mere generelle tilfælde.

I den model, der opstilles for at belyse dette spørgsmål, anvendes følgende notation:

X_ν : Værdien af kendetegnet for ν'te enhed i populationen, $\nu = 1,2,...,N$.

x_i : Værdien af kendetegnet for i'te enhed i den planlagte stikprøve, $i = 1,2,...,n$.

x_j^*: Værdien af kendetegnet for j'te enhed i den realiserede stikprøve, $j = 1,2,...,n^*$.

θ_ν : Sandsynligheden for, at v'te enhed i populationen udtages til den planlagte stikprøve, $\nu = 1,2...,N$. $0 < \theta_\nu < 1$.

ρ_ν : Sandsynligheden for, at v'te enhed i populationen *ikke* indgår i den realiserede stikprøve, skønt den er udtaget til den planlagte stikprøve, $\nu = 1,2...,N$. $0 < \rho_\nu < 1$.

Vi antager, at formålet med stikprøven er at estimere middeltallet i populationen:

$$\overline{X} = \frac{1}{N} \sum_{\nu=1}^{N} X_\nu \qquad (1)$$

Simpel tilfældig udvælgelse

Det forudsættes, at den planlagte stikprøve udtages ved *simpel tilfældig udvælgelse*. Dette indebærer, at alle enheder i populationen har samme sandsynlighed for at blive valgt:

$$\theta_\nu = \frac{n}{N} \; ; \; \nu = 1,2,...,N \qquad (2)$$

Da gælder følgende sætning i *den traditionelle model:*

Sætning 1: Middeltallet i stikprøven

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (3)$$

er en *middelret* estimator over middeltallet i populationen $\overline{X}$.

Bevis: Lad indikatorvariablen α_ν være defineret ved

$$\alpha_\nu = \begin{cases} 1, \text{ hvis } X_\nu \text{ indgår i stikprøven} \\ 0, \text{ ellers} \end{cases} \tag{4}$$

Da kan estimatoren $\bar{x}$ skrives

$$\bar{x} = \frac{1}{n} \sum_{\nu=1}^{N} \alpha_\nu \cdot X_\nu \tag{5}$$

Den forventede værdi af $\bar{x}$, $E(\bar{x})$, kan findes:

$$E(\bar{x}) = \frac{1}{n} \sum_{\nu=1}^{N} E(\alpha_\nu) \cdot X_\nu . \tag{6}$$

Idet

$$E(\alpha_\nu) = 1 \cdot Pr\{\alpha_\nu = 1\} + 0 \cdot Pr\{\alpha_\nu = 0\} = \frac{n}{N}$$

fås:

$$E(\bar{x}) = \frac{1}{n} \Sigma \frac{n}{N} X_\nu = \bar{X} \quad \blacktriangleleft$$

Uden bevis[2]) anføres.

Sætning 2: Den teoretiske varians på estimatoren $\bar{x}$ er givet ved

$$V(\bar{x}) = \frac{\sigma^2}{n} \frac{N-n}{N-1} \simeq \sigma^2 \left(\frac{1}{n} - \frac{1}{N}\right) \tag{7}$$

hvor

$$\sigma^2 = \frac{1}{N} \sum_{\nu=1}^{N} (X_\nu - \bar{X})^2 \tag{8}$$

er variansen i populationen.

---

[2]) Et bevis kan f.eks. findes hos Jensen (1974).

For $N \to \infty$ fremkommer den velkendte formel for variansen på middeltallet af uafhængige observationer fra en uendelig population.

