MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 162 (Ov 4)

Overzicht van een aantal artikelen over de theorie van samengestelde steekproeven.

door
J.F.van Haastrecht.

1954.

1. Inleiding.

In dit kleine overzichtsrapport worden de resultaten van enkele in de literatuur vermelde steekproefmethoden voor samengestelde populaties beschreven, waarmee bij voorbeeld de kosten van het onderzoek onder bepaalde voorwaarden geminimaliseerd worden.

Uitgegaan wordt van een populatie van elementen, bij voorbeeld bestaande uit alle eenheden van een product die op één dag vervaardigd zijn. Hierin onderscheidt men een aantal sub-populaties, zoals de groepen eenheden van dat product welke op die dag door de verschillende machines zijn gemaakt. Wij beschouwen een bepaalde grootheid die bij ieder van de elementen kan worden gemeten, b.v. het gewicht van elke eenheid. De aantallen elementen in de sub-populaties alsmede de spreidingen van de beschouwde grootheid binnen de sub-populaties zijn in het algemeen niet gelijk.

In de statistische literatuur wordt het probleem behandeld hoe men op de "beste wijze" een steekproef uit die populatie kan nemen indien men daaruit het gemiddelde van de beschouwde grootheid over de populatie wil schatten. Die "beste wijze" kan men op verschillende manieren definiëren. In dit overzicht wordt hij in de regel opgevat als een zodanige keuze van de aantallen der elementen uit de sub-populaties dat aan één der beide volgende voorwaarden voldaan is:

Als bepaalde kostenfactoren gegeven zijn,

a) is bij een gegeven beschikbaar geldsbedrag de spreiding in de schatting van het totaal gemiddelde der beschouwde grootheid zo klein mogelijk gemaakt, of

b) zijn bij een voorgeschreven waarde van die spreiding de kosten van het experiment zo klein mogelijk gemaakt.

Naast een dergelijke optimale inrichting van het steekproefexperiment in diverse gevallen worden in de literatuur ook veelal de formules gegeven voor de spreiding in de schatting van het totaal gemiddelde, uitgedrukt in de steekproefuitgebreidheden en in de spreidingen van de grootheid binnen en tussen de sub-populaties.

In de literatuur worden twee methoden behandeld die wij, daar hiervoor geen gangbare Nederlandse termen bekend zijn, met de Engelse benamingen zullen aanduiden: "two-stage sampling" en "stratified sampling". Bij two-stage sampling wordt eerst aselect(random) een aantal sub-populaties

gekozen en daarna worden uit ieder van die sub-populaties aselect een aantal elementen getrokken. In het bovengenoemde voorbeeld komt dit dus hierop neer dat men uit elk van de productiehoeveelheden van een door loting aangewezen gedeelte van de machines aselect een steekproef neemt. Bij stratified sampling wordt aselect een steekproef genomen uit alle sub-populaties, dus bij ons voorbeeld uit de productie van elk van de machines.

Wij zullen hier de volgende notatie aanhouden voor de waarden, die de beschouwde grootheid op de populatie aanneemt:

Sub-populatie 1: $X_{11}, X_{12}, \ldots\ldots, X_{1N_1}$

" " 2: $X_{21}, X_{22}, \ldots, X_{2N_2}$

" " M: $X_{M1}, X_{M2}, \ldots\ldots\ldots, X_{MN_M}$

Dus M is het aantal sub-populaties, en N_i het aantal elementen van sub-populatie i.

$$X_i \overset{\text{def}}{=} \frac{1}{N_i} \sum_{j=1}^{N_i} X_{ij} \qquad (1)$$

$$\sigma_i^2 \overset{\text{def}}{=} \frac{1}{N_i} \sum_{j=1}^{N_i} (X_{ij} - X_i)^2 \qquad (2)$$

$$X \overset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij}, \quad \text{waarbij} \quad N = \sum_{i=1}^{M} N_i \qquad (3)$$

$$\sigma_t^2 \overset{\text{def}}{=} \frac{1}{M} \sum_{i=1}^{M} (X_i - X)^2 \qquad (4)$$

De variantie van de beschouwde grootheid binnen de sub-populatie i wordt dus aangegeven met σ_i^2 en de variantie van het gemiddelde X_i van die grootheid tussen de sub-populaties met σ_t^2.

