EEN EENVOUDIGE AFLEIDING VAN DE FORMULES, GEBRUIKT BIJ GELAAGDE BEMONSTERING VOLGENS TOEVAL

A SIMPLE DERIVATION OF THE FORMULAE USED IN STRATIFIED RANDOM SAMPLING

[524.63]

door/*by*

P. G. DE VRIES

(I.B.O. afd. Houtmeetkunde Landbouwhogeschool)

SUMMARY

This article contains derivations of the expressions for the estimated total and mean population values and their standard deviations in stratified random sampling. To this end use is made of the corresponding expressions in simple random sampling and of the rules for the standard deviations of a sum and of a product.

In de bosbouw wordt veelvuldig het principe van bemonstering toegepast, waarbij men dus aan de hand van metingen aan slechts een deel van de populatie conclusies omtrent het geheel trekt. Hoewel de moderne bemonsterings-theorie, (die pas gedurende de laatste twee decennia is ontwikkeld) uiteraard op een vrijwel onbeperkt aantal gebieden toepassing vindt, is zij voor de bosbouw wel van speciaal belang. Hier immers is bemonstering vaak een noodzaak, zoals bij de inventarisatie van uitgestrekte, moeilijk toegankelijke bosgebieden, terwijl in geval van technisch overigens uitvoerbare totale opnamen een aanzienlijke arbeidsbesparing kan worden verkregen, waarbij de mate van nauwkeurigheid van de gegevens omtrent het geheel met voldoende zekerheid bekend kan worden.

De literatuur over de bemonsterings-theorie en die van de toepassingen op het gebied van de bosbouw is inmiddels zeer uitgebreid geworden. Om enige kennis omtrent de achtergrond van veel gebruikte formules te verkrijgen, is het meer dan eens nodig, vrij moeilijk leesbare, in compacte symboliek gestelde verhandelingen door te werken, die de in principe geïnteresseerde wel eens afschrikken. In het onderstaande zal getracht worden de formules, die bij de algemeen toegepaste zogenaamde gelaagde bemonstering volgens toeval (*stratified random sampling*) gebruikt worden, op eenvoudige en overzichtelijke wijze af te leiden, waarbij de lezer verondersteld wordt bekend te zijn met de elementaire begrippen en werkwijzen uit de wiskundige statistiek.

Om de gedachten te bepalen wordt een bosareaal van A ha beschouwd, waarvan het totale stamtal door bemonstering moet worden vastgesteld, en wel door middel van volgens toeval (*at random*) over het areaal verdeelde

monster-vlakjes van a ha elk. Er kunnen zich twee gevallen voordoen:

1) het bosareaal kan, ondanks zijn natuurlijke variaties, zonder bezwaar als één homogeen type worden beschouwd. Men kan dan eenvoudige bemonstering volgens toeval (*simple random sampling*) toepassen;

2) het areaal valt uiteen in oppervlakten van verschillende bostypen, welke deel-oppervlakten op zichzelf homogeen zijn. Dit is bijvoorbeeld het geval wanneer in een moerasbosgebied tevens „drooglandbos" op hoger gelegen schollen voorkomt; men onderscheidt daar dan twee bostypen, of algemeen uitgedrukt: twee *strata* (enkelvoud *stratum* d.i. letterlijk „laag", hoewel de typen ruimtelijk *naast* elkaar voorkomen). Onder zulke omstandigheden winnen de bemonsteringsresultaten meestal aan nauwkeurigheid, indien men de strata afzonderlijk bemonstert (*stratified random sampling*), en wel is de winst des te groter, naarmate de strata duidelijker onderscheiden kunnen worden. Bij het vaststellen van de grenzen van dit soort strata speelt de luchtfoto een belangrijke rol.

Geval 1. Homogeen bosareaal. Eenvoudige bemonstering volgens toeval.