I *modellen med bortfald* skal vi nu betragte

$$\bar{x}^* = \frac{1}{n^*} \sum_{j=1}^{n^*} x_j^* \tag{9}$$

d.v.s. middeltallet i den *gennemførte* stikprøve som estimator for $\bar{X}$. Med α_ν defineret ved (4) kan (9) skrives:

$$\bar{x}^* = \frac{1}{n^*} \sum_{\nu=1}^{N} \alpha_\nu X_\nu \tag{10}$$

Her er fordelingen af α_ν imidlertid ændret. Vi antager, at sandsynligheden for at indgå i den realiserede stikprøve kan skrives:

$$\Pr\{\alpha_\nu = 1\} = \theta_\nu \cdot (1-\rho_\nu) = \frac{n}{N}(1-\rho_\nu) \tag{11}$$

d.v.s. som produktet af sandsynligheden for at blive udtaget til den planlagte stikprøve og tilbøjeligheden (sandsynligheden) for at indgå i den realiserede stikprøve. Man forestiller sig altså, at populationen ikke blot er karakteriseret ved kendetegnet X_ν, men også ved bortfaldstilbøjelighederne ρ_ν . En enhed består af et talpar (X_ν, ρ_ν), som ved udtrækning repræsenteres ved (x_i, ρ_i). x_i vil da indgå i den realiserede stikprøve med sandsynligheden $1-\rho_i$.

Mens θ_ν alene afhænger af den valgte stikprøveplan, er ρ_ν i vidt omfang uden for planlæggerens kontrol.

For $\rho_\nu = 0$ fremkommer den traditionelle model som et specialtilfælde.

Vi skal nu vise følgende:

Sætning 1:* Middeltallet $\bar{x}^*$ i den realiserede stikprøve er *som hovedregel ikke* en middelret estimator for middeltallet i $\bar{X}$ populationen.

Bevis: Med α_ν defineret ved (4) kan antallet af enheder n^* i den gennemførte stikprøve skrives:

$$n^* = \sum_{\nu=1}^{N} \alpha_\nu$$

d.v.s. n^* er en stokastisk variabel. Den forventede værdi af $\bar{x}^*$ er givet som:

$$E(\bar{x}^*) = E(\frac{1}{n^*} \sum_{\nu=1}^{N} \alpha_\nu \cdot X_\nu) = \sum_{\nu=1}^{N} E(\frac{\alpha_\nu}{n^*}) \cdot X_\nu \qquad (12)$$

$\bar{x}^*$ vil derfor kun være middelret, såfremt

$$\sum_{\nu=1}^{N} E(\frac{\alpha_\nu}{n^*}) \cdot X_\nu = \frac{1}{N} \sum_{\nu=1}^{N} X_\nu$$

$$\Leftrightarrow \quad \sum_{\nu=1}^{N} (E(\frac{\alpha_\nu}{n^*}) - \frac{1}{N}) \cdot X_\nu = 0 \qquad (13)$$

hvilket ikke kan antages at gælde i almindelighed. ◀

Vi skal nu undersøge (13) i nogle specialtilfælde. Hvis n^* er rimelig stor, vil α_ν kunne betragtes som uafhængig af $n^* = \Sigma\alpha_\nu$. Da gælder:

$$E(\frac{\alpha_\nu}{n^*}) = E(\frac{1}{n^*}) \cdot E(\alpha_\nu)$$

Anvendelse af (11) giver

$$E(\alpha_\nu) = \frac{n}{N}(1-\rho_\nu)$$

og dermed kan betingelsen (13) skrives

$$E(\frac{1}{n^*}) \cdot \frac{n}{N} \sum_{\nu=1}^{N} (1-\rho_\nu) \cdot X_\nu - \bar{X} = 0 \qquad (14)$$

Introduceres nu mere formelt antagelsen om *uafhængighed mellem bortfaldssandsynlighed og værdien af kendetegnet*, haves

$$\sum_{\nu=1}^{N} (\rho_\nu - \bar{\rho})(X_\nu - \bar{X}) = 0$$

$$\Rightarrow \quad \sum_{\nu=1}^{N} (1-\rho_\nu) X_\nu = N \cdot \bar{X}(1-\bar{\rho})$$