Bij two-stage sampling neemt men nu aselect m sub-populaties uit het totale aantal M; wij nummeren deze sub-populaties opnieuw $1, \ldots, m$. Men neemt nu uit ieder van deze sub-populaties aselect een aantal elementen. Deze aantallen geven wij aan met $n_1, n_2, \ldots\ldots, n_m$ zodat dus het aantal waarnemingen uit de i-de sub-populatie, die in de steekproef betrokken is, gelijk is aan n_i. De waarden die de beschouwde grootheid op de steekproef aanneemt geven wij dan als volgt aan:

Deelsteekproef 1: $x_{11}, x_{12}, \ldots\ldots, x_{1n_1}$

" " 2: $x_{21}, x_{22}, \ldots\ldots, x_{2n_2}$

m: $x_{m1}, x_{m2}, \ldots\ldots, x_{mn_m}$

Het gemiddelde van de i-e deelsteekproef wordt aangegeven met

$$x_i \overset{\text{def}}{=\!=} \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \qquad (5)$$

Wij zullen hieronder de inhoud van een aantal artikelen over deze onderwerpen in het kort weergeven.

2. SOPHIE MARCUSE, Optimum allocation and variance components in nested sampling with an application to chemical analysis, Biometrika, 5 (1949) p. 189-206.

Een speciaal geval van hetgeen in dit artikel beschouwd wordt is het volgende:
Het aantal sub-populaties M en de aantallen elementen binnen de sub-populaties $N_1, N_2, \ldots$ zijn oneindig. Van m sub-populaties wordt een steekproef genomen, elk met uitgebreidheid $\bar{n}$. De variantie van de beschouwde grootheid binnen elke sub-populatie is gelijk en wordt aangegeven met $\bar{\sigma}^2$ en die van het gemiddelde tussen de sub-populaties wordt aangegeven met σ_t^2. Zijn $\bar{\sigma}^2$ en σ_t^2 niet bekend dan moeten zij uit een voorafgaand onderzoek geschat worden. De vaste kosten van het nemen van een steekproef uit een sub-populatie, ongeacht het aantal waarnemingen, worden voor elke sub-populatie gelijk gesteld aan c_1. De extra kosten van het in de steekproef betrekken van een element uit een sub-populatie worden voor alle elementen uit alle sub-populaties gelijk gesteld aan c_2. De totale kosten van het experiment kunnen dan worden voorgesteld door $C = m c_1 + m\bar{n} c_2$.

Zijn nu σ_t^2, $\bar{\sigma}^2$, c_1 en c_2 bekend dan wordt bij voorgeschreven totale kosten C de spreiding in het totale steekproefgemiddelde x, waarbij $x \overset{\text{def}}{=\!=} \frac{1}{m\bar{n}} \sum_{i=1}^{m} \sum_{j=1}^{\bar{n}} x_{ij}$, zo klein mogelijk wanneer men voor m en $\bar{n}$ gehele waarden kiest welke dicht liggen bij

$$m' = \frac{C \cdot \sigma_t}{\sqrt{c_1}\,(\sigma_t \sqrt{c_1} + \bar{\sigma}\sqrt{c_2})} \qquad \text{en} \qquad \bar{n}' = \frac{\bar{\sigma}}{\sigma_t} \sqrt{\frac{c_1}{c_2}} \qquad (6,7)$$

De spreiding σ_x heeft dan een waarde welke dicht ligt bij:

$$\sigma_x' = \sigma_t \sqrt{\frac{c_1}{C}} + \bar{\sigma} \sqrt{\frac{c_2}{C}} \qquad (8)$$

Bij voorgeschreven waarde S van de spreiding in het totaal gemiddelde worden de kosten zo klein mogelijk indien men voor m en $\bar{n}$ gehele waarden kiest die dicht liggen bij:

$$m' = \frac{\sigma_t \left(\sigma_t \sqrt{c_1} + \bar{\sigma} \sqrt{c_2} \right)}{S^2 \sqrt{c_1}} \quad \text{en} \quad \bar{n}' = \frac{\bar{\sigma}}{\sigma_t} \sqrt{\frac{c_1}{c_2}} \qquad (9)$$