Het gehele bosgebied kan theoretisch worden verdeeld in $N = \frac{A}{a}$ monster-vlakjes, op elk waarvan het stamtal geteld zou kunnen worden. Men zou dan een serie van N stamtalwaarnemingen: $x_1, x_2, \ldots, x_i, \ldots, x_n$ verkrijgen, waarin x_i het stamtal in het i^e vakje is. Men trekt echter een monster van n (kleiner dan N) vlakjes, dus neemt men de stamtallen: $x_1, x_2, \ldots, x_i, \ldots, x_n$ waar (de indices 1, 2, ...i....n slaan nu op het monster), zodat als schatting van het gemiddelde stamtal per vlakje wordt gevonden:

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

met als standaard-deviatie van het monster:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (2)$$

en als standaard-deviatie van $\bar{x}$:

$$s_{\bar{x}} = \sqrt{\frac{N-n}{N}\,\frac{s^2}{n}} \qquad (3)$$

Hierin is (N—n)/N de „correctie voor eindige populatie" (c.e.p.; *finite population correction*), die bij volledige opname (namelijk als n = N) gelijk aan nul wordt. Inderdaad is dan $s_{\bar{x}} = 0$, terwijl s overgaat in σ, de standaard-deviatie van de gehele populatie. Meestal is n klein ten opzichte van N, zodat de c.e.p. practisch gelijk is aan 1.

Daar er in totaal N vlakjes zijn, met een geschat gemiddeld stamtal van $\bar{x}$, wordt het totale stamtal in het bosareaal geschat op

$$X = N.\bar{x} \qquad (4)$$

De standaard-deviatie S_x van X kan men berekenen met behulp van de formule voor de standaard-deviatie van een product. De standaard-deviatie

S_p van het product $p = y.z$, waarvan de termen y en z respectievelijk standaard-deviaties van s_y en s_z bezitten, is nl.:

$$S_p = \sqrt{(y.s_z)^2 + (z.s_y)^2} \qquad (5)$$

Het totale stamtal X kan nu worden beschouwd als het product van 2 termen: $\bar{x}$ (met st.dev. $s_{\bar{x}}$) en N (met st.dev. $^sN = 0$), zodat uit (5) volgt:

$$S_x = \sqrt{(N.s_{\bar{x}})^2} = N.s_{\bar{x}} = \sqrt{N(N-n)\frac{s^2}{n}} \qquad (6)$$

Uit X en S_x kunnen nu op de bekende wijze met behulp van „Student's t-tabel" de grenzen van het totale stamtal op het vereiste betrouwbaarheidsniveau gevonden worden:

Bovenste grens: $X_b = X + t.S_x$

Onderste grens: $X_0 = X - t.S_x$

Meestal echter gaat men als volgt te werk: men eist dat het absolute verschil tussen X en zijn betrouwbaarheidsgrenzen op het 95%-niveau niet meer dan een bepaald percentage (E%) van deze geschatte X mag bedragen, en vraagt dan, hoeveel random-monstervlakjes er geteld moeten worden om aan deze eis te voldoen. Dus:

$$X_b - X = X - X_0 = t_{95}.S_x \overset{<}{=} \frac{E}{100} X$$

of: $t^2_{95}.s^2_{\bar{x}} = t^2_{95}.\frac{s^2}{n} \overset{<}{=} (\frac{E}{100}\bar{x})^2$ (waarin c.e.p. = 1 is gesteld)

zodat:

$$n \overset{>}{=} \left(\frac{t_{95}.\ 100.s}{E.\ \bar{x}}\right)^2 \qquad (7)$$

Kiest men ruwweg $t_{95} = 2$ en $E = 20\%$, dan gaat (7) over in

$$n \overset{>}{=} \left(\frac{10.s}{\bar{x}}\right)^2 \qquad (8)$$

Uit een proefbemonstering ter grootte van n' vlakjes worden nu eerst s en $\bar{x}$ voorlopig geschat. Vindt men door invulling van deze s en $\bar{x}$ in (8) dat $n > n'$, dan breidt men het monster uit tot n vlakjes, waaruit opnieuw een schatting voor s en $\bar{x}$ volgt, welke waarden men opnieuw in (8) substitueert. Met deze iteratie vindt men dan meestal spoedig dat n uit (8) kleiner wordt dan de laatstgebruikte monstergrootte, zodat deze laatste dan aan de gestelde eis voldoet.

Geval 2. Bosareaal bestaat uit een aantal verschillende, op zichzelf homogene strata. Gelaagde bemonstering volgens toeval.

In elk stratum vindt eenvoudige bemonstering volgens toeval plaats. De resultaten daarvan worden vervolgens voor alle strata (het gehele bosareaal) samengevat.

