

HET EFFECT VAN MEETFOUTEN OP DICHTHEDEN

C.A.J. Koeijers *)

Samenvatting

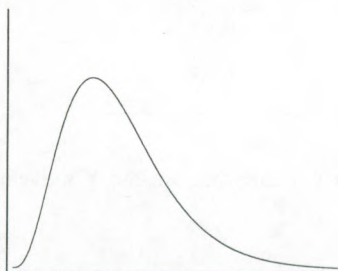
In dit artikel wordt met een gering aantal veronderstellingen getoond dat meetfouten bij kansdichtheden kunnen leiden tot afvlakking van de top(pen). Er worden twee voorbeelden gegeven. De theorie is te generaliseren voor meerdimensionale kansdichtheden.

1. Inleiding

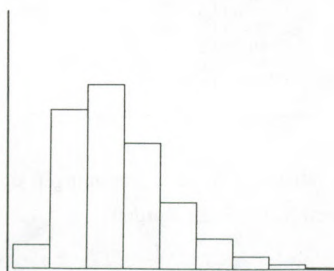
In de literatuur over steekproeftheorie, bijvoorbeeld Cochran (1977), wordt veel aandacht besteed aan het zo goed mogelijk schatten van parameters, zoals het gemiddelde en de variatie. Minder vaak wordt het schatten van kansdichtheden beschouwd zonder dat er direct sprake is van parameters. In dit artikel wordt het effect van meetfouten op kansdichtheden beschreven. Meer over dit onderwerp is te vinden in Chesher (1991).

De theorie wordt toegelicht aan de hand van voorbeelden afkomstig uit het Statistisch Jaarboek 1992. De voorbeelden zijn in de vorm van tabellen, maar ze zouden als histogrammen voorgesteld kunnen worden. Een histogram gebaseerd op een groot aantal waarnemingen is meestal bedoeld om een indruk te geven van een onbekende kansdichtheid. In figuur 1 en 2 is de overeenkomst tussen een kansdichtheid en een histogram nog eens geschetst.

Figuur 1. Kansdichtheid



Figuur 2. Histogram



*) Centraal Bureau voor de Statistiek, Hoofdafdeling Statistische Methoden, Postbus 959, 2270 AZ Voorburg. De in dit artikel weergegeven opvattingen zijn die van de auteur en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek.

Hoewel in dit artikel wordt uitgegaan van kansdichtheden, kan het effect dat wordt beschreven zich ook voordoen bij histogrammen.

2. Voorbeelden

Er worden twee voorbeelden gegeven. Ze hebben gemeen dat de variabele waarnaar wordt getabelleerd een continue variabele is die niet exact te meten is. In tabel 1 betreft dit het aantal ha cultuurgrond van een bedrijf en in tabel 2 het lichaamsgewicht van een dienstplichtige. De tabellen zijn te vinden in het Statistisch Jaarboek 1992 op p. 132 en 459. Om ze te laten aansluiten op figuur 2 zijn de relatieve frequenties toegevoegd.

Tabel 1. Land- en tuinbouwbedrijven naar grootteklasse cultuurgrond

	1990	relatieve frequentie
0,00 ha	2 714	0.022
0,01 tot 1 ha	10 046	0.080
1 tot 2 ha	9 945	0.080
2 tot 5 ha	18 715	0.150
5 tot 10 ha	20 896	0.167
10 tot 15 ha	14 149	0.113
15 tot 20 ha	11 079	0.089
20 tot 30 ha	16 959	0.136
30 tot 50 ha	14 390	0.115
50 tot 100 ha	5 390	0.043
100 ha en meer	681	0.005

Tabel 2. Dienstplichtigen naar lichaamsgewicht, 18-jarigen

	1990	relatieve frequentie
Gewicht	%	
≤ 49 kg	0.5	0.005
50-59 kg	10.2	0.102
60-69 kg	39.7	0.397
70-79 kg	33.8	0.338
80-89 kg	11.1	0.111
90-99 kg	3.1	0.031
100 kg en meer	1.5	0.015

Bij beide tabellen zijn de waarnemingen op te vatten als $Y + \text{meetfout}$, waarbij Y werkelijk getabelleerd had moeten worden.

3. Het effect van meetfouten

Om het effect van meetfouten te beschrijven wordt hier uitgegaan van een zeer eenvoudig model. Het doel is het meten van de waarde van een variabele Y , b.v. het aantal ha

cultuurgrond of het lichaamsgewicht. De waarde van Y is de werkelijke, maar onbekende waarde. In plaats daarvan komt de waarde van een variabele Z ter beschikking. Het verschil tussen de waarden van Y en Z is de meetfout. Er wordt verondersteld dat de verdeling van de meetfout verwachting 0 en variantie σ^2 heeft. De meetfout wordt bovendien onafhankelijk verondersteld van Y . Er wordt uitgegaan van continue stochastische variabelen. Samengevat:

$$Z = Y + \sigma W, \text{ waarbij } W \text{ onafhankelijk is van } Y \\ \text{en } E(W) = 0, V(W) = 1, \sigma \geq 0. \quad (1)$$

Direct kan worden opgemerkt dat uit (1) volgt dat $E(Z) = E(Y)$ en dat $V(Z) \geq V(Y)$, immers $V(Z) = V(Y) + \sigma^2$. Het is uiteindelijk de bedoeling om een uitspraak te doen over de kansdichtheid f_z van Z . Daartoe wordt eerst de simultane verdelingsfunctie van Z en W beschouwd:

$$F_{ZW}(z, w) = P\{Z \leq z, W \leq w\} \\ = P\{Z - \sigma W \leq z - \sigma w, W \leq w\} \\ = P\{Y \leq z - \sigma w, W \leq w\} \\ = F_{YW}(z - \sigma w, w). \quad (2)$$