De kosten liggen dan dicht bij

$$C' = \left(\frac{\sigma_t \sqrt{c_1} + \bar{\sigma} \sqrt{c_2}}{S} \right)^2 \qquad (10)$$

Indien $\bar{\sigma}$ en σ_t niet nauwkeurig bekend zijn kan men in beide behandelde gevallen een zuivere schatting krijgen van de variantie in het totale steekproefgemiddelde door te nemen:

$$s_x^2 = \frac{\sum_{i=1}^{m} (x_i - x)^2}{(m-1)\, m} \qquad (11)$$

Indien $c_1 = 0$ wordt het resultaat in beide gevallen: $m = \infty$ en $\bar{n} = 0$. Daar de kleinste waarde van $\bar{n}$, die in aanmerking komt, gelijk is aan 1, ligt het voor de hand om dan in deze gevallen $\bar{n} = 1$ en m dicht bij $\frac{C}{c_2}$, resp. $\frac{\sigma_t^2 + \bar{\sigma}^2}{S^2}$ te kiezen.

In het eerste van deze gevallen ligt dan de variantie dicht bij:

$$\sigma_x'^2 = \frac{c_2}{C} \left(\bar{\sigma}^2 + \sigma_t^2 \right)$$

en in het tweede liggen de kosten C dicht bij:

$$C' = \frac{c_2}{S^2} \left(\bar{\sigma}^2 + \sigma_t^2 \right)$$

Het artikel behandelt uitvoerig het geval van 3-stage sampling. Hierbij onderscheidt men één onderverdeling méér dan bij 2-stage sampling. Dus om een voorbeeld te geven: Men bepaalt binnen een concern door aselecte keuze een aantal fabrieken (stage 1), bij elk van die fabrieken een aantal dagen in het jaar (stage 2) en op elk van die dagen een aantal tijdstippen (stage 3). In de aangewezen fabrieken telt men nu op de aangewezen tijden het aantal werkende machines om een schatting te krijgen van het percentage van de tijd, dat een machine in dat jaar in gebruik is.

Dit voorbeeld valt echter niet onder hetgeen in het hier beschouwde artikel behandeld wordt over 3-stage sampling. Daarin wordt namelijk ondersteld dat elke steekproef in elk der 3 stages uit een oneindige populatie gekozen wordt, ter-

wijl dit in het voorbeeld bij het aantal fabrieken en het aantal dagen zeker niet het geval is. De resultaten in het eindige geval verschillen echter slechts weinig van die in het oneindige geval indien de uitgebreidheden van de populaties maar groot zijn.

In de generalisatie voor k-stage sampling wordt eveneens het oneindige geval behandeld. Bij k-stage sampling, en dus ook bij 3-stage sampling, verkrijgt men resultaten analoog aan die bij 2-stage sampling. Verder wordt nog ingegaan op k-stage sampling waarbij de grootten van de steekproeven in de laatste k_1 stages zijn voorgeschreven.

In het geval dat de populaties eindig zijn beschouwt SOPHIE MARCUSE een probleem van stratified sampling (dus $m=M$). Het belangrijkste speciale geval hierbij is meer volledig behandeld in het hieronder besproken artikel van J.NEYMAN.

3. J.NEYMAN, Lectures and conferences on mathematical statistics and probability, Department of Agriculture, Washington 1952, p.103-128.

Behandeld wordt een geval van stratified sampling, waarbij dus steekproeven uit alle sub-populaties worden genomen. De kosten per waarneming uit de i-de sub-populatie (stratum) bedragen c_i. De totale kosten C van het experiment kunnen dan geschreven worden als

$$C = \sum_{i=1}^{m} n_i c_i \qquad (12)$$

Er wordt hier ondersteld dat het aantal M $(=m)$ en de uitgebreidheden N_i van de subpopulaties eindig zijn.