Stel men onderscheidt in een A ha groot areaal F strata met oppervlakten $A_1, A_2, \ldots, A_h, \ldots, A_F$, zodat

$$A_1 + A_2 + \ldots + A_h + \ldots + A_F = A$$

In het willekeurige h^e stratum is het aantal mogelijke monstervlakjes (van a ha elk):

$$N_h = \frac{A_h}{a} \tag{9}$$

zodat:

$$N_1 + N_2 + \ldots + N_h + \ldots + N_F = N = \frac{A}{a} \tag{10}$$

Trekt men in het h^e stratum een monster van n_h vlakjes, dan is dus het totale aantal monstervlakjes over het gehele areaal:

$$n_1 + n_2 + \ldots + n_h + \ldots n_F = n \tag{11}$$

Stel de stamtalwaarnemingen in de n_h monstervlakjes van het h^e stratum zijn:

$$x_{h.1},\ x_{h.2},\ \ldots,\ x_{h.i},\ \ldots,\ x_{h.n_h}$$

dan is de schatting van het gemiddelde stamtal per vlakje in het h^e stratum:

$$\text{(cf.1)}\quad \bar{x}_h = \frac{\sum_{i=1}^{n_h} x_{h.i}}{n_h} \tag{12}$$

De standaard-deviatie van dit monster bedraagt (cf.2):

$$s_h = \sqrt{\frac{\sum_{i=1}^{n_h} (x_{h.i} - \bar{x}_h)^2}{n_h - 1}} \tag{13}$$

en de standaard-deviatie van $\bar{x}_h$ (cf.3):

$$s_{\bar{x}_h} = \sqrt{\frac{N_h - n_h}{N_h} \cdot \frac{s_h^2}{n_h}} \tag{14}$$

Als schatting van het totale stamtal X_h van het h^e stratum en de standaarddeviatie S_h daarvan, vindt men naar analogie van (4) en (6):

$$X_h = N_h \bar{x}_h \tag{15}$$

en

$$S_h = N_h s_{\bar{x}_h} \tag{16}$$

Als schatting van het totale stamtal X van het gehele bosareaal heeft men dan:

$$X = X_1 + X_2 + \ldots + X_h + \ldots + X_F = N_1 . \bar{x}_1 + N_2 . \bar{x}_2 + .. + N_h . \bar{x}_h + .. + N_F . \bar{x}_F$$

of:
$$X = \sum_{h=1}^{F} N_h . \bar{x}_h \qquad (17)$$

Als schatting van het gemiddelde stamtal per vlakje voor het gehele gebied heeft men:

$$\bar{X} = \frac{X}{N} \qquad (18)$$

De standaard-deviatie van S_x van X kan men berekenen als de standaard-deviatie van een som:

$$S_x = \sqrt{S_1^2 + S_2^2 + .. + S_h^2 + .. + S_F^2} \qquad (19)$$

waaruit men na substitutie van (16), (14) en (13) verkrijgt:

$$S_x = \sqrt{\sum_{h=1}^{F} \frac{N_h (N_h - n_h)}{n_h (n_h - 1)} \sum_{i=1}^{n^h} (x_{h.i.} - \bar{x}_h)^2} \qquad (20)$$

Ter verkrijging van de standaard-deviatie $S_{\bar{X}}$ van $\bar{X}$ moet volgens (5) S_X door N worden gedeeld.

Hiermede zijn de formules voor het totale en het gemiddelde stamtal en hun standaard-deviaties bij gelaagde bemonstering volgens toeval in hun meest algemene gedaante afgeleid. De formules voor bijzondere gevallen ($n_h << N_h$; $n_h = 2$; evenredige monsterverdeling) kunnen hieruit door eenvoudige substitutie worden verkregen.

Literatuur

Baten, W. D.; Elementary mathematical statistics. Chapman & Hall; London 1938.

Cochran, W. G.; Sampling Techniques. Chapman & Hall; London 1953.

Dawkins, H. C.; Experiments in low percentage enumerations of tropical high-forest. Emp. For. Rev. 31, 1952.

——; The extensive sampling of closed high-forest as developed in Uganda. Proc. IVth World For. Congr. Dehra-Dun, 1954.

F.A.O.; Survey methods of tropical forests. Rome 1961.

Husch, B.; Forest mensuration and statistics. Ronald Press Cy.; New York 1963.

Schumacher, F. X., and Chapman, R. A.; Sampling methods in forestry and range management. Duke Univ. School of Forestry, Bull. 7; 1954.