som indsat i (14) giver

$$E\left(\frac{1}{n^*}\right) = \frac{1}{n(1-\bar{\rho})} \tag{15}$$

Da $1/n^*$ er en konveks funktion af n^* gælder det ifølge Jensens ulighed, at

$$E\left(\frac{1}{n^*}\right) > \frac{1}{E(n^*)}$$

Vi har tillige

$$E(n^*) = \sum_{\nu=1}^{N} E(\alpha_\nu) = \sum_{\nu=1}^{N} \frac{n}{N}(1-\rho_\nu) = n(1-\bar{\rho})$$

og dermed

$$E\left(\frac{1}{n^*}\right) > \frac{1}{n(1-\bar{\rho})}$$

Selv når n^* er stor gælder sætning 1* stadig. Variationen i det gennemførte antal observationer er medvirkende til, at $\bar{x}^*$ er et skævt skøn over $\bar{X}$. Skævheden vokser med variansen på n^*.

Vi ser dernæst på *en vigtig undtagelse* fra hovedreglen i sætning 1*:

Korollar 1:* Såfremt bortfaldssandsynligheden ρ_ν er konstant, $\rho_\nu = \bar{\rho}$, vil $\bar{x}^*$ være en middelret estimator for $\bar{X}$.

Bevis: Når $\rho_\nu = \bar{\rho}$, er $\Pr\{\alpha_\nu = 1\} = \frac{n}{N}(1-\bar{\rho})$ konstant, uafhængig af ν. Antages der tillige uafhængighed mellem enhederne, hvad angår deres tilstedeværelse i den realiserede stikprøve, vil fordelingen af n^* være

givet ved en binomialfordeling med antalsparameteren N og sandsynlighedsparameteren $\frac{n}{N}(1-\bar{\rho}\ )$. Der gælder nu generelt[3]):

$$E\left(\frac{\alpha_\nu}{n^*}\right) = E\left(\frac{1}{n^*} \cdot E(\alpha_\nu \mid n^*)\right) \tag{16}$$

hvor $E(\alpha_\nu|n^*)$ betegner den betingede middelværdi af α_ν, givet summen $n^* = \Sigma\alpha_\nu$. Da n^* en binomialfordelt gælder[4]):

$$E(\alpha_\nu \mid n^*) = Pr\{\alpha_\nu = 1 \mid n^*\} = \frac{n^*}{N} \tag{17}$$

$$\Rightarrow \quad E\left(\frac{\alpha_\nu}{n^*}\right) = E\left(\frac{1}{n^*} \cdot \frac{n^*}{N}\right) = \frac{1}{N}$$

og dermed

$$E(\bar{x}^*) = \frac{1}{N} \Sigma X_\nu = \bar{X}$$ ◀

En intuitiv forklaring på dette resultat er:

Når bortfaldssandsynligheden er konstant over enhederne i populationen er den »mekanisme«, som udvælger enheder fra den planlagte til den realiserede stikprøve helt analog med udvalgsproceduren fra populationen til den planlagte stikprøve.

Usikkerheden på skønnet $\bar{x}^*$ skal nu vurderes. Som det kunne forventes gælder:

Sætning 2:* Bortfald bevirker en forøgelse af usikkerheden, målt ved variansen på estimatoren $\bar{x}^*$.

Delvist bevis: Variansen på $\bar{x}^*$, $V(\bar{x}^*)$, kan skrives[5])

$$V(\bar{x}^*) = E[V(\bar{x}^* \mid n^*)] + V[E(\bar{x}^* \mid n^*)] \tag{18}$$

For givet n^* vil $E(\bar{x}^*|n^*)$ i det generelle tilfælde med ρ_ν forskellig afvige fra $\bar{X}$ og afhænge af n^*. Derfor er andet led i (18) positivt. Ligeledes vil den betingede varians $V(\bar{x}^*|n^*)$ være større end variansen i den traditionelle model med samme stikprøvestørrelse:

[3]) For produktet af to stokastiske variable X og Y gælder det, at E(XY) = E(XE(Y|X)) = E(YE(X|Y)), se f.eks. Jensen (1974), p. 55.