Om bij gegeven toegelaten totale kosten C de spreiding in het totale gemiddelde $\bar{x}$ (gewogen met de aantallen elementen in de sub-populaties),

$$\bar{x} = \frac{\sum_{j=1}^{M} N_j \frac{\sum_{k=1}^{n_j} x_{jk}}{n_j}}{\sum_{i=1}^{M} N_i},$$

zo klein mogelijk te maken moet men voor n_i een gehele waarde kiezen die dicht ligt bij

$$n_i' = \frac{C N_i \sigma_i}{\sqrt{c_i} \sum_{j=1}^{M} N_j \sigma_j \sqrt{c_j}} \qquad i = 1,\ldots,M \qquad (13)$$

De variantie in het steekproefgemiddelde ligt dan dicht bij:

$$\sigma_{\bar{x}}'^2 = \frac{1}{\left(\sum_{i=1}^{M} N_i\right)^2} \sum_{j=1}^{M} N_j \sigma_j \left(\frac{\sqrt{c_j}}{C} \sum_{k=1}^{M} N_k \sigma_k \sqrt{c_k} - \sigma_j \right) \qquad (14)$$

Als de σ_i 's niet exact bekend zijn kan men als schatting van de variantie in het steekproefgemiddelde nemen:

$$s_{\bar{x}}^2 = \frac{1}{\left(\sum_{i=1}^{M} N_i\right)^2} \sum_{j=1}^{M} \frac{N_j(N_j - n_j)}{n_j(n_j - 1)} \sum_{k=1}^{n_j} (x_{jk} - \bar{x}_j)^2 \qquad (15)$$

Over de hier beschreven methode ("NEYMAN-sampling") kunnen de volgende opmerkingen worden gemaakt:

1. In het algemeen worden meer elementen van een sub-populatie getrokken naarmate de spreiding en het aantal elementen ervan groter en de kosten per waarneming eruit kleiner zijn. In het geval dat het bedrag C te klein is om alle sub-populaties in de steekproef te betrekken (wij hebben dan uiteraard niet meer met stratified sampling in de strikte zin van het woord te doen)en bovendien alle N_i en c_i gelijk zijn geldt dit echter zeker niet. Men kan dan het best juist van de sub-populaties met de kleinste spreidingen ieder één waarneming doen.

2. Indien van de elementen meerdere grootheden beschouwd worden kan men de n_i's het best berekenen met behulp van de varianties σ_i^2 van de belangrijkste grootheid. Vooral wanneer de varianties van de grootheden binnen de sub-populaties positief gecorreleerd zijn, zal men ook van de totale gemiddelden van de andere grootheden goede schattingen verkrijgen, hoewel deze in het algemeen natuurlijk niet de beste zijn die gegeven C. $\widehat{\text{bij}}$

3. Als de σ_i's niet bekend zijn moeten zij uit een voorlopig onderzoek geschat worden waarbij de aantallen waarnemingen uit alle sub-populaties evenredig met de N_i gekozen worden. Hieraan wordt dus een gedeelte van de toegelaten kosten C besteed. Het kan nu achteraf blijken dat van enkele sub-populaties bij dit voorlopig onderzoek reeds meer waarnemingen verricht zijn dan overeenkomt met formule (13). In dat geval berekent men eerst de kosten besteed aan alle waarnemingen van deze sub-populaties; laat dit bedrag C_v zijn. De totale kosten die nog besteed kunnen worden aan de waarnemingen uit de overige sub-populaties (dus met inbegrip van de daaraan bij het vooronderzoek reeds bestede bedragen) zijn

nu $C_r = C - C_v$. Men past nu formule (13) toe met C_r in plaats van C echter alleen voor die waarden van i, en men strekt de sommatie alleen uit over die waarden van j, welke corresponderen met de sub-populaties waarvan nog meer waarnemingen genomen moeten worden. Men vindt dan uiteindelijk de "beste" totale aantallen waarnemingen, waaruit, na aftrekking van de aantallen reeds verrichte waarnemingen, de aantallen nog te verrichten waarnemingen bepaald kunnen worden.

4. Zijn alle c_i's aan elkaar gelijk dan verkrijgt men de resultaten van F.N.DAVID, die hieronder apart zijn weergegeven.

4. F.N.DAVID, Probability theory for statistical methods, Cambridge Un.Press, 1949, p.179-187.

Een speciaal geval van stratified sampling wordt behandeld. De spreiding σ_i en de (eindige) uitgebreidheid N_i van elke sub-populatie zijn gegeven. De kosten per waarneming uit alle sub-populaties worden gelijk ondersteld.