Voor de bijbehorende simultane kansdichtheid kan nu geschreven worden:

$$f_{ZW}(z, w) = f_{YW}(z - \sigma w, w). \quad (3)$$

Omdat is aangenomen dat Y en W onafhankelijk zijn, is het rechterlid van (3) het product van de desbetreffende kansdichtheden. Dus geldt:

$$f_{ZW}(z, w) = f_Y(z - \sigma w) f_W(w). \quad (4)$$

Om de kansdichtheid f_z te vinden moet in formule (4) over w worden geïntegreerd, maar eerst wordt f_Y nog in een Taylorreeks ontwikkeld. De algemene vorm van de eerste drie termen van die reeks bij ontwikkelen rond een willekeurig punt y_0 luiden:

$$f_Y(y) \approx f_Y(y_0) + (y - y_0) f_Y^{(1)}(y_0) + (y - y_0)^2 f_Y^{(2)}(y_0)/2. \quad (5)$$

Hierin is $f_Y^{(j)}$ de j^{e} afgeleide van f_Y . Hoe dichter de punten y en y_0 bij elkaar liggen, hoe beter de benadering (5) in het algemeen zal voldoen. In (5) wordt nu ingevuld $y = z - \sigma w$ en $y_0 = z$. Als σ klein is dan liggen y en y_0 dicht bij elkaar. Gevonden wordt dan de zogenoemde 'small variance approximation' van f_Y . In formule:

$$f_Y(z - \sigma w) \approx f_Y(z) - \sigma w f_Y^{(1)}(z) + (\sigma w)^2 f_Y^{(2)}(z)/2. \quad (6)$$

Invullen van (6) in (4) leidt tot:

$$f_{ZW}(z, w) \approx [f_Y(z) - \sigma w f_Y^{(1)}(z) + (\sigma w)^2 f_Y^{(2)}(z)/2] f_W(w). \quad (7)$$

Aangezien $\int f_w(w)dw = 1$, $\int wf_w(w)dw = E(W) = 0$ en $\int w^2 f_w(w)dw = V(W) = 1$ geeft integratie van het rechterlid van (7) over w :

$$f_y(z) \approx f_y(z) + \sigma^2 f_y^{(2)}(z)/2. \quad (8)$$

Uit formule (8) blijkt dat $f_z(z)$ bij benadering gelijk is aan $f_y(z)$ als $f_y^{(2)}(z)$ gelijk is aan 0. Als f_y dus in een bepaald punt z een buigpunt heeft dan zullen $f_z(z)$ en $f_y(z)$ elkaar in de omgeving van het punt z niet veel ontlopen. Ook blijkt dat $f_z(z)$ kleiner is dan $f_y(z)$ bij negatieve waarden van $f_y^{(2)}(z)$. In gebieden waar f_y concaaf (bol) is liggen de waarden van f_z dus onder die van f_y . Het omgekeerde geldt in gebieden waar f_y convex (hol) is. Bij een een-toppige kansdichtheid wordt de top lager, terwijl de staarten dikker worden. Er treedt dus een soort erosie op.

Het bovenstaande impliceert voor een histogram dat de hoogste toppen in het algemeen zullen afvlakken ten gunste van de klassen met minder waarnemingen. Voor de tabellen van paragraaf 2 kan dus worden gesteld dat de bijbehorende werkelijke histogrammen, dus zonder meetfouten, waarschijnlijk hogere toppen hebben en dunnere staarten. Het aantal bedrijven met minder dan 1 ha cultuurgrond of het aantal dienstplichtigen dat meer dan 90 kg weegt wordt met de gegeven tabellen vermoedelijk enigszins overschat.

In het geval dat W een standaardnormale verdeling heeft, geldt bovendien dat $\int w^3 f_w(w)dw = 0$, $\int w^4 f_w(w)dw = 3$, $\int w^5 f_w(w)dw = 0$ en $\int w^6 f_w(w)dw = 15$, zie Kendall & Stuart (1977), p. 62. De Taylorreeksontwikkeling van formule (5) kan dan verder worden voortgezet, waardoor formule (8) kan worden uitgebreid tot:

$$f_z(z) \approx f_y(z) + \sigma^2 f_y^{(2)}(z)/2 + \sigma^4 f_y^{(4)}(z)/8 + \sigma^6 f_y^{(6)}(z)/48. \quad (9)$$

Uit (8) en (9) blijkt tenslotte dat bij een kleine waarde van σ het effect van meetfouten geringer zal zijn dan bij een grotere. In het triviale geval dat $\sigma = 0$ vallen f_z en f_y samen; Z en Y zijn dan identiek.

Slotopmerking

Het beschreven erosie-verschijnsel kan zich ook voordoen bij meerdimensionale kansdichtheden, zie hiervoor Chesher (1991). Hierin wordt ook een aantal toepassingen behandeld.

Referenties

- Chesher, A, 1991, The Effect of Measurement Error, *Biometrika* 78, 3, p 451-462.
 Cochran, W.G., *Sampling Techniques*, 3^e druk, 1977, Wiley, New York.
 Kendall, M. & Stuart, A., *The Advanced Theory of Statistics*, 1977, deel 1, Griffin, Londen.
 Statistisch Jaarboek 1992, CBS-publikatie, SDU/uitgeverij, 's-Gravenhage.