[4]) Den betingede sandsynlighed for, at et vilkårligt binomiallovs forsøg i en forsøgsrække på N er »gunstigt«, *givet* n^* »gunstige« ialt, er n^*/N, uafhængigt af den *marginale* sandsynlighed for »gunstigt« udfald i et forsøg.

[5]) se f.eks. Jensen (1974), p. 59.

$$V(\bar{x}^*|n^*) > \sigma^2 \left(\frac{1}{n^*} - \frac{1}{N}\right)$$

$$\Rightarrow \quad E[V(\bar{x}^*|n^*)] > \sigma^2 \left(E\left(\frac{1}{n^*}\right) - \frac{1}{N}\right)$$

Da $E\left(\frac{1}{n^*}\right) > 1/E(n^*) = 1/n(1-\bar{\rho})$ fås

$$E[V(\bar{x}^*|n^*)] > \sigma^2 \left(\frac{1}{n(1-\bar{\rho})} - \frac{1}{N}\right) > V(\bar{x}) \quad \blacktriangleleft$$

Specialtilfældet med $\rho_\nu = \bar{\rho}$ giver følgende

Korollar 2:* Såfremt bortfaldssandsynligheden ρ_ν er konstant, $\rho_\nu = \bar{\rho}$, vil

$$V(\bar{x}^*) \simeq \sigma^2 \left(E\left(\frac{1}{n^*}\right) - \frac{1}{N}\right) \tag{19}$$

Intuitivt bevis: Som nævnt er situationen med konstant bortfaldssandsynlighed analog til simpel tilfældig udvælgelse i den traditionelle model. For givet n^* har vi derfor ifølge sætning 1 og 2:

$$E(\bar{x}^*|n^*) = \bar{X}$$

$$V(\bar{x}^*|n^*) = \frac{\sigma^2}{n^*}\left(\frac{N-n^*}{N-1}\right) \simeq \sigma^2 \left(\frac{1}{n^*} - \frac{1}{N}\right)$$

og dermed

$$V(\bar{x}^*) \simeq \sigma^2 \left(E\left(\frac{1}{n^*}\right) - \frac{1}{N}\right) \quad \blacktriangleleft \tag{20}$$

Udnytter vi, at[6])

$$E\left(\frac{1}{n^*}\right) \simeq \frac{1}{E(n^*)} + \frac{V(n^*)}{[E(n^*)]^3} = \frac{1+\gamma^2_{n^*}}{E(n^*)} \tag{21}$$

hvor $\gamma^2_{n^*} = V(n^*)/[E(n^*)]^2$ er *den relative varians*, kan (20) skrives

[6]) (21) forekommer ved rækkeudvikling af $f(x) = \frac{1}{x}$ omkring $x = E(x)$ og efterfølgende middelværdidannelse.

$$V(\bar{x}^*) \simeq \sigma^2 \left(\frac{1+\gamma^2_{n^*}}{n(1-\bar{\rho})} - \frac{1}{N}\right) \qquad (22)$$

hvoraf det klart fremgår, at usikkerheden i modellen med bortfald øges, ikke blot som følge af det reducerede antal observationer, $n > n(1-\bar{\rho})$, men også på grund af variationen i n^*, målt ved den relative varians $\gamma^2_{n^*}$.

Stratificeret tilfældig udvælgelse

Hovedkonklusionen af analysen over bortfaldets virkninger ved simpel tilfældig udvælgelse er, at anvendeligheden af den traditionelle model afhænger af, hvor *homogen* populationen er m.h.t. bortfaldssandsynligheden. Hvor denne forudsætning ikke er opfyldt, fører dette naturligt tanken hen på en stratifikation af populationen med henblik på dannelse af delpopulationer med (tilnærmelsesvis) konstant bortfaldssandsynlighed, hvorfra der udtages en stikprøve ved simpel tilfældig udvælgelse.
Lad populationen være opdelt i L strata, som hver karakteriseres ved