Bij voorgeschreven totale omvang n van de steekproef (dus $n = \sum_{i=1}^{M} n_i$) is de spreiding in de schatting van het totaal gemiddelde zo klein mogelijk als voor n_i een gehele waarde genomen wordt dicht bij

$$n_i' = \frac{n N_i \sigma_i}{\sum_{j=1}^{M} N_j \sigma_j} \qquad i = 1, \ldots, M \qquad (16)$$

De variantie in het totale gemiddelde x ligt dan dicht bij

$$\sigma_x'^2 = \frac{\left(\sum_{i=1}^{M} N_i \sigma_i\right)^2 - n \sum_{j=1}^{M} N_j \sigma_j^2}{n \left(\sum_{k=1}^{M} N_k\right)^2} \qquad (17)$$

5. W.E. DEMING, Some theory of sampling, Wiley and Sons, New York, 1950, p. 135-165 en p. 213-246.

Hierin wordt o.a. het hierboven behandelde geval van SOPHIE MARCUSE beschouwd (zie § 2), nu echter voor een eindig aantal (M) eidige en even grote sub-populaties met uitgebreidheid $\bar{N}$. Alle σ_i^2 zijn hier wederom gelijk aan $\bar{\sigma}^2$. De totale kosten van het experiment worden evenals in § 2 voorgesteld door

$$C = m c_1 + m \bar{n} c_2 \qquad (18)$$

in welke formule c_1, c_2, m en $\bar{n}$ ook dezelfde betekenis hebben als in § 2.

Bij gegeven toegelaten C is de spreiding in de schatting van het totale gemiddelde x, $x = \frac{1}{m\bar{n}} \sum_{i=1}^{m} \sum_{j=1}^{\bar{n}} x_{ij}$, zo klein mogelijk als voor $\bar{n}$ een gehele waarde genomen wordt dicht bij

$$\bar{n}' = \sqrt{\frac{c_1}{c_2}} \frac{\bar{\sigma}}{\sqrt{\sigma_t^2 - \frac{\bar{\sigma}^2}{\bar{N}}}} \quad , \tag{19}$$

als voldaan is aan de voorwaarde

$$\bar{N}\sigma_t^2 > \bar{\sigma}^2 \quad . \tag{20}$$

m Wordt gevonden als een gehele waarde die zo goed mogelijk aan (18) voldoet. De variantie van het gemiddelde x ligt dan dicht bij

$$\sigma_x'^2 = \frac{\sigma_t^2}{C(M-1)} \left(M c_1 + M \bar{n} c_2 - C \right) + \frac{\bar{\sigma}^2 (\bar{N} - \bar{n})}{C(\bar{N}-1)\bar{n}} \left(c_1 + \bar{n} c_2 \right) \tag{21}$$

waarbij $\bar{n}$ de gehele waarde dicht bij $\bar{n}'$ voorstelt. Men neemt $\bar{n} = \bar{N}$ als $\bar{N}\sigma_t^2 \leq \bar{\sigma}^2$. Met behulp van formule (18) vindt men weer m en met formule (21), nu met $\bar{n} = \bar{N}$, de waarde van $\sigma_x'^2$

Bij voorgeschreven waarde S van de spreiding in het totale gemiddelde x worden de kosten zo klein mogelijk indien men voor $\bar{n}$ een gehele waarde neemt dicht bij

$$\bar{n}' = \sqrt{\frac{c_1}{c_2}} \frac{\bar{\sigma}}{\sqrt{\sigma_t^2 - \frac{\bar{\sigma}^2}{\bar{N}}}} \quad , \tag{22}$$

als voldaan is aan de voorwaarde

$$\bar{N}\sigma_t^2 > \bar{\sigma}^2 \quad . \tag{23}$$

m Wordt gevonden als een gehele waarde dicht bij

$$m' = \frac{M\left(\frac{\bar{N} - \bar{n}}{\bar{n}} \bar{\sigma}^2 + \bar{N}\sigma_t^2 \right)}{\bar{N}\left(MS^2 + \sigma_t^2 \right)} \tag{24}$$

De kosten zijn dan ongeveer gelijk aan

$$C' = \frac{M\left(\frac{\bar{N}-\bar{n}}{\bar{n}}\bar{\sigma}^2 + \bar{N}\sigma_t^2\right)}{\bar{N}\left(MS^2 + \sigma_t^2\right)}\left(c_1 + \bar{n}c_2\right) \qquad (25)$$

Zowel in (24) als in (25) stelt $\bar{n}$ weer de gehele waarde dicht bij $\bar{n}'$ voor.

Men neemt $\bar{n} = \bar{N}$ als $\bar{N}\sigma_t^2 \leq \bar{\sigma}^2$. De formules (24) en (25), waarbij nu $\bar{n} = \bar{N}$, blijven ook voor dit geval geldig.

Zijn $\bar{\sigma}$ en σ_t niet exact bekend, dan kan men als zuivere schatting van de variantie in het steekproefgemiddelde x gebruiken:

$$s_x^2 = \frac{M-m}{mM(m-1)}\sum_{i=1}^{m}(x_i - x)^2 + \frac{\bar{N}-\bar{n}}{m\bar{n}M\bar{N}(\bar{n}-1)}\sum_{i=1}^{m}\sum_{j=1}^{\bar{n}}(x_{ij} - x_i)^2 \qquad (26)$$

Als men in de formules (19), (21),(22),(24),(25) en (26) $\bar{N}$ en M naar oneindig laat gaan (aan de voorwaarden (20) en (23) wordt dan zeker voldaan), verkrijgt men, zoals te verwachten was, de overeenkomstige resultaten van SOPHIE MARGUSE (§2).

Naast de resultaten die NEYMAN verkreeg in het geval van stratified sampling (§3) leidt DEMING nog af, dat bij voorgeschreven waarde S^2 van de variantie in het totale gemiddelde x de kosten zo klein mogelijk worden als men voor n_i een gehele waarde neemt, dicht bij

$$n_i' = \frac{N_i\sigma_i\sum_{j=1}^{M}N_j\sigma_j\sqrt{c_j}}{\sqrt{c_i}\left\{S^2\left(\sum_{k=1}^{M}N_k\right)^2 + \sum_{\ell=1}^{M}N_\ell\sigma_\ell^2\right\}} \qquad i = 1, \ldots, M \qquad (27)$$

De kosten worden dan bij benadering: $C' = \frac{\left(\sum_{i=1}^{M}N_i\sigma_i\sqrt{c_i}\right)^2}{S^2\left(\sum_{j=1}^{M}N_j\right)^2 + \sum_{k=1}^{M}N_k\sigma_k^2}$ (28)

6. L.H.C. TIPPETT, The methods of statistics, 3rd ed., Williams and Norgate London, 1948, p. 207-209.

De voorwaarde welke hier opgelegd wordt aan de inrichting van een two-stage sampling experiment is de volgende: De totale tijd T, die voor het steekproefexperiment nodig is, moet onder een gegeven grens blijven. De tijd benodigd voor het in de steekproef betrekken van een sub-populatie zij gelijk aan kt en de extra tijd nodig voor het nemen van één element uit een sub-populatie gelijk aan t . Behoudens het verschil dat hier C door T, c_1 door kt en c_2 door t is vervangen, is het probleem hetzelfde als in § 2 behandeld is. De resultaten zijn dan ook, als men die veranderingen aanbrengt, geheel gelijk.