1) stratumstørrelsen N_h
2) stratummiddeltallet $\bar{X}_h$
3) stratumvariansen σ^2_h
4) bortfaldssandsynligheden ρ_h, som er konstant inden for et stratum,

hvor h = 1,2,...,L. Fra hvert stratum udvælges ved simpel tilfældig udvælgelse n_h enheder, der efter bortfald reduceres til n_h^*. Da vil $\bar{x}_h^* \approx \bar{X}_h$, iflg. det foregående og

$$\bar{x}^*_{str} = \sum_{h=1}^{L} \frac{N_h}{N} \cdot \bar{x}_h^* \; ; \quad N = \sum_{h=1}^{L} N_h \qquad (23)$$

vil dermed være en middelret estimator over populationsgennemsnittet $\bar{X} = \Sigma (N_h/N) \cdot \bar{X}_h$. Variansen på $\bar{x}^*_{str}$ er givet ved

$$V(\bar{x}^*_{str}) = \sum_{h=1}^{L} \frac{N_h^2}{N^2} V(\bar{x}_h^*) = \sum_{h=1}^{L} \frac{N_h^2}{N^2} \sigma_h'^2 \left(E\left(\frac{1}{n_h^*}\right) - \frac{1}{N}\right)$$

$$\simeq \sum_{h=1}^{L} \frac{N_h^2}{N^2} \cdot \sigma_h^2 \left(\frac{1+\gamma^2_{n_h^*}}{n_h(1-\bar{\rho}_h)} - \frac{1}{N}\right) \qquad (24)$$

jf. (22). Af (24) ses, at stratifikationen bør ske således at strata er *homogene,* hvad angår det kendetegn, der ønskes undersøgt. Da homogenitet med hensyn til bortfald inden for et stratum er forudsætningen for modellen, kan der opstå problemer, hvis disse to kriterier er *konkurrerende.* Da må stratifikationen ske ved en afvejning mellem homogenitet m.h.t. kendetegnet og bevarelse af den simple statistiske model. Det er åbenbart, at en sådan afvejning vil være uhyre vanskelig at gennemføre i praksis.

Som det også fremgår af (24) afhænger usikkerheden på skønnet tillige af *stikprøvens fordeling på strata.* Dette problem skal nu drøftes nøjere.

Fordelingen på strata af en given stikprøve n i den traditionelle model siges at være *optimal,* hvis variansen $V(\bar{x}_{str})$ minimeres under bibetingelsen $\Sigma n_h = n$. Det kan vises[7]), at løsningen resulterer i fordelingen

$$n_h = \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h} \cdot n \qquad (25)$$

I modellen med bortfald kan det tilsvarende problem formuleres med udgangspunkt i (24) og løsningen bestemmes efter samme principper som i den traditionelle model. Antager vi for simpelheds skyld, at den relative varians er konstant over strata, $\gamma^2_{n_h^*} = \gamma^2$, kan det vises, at

$$n_h = \frac{N_h \sigma_h / \sqrt{1-\bar{\rho}_h}}{\Sigma N_h \sigma_h / \sqrt{1-\bar{\rho}_h}} \cdot n \qquad (26)$$

---

[7]) Jensen (1974), p. 124

giver den optimale fordeling af den *planlagte* stikprøve på strata i modellen med bortfald. Det ses, at jo større bortfaldssandsynligheden $\bar{p}_h$ er, desto større andel af stikprøven skal udtages fra stratum h, men da $\bar{p}_h$ optræder under kvadratrodstegn, skal forskellene mellem $\bar{p}_h$ være store, før det giver væsentlige udslag i stikprøvens fordeling[8]).

Korrektion for bortfald

I den nævnte artikel af Flemming Hansen diskuteres forskellige metoder til *opvejning* med det primære formål at korrigere for bortfaldet. Den sammenhæng, som herved tilsyneladende etableres mellem bortfaldsproblemet og behovet for opvejning, er uheldig. Vejning er påkrævet, når den udtagne stikprøve ikke er repræsentativ for populationen med hensyn til karakteristika, som har eller antages at have indflydelse på det kendetegn, man ønsker at undersøge. Stikprøvens information om kendetegnet, er *betinget* af disse karakteristika, og for at kunne skønne over kendetegnets udbredelse i *hele* populationen må man veje med vægte, som er bestemt af *de relative stratumstørrelser,* N_h/N. Dette gælder uafhængigt af den måde, hvorpå stikprøven er udvalgt og uanset omfanget af bortfald. Med kendskab til vægtene N_h/N repræsenterer denne vejning ikke noget *statistisk* problem. Er vægtene ikke kendte, kan der være tale om at stratificere den udtagne stikprøve, hvis $n_h^*/n^* \approx N_h/N$, og da indføres endnu en kilde til usikkerhed i den statistiske inferens.

Bortfaldet som statistisk problem bør i princippet inddrages allerede ved tilrettelæggelsen af undersøgelsen. Man kan søge at udnytte information fra tidligere undersøgelser af samme art. Det er ikke utænkeligt, at bortfaldssandsynligheden afhænger af en række generelle kriterier så som alder, køn, erhverv, uddannelse m.m., som hyppigt anvendes ved stratifikation af populationen. Forskelle i bortfaldssandsynligheder *mellem* disse strata kan testes med anvendelse af oplysninger om det faktiske bortfald, $n_h - n_h^*$, i tidligere undersøgelser, som da kan indgå i den planlagte stikprøves fordeling på strata. Man antager her implicit, at ρ_h er konstant over individer *inden for* hvert stratum.

---

[8]) Det bemærkes, at $(1-\rho_h)$ indgår i (26) som grænseomkostningerne i en traditionel model, der optimeres under hensynstagen til en lineær omkostningsfunktions $C = c_0 + \sum_{h=1} c_h n_h$, se Jensen (1974), pp. 130-131.

For så vidt angår den del af bortfaldet, der skyldes ikke-trufne respondenter, er det muligt at tage hensyn til bortfaldet ved at forlade den simple tilfældige udvælgelse med $\theta_\nu = \frac{n}{N}$ og tillade θ_ν at variere. Med kendskab til ρ_ν kan θ_ν fastsættes således, at sandsynligheden for at indgå i den gennemførte stikprøve, $\Pr\{\alpha_\nu = 1\} = \theta_\nu (1-\rho_\nu)$, er konstant. Herved opnås at den traditionelle model kan anvendes på trods af den tilfældige variation i bortfaldet. Personparameteren ρ_ν må også her søges forklaret og estimeret ved en række generelle egenskaber ved enhederne i populationen.

Litteratur:

Hansen, Flemming: Sampling teori og anvendt statistik. Erhvervsøkonomisk Tidsskrift nr. 4, 1977.

Jensen, Niels Erik (red.): Stikprøveteori. København 1974.

For så vidt angår den del af bortfaldet, der skyldes ikke-trufne respondenter, er det muligt at tage hensyn til bortfaldet ved at forlade den simple tilfældige udvælgelse med $\theta_\nu = \frac{n}{N}$ og tillade θ_ν at variere. Med kendskab til ρ_ν kan θ_ν fastsættes således, at sandsynligheden for at indgå i den gennemførte stikprøve, $\Pr\{\alpha_\nu = 1\} = \theta_\nu (1-\rho_\nu)$, er konstant. Herved opnås at den traditionelle model kan anvendes på trods af den tilfældige variation i bortfaldet. Personparameteren ρ_ν må også her søges forklaret og estimeret ved en række generelle egenskaber ved enhederne i populationen.

Litteratur:

Hansen, Flemming: Sampling teori og anvendt statistik. Erhvervsøkonomisk Tidsskrift nr. 4, 1977.

Jensen, Niels Erik (red.): Stikprøveteori. København 1974.