7. M.H. HANSEN, W.N. HURWITZ and W.G.MADOW, Sampling survey methods and theory I, Wiley and Sons New York, 1953.

Dit werk behandelt algemeen de steekproeftechniek, waarbij echter gedacht is aan een zodanige ontwikkeling van de theorie dat directe toepassingen op het gebied van bevolkingsonderzoek mogelijk zijn. O.a. wordt beschouwd:
p. 205-228, Stratified sampling uit eindige populaties; het behandelde komt overeen met NEYMAN (§3) en DEMING (laatste deel van § 5).
p. 284-302, Two-stage sampling uit eindige populaties, o.a. met kostenfunctie $C = mc_1 + m\bar{n}c_2$, zoals in het eerste deel van § 5 (DEMING) is weergegeven. Verder wordt hier nog een kostenfunctie gebruikt van een vorm die speciaal van belang is bij het houden van een enquête onder personen die over een groot gebied verspreid wonen: $C = c_0\sqrt{m} + c_1 m + c_2 m\bar{n}$ en in het geval van 3-stage sampling (p. 376-414) analoog:
$C = c_0\sqrt{m} + c_1 m + c_2 m\bar{n} + c_3 m\sqrt{\bar{n}} + c_4 m\bar{n}\bar{\bar{q}}$ waarbij $\bar{\bar{q}}$ het gemiddelde aantal elementen in stage 3 voorstelt, dat in de steekproefeenheid van stage 2 bevat is.
p. 314-335, Stratified 2-stage sampling uit eindige verzamelingen. Hierbij wordt binnen elke van de L strata (primaire subpopulaties) een 2-stage steekproef genomen. De kostenfactor c_1 uit § 2 (en het eerste deel van §5) wordt verschillend onderstelt voor elke stratum, voor stratum i wordt hij voorgesteld door c_{1i} . De factor c_2 wordt echter in alle strata gelijk ondersteld. De totale kosten C kunnen dan worden voorgesteld door $C = \sum_{i=1}^{L} c_{1i} m_i + c_2 \sum_{i=1}^{L} m_i \bar{n}_i$

In de laatste 3 gevallen, dus die met de speciale kostenfuncties en stratified 2-stage sampling, kunnen de optimale aantallen m of m_i, $\bar{n}$ of $\bar{n}_i$ en eventueel $\bar{\bar{q}}$ bepaald worden met behulp van een benaderingsproces.

8. Een aantal speciale artikelen.

Hieronder noemen wij nog een aantal artikelen die handelen over speciale zijden van het in dit overzicht beschouwde probleem van sub-sampling.
a) De indeling van de populatie. Het kan voorkomen dat de onderzoeker zelf de mogelijkheid heeft om te beslissen hoe de populatie onder te verdelen. Over de handelwijze die men dan het best kan volgen gaan de volgende 5 artikelen:

T. DALENIUS, The problem of optimum stratification, Skand. Akt. Tidsk. 33 (1950) p. 203-213, Statistics Upsala 5,

T. DALENIUS and M. GURNEY, The problem of optimum stratification II, Skand.Akt.Tidsk. 34 (1952), p. 133-148, Statistics Upsala 8.

Beide artikelen gaan over het probleem op welke wijze men de strata moet bepalen als het aantal strata gegeven is.

T. DALENIUS, The economics of one-staged stratified sampling, Sankhyā 12 (1953) p.351-356. Dit artikel behandelt de keuze van het aantal strata waarbij met kostenfactoren rekening gehouden wordt.

M.H. HANSEN and W.H. HURWITZ, On the theory of sampling from finite populations, Annals of Math. Stat. 14 (1943) p. 333-362.

De invloed van variatie in de uitgebreidheid van de sub-populaties in bevolkingsonderzoek wordt nagegaan.

Y. TAGA, On optimum balancing between sample size and number of strata in sub-sampling, Ann. of the Inst. of Stat.Math. 4 (1952-53) 95-102.

Dit artikel behandelt de bepaling van het optimale aantal strata en van de onderverdeling ervan bij stratified 2-stage sampling.

b) Vergelijking van NEYMAN-sampling met andere steekproefmethoden

In een aantal artikelen wordt "NEYMAN-sampling" (§3) vergeleken met het nemen van een aselecte steekproef uit de gehele populatie of met "evenredige" steekproeven uit de sub-populaties (evenredig met de uitgebreidheid van die sub-populaties). Daar de kosten niet in beschouwing worden genomen heeft men met een speciaal geval van NEYMAN te doen, nl. alle c_i's gelijk hetgeen onder § 4 (DAVID) behandeld is.

P. ARMITAGE, A comparison of stratified with unrestricted random sampling from a finite population, Biometrika 34 (1947) p. 273-280.

Vergelijkt NEYMAN - sampling met het nemen van een aselecte steekproef.

W.D. EVANS, On stratification and optimum allocations, Journal of the Amer.Stat. Association 46 (1951) p. 95-104.

Vergelijkt NEYMAN-sampling met het nemen van een evenredige steekproef. Gaat o.a. na hoe uitgebreid het vooronderzoek (om de varianties in de sub-populaties te schatten) minstens moet zijn om NEYMAN-sampling aantrekkelijker te doen zijn dan het nemen van een evenredige steekproef.

K. YONEDA, On the use of NEYMAN's allocation, Yokohama Math. Journal 1 (1953) p. 117-123.

Hierin wordt een criterium gegeven naar aanleiding waarvan men kan beslissen in de keuze tussen NEYMAN-sampling en het nemen van een evenredige steekproef.

c) Formules voor de spreiding. De afleiding van formules voor de spreiding in diverse gevallen van sub-sampling vormt met enkele andere onderwerpen de inhoud van de volgende artikelen.

J. CORNFIELD, On samples of finite populations, Journal of the Amer.Stat.Ass. 39 (1944)p. 236-239.

Dit artikel geeft de afleiding van enkele eenvoudige formules voor verwachting en spreiding o.a. in stratified sampling.

H.FAIRFIELD SMITH, Standard errors of means in sampling surveys with two stage sampling, Journal of the Roy. Stat.Soc. 110 (1947) p.257-259.

De spreiding in het totale gemiddelde $\bar{x}$ bij two-stage sampling in het eindige geval wordt afgeleid.

G.F. PEAKER, A sampling design used by the ministry of education, Journal of the Roy.Stat.Soc. 116 (1953) p. 140-166.

Dit artikel geeft de behandeling (en de formule voor de spreiding in het steekproefgemiddelde $\bar{x}$) van een voorbeeld van stratified two-stage sampling.

A.M. VANDEN ABEELE, Estimations sans biais, d'après échantillon, des composantes de la variance de populations dont les éléments font l'objet d'une classification simple ou multiple, Bull. de la Soc Belge de Stat. 2^2 (1953) p. 8-38.

Hierin worden o.a. formules gegeven (zonder bewijs) voor de

variantie in het steekproefgemiddelde $\bar{x}$ in het geval van two-stage sampling uit eindige populaties.

J.W. VOS, Nauwkeurigheid van gestratificeerde steekproeven, Statistische en econometrische onderzoekingen. CBS 9 (1954) p. 109-115.

Dit artikel behandelt de invloed, die afwijkingen van de optimale aantallen bij NEYMAN-sampling hebben op de variantie.

d) Diversen. Enkele artikelen over diverse onderwerpen, in een gedeelte waarvan onderstellingen worden gemaakt over de verdeling van de te schatten grootheid.

P.H. ANDERSON, Distributions in stratified sampling, Ann. of Math.Stat. 13 (1942) p. 42-52.

De verdelingen van het steekproefgemiddelde $\bar{x}$ en steekproefspreiding $s_{\bar{x}}$ worden gegeven voor gevallen waarbij men met stratified sampling uit bepaalde gegeven verdelingen te maken heeft.

S.G. GHURYE and H. ROBBINS, Two-stage procedures for estimating the difference between means, Biometrika 41(1954) p. 146-152.

Dit artikel behandelt de schatting van het verschil van twee gemiddelden bij optimaal two-stage sampling uit normale verdelingen.

J. PUTTER, Sur une méthode de double échantillonnage pour estimer la moyenne d'une population Laplacienne stratifiée, Revue de l'Inst. Int. de Stat 19 (1951) p. 231-238.

Beschouwd wordt het NEYMAN-geval met een iets gewijzigde strategie en de resultaten worden nagegaan indien de steekproef uit een normale verdeling afkomstig is.

F. YATES and P.M. GRUNDY, Selection without replacement from within strata with probability proportional to size, Journal of the Roy.Stat.Soc. B 15 (1953) p. 253-261.

In een geval van stratified two-stage sampling wordt uit iedere sub-populatie van een stratum hoogstens één waarneming gedaan. Aangetoond wordt dat de onzuiverheid die hierdoor in de schatting van het totale gemiddelde $\bar{x}$ en van de spreiding $s_{\bar{x}}$ in het steekproefgemiddelde ontstaat, gering is.