Inhoud

7. Functies van continue stochastische variabelen

7.1. Inleiding

Laat $\underline{x}$ een continue stochastische variabele zijn met dichtheidsfunctie f(x) en verdelingsfunctie $F(x) = \int_{-\infty}^{x} f(t)\,dt$; zoals bekend is $f(x) = F'(x)$. Stel dat aan elke realisatie x van $\underline{x}$ een getal $y = ax + b$ (waarin a en b gegeven constanten) of algemener $y = \varphi(x)$, wordt toegevoegd. Deze y is dan te beschouwen als de realisatie van een stochastische variabele $\underline{y}$, gedefinieerd als $\underline{y} = a\underline{x} + b$ respectievelijk $\underline{y} = \varphi(\underline{x})$. We bepalen de verdeling van $\underline{y} = a\underline{x} + b$ in de vorm van de verdelingsfunctie $G(y) = \Pr\{\underline{y} \le y\}$:

$$G(y) = \Pr\{a\underline{x} + b \le y\} = \Pr\left\{\underline{x} \le \frac{y-b}{a}\right\} = F\left(\frac{y-b}{a}\right)$$

althans indien $a > 0$ verondersteld wordt. Voor de dichtheidsfunctie van $\underline{y}$ geldt dus

$$g(y) = \frac{dG(y)}{dy} = \frac{1}{a} f\left(\frac{y-b}{a}\right)$$

Is in het bijzonder $\underline{x}$ normaal verdeeld met parameters μ en σ^2, dan

$$g(y) = \frac{1}{a}\,\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(\frac{y-b}{a} - \mu\right)^2\right) =$$

$$= \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2a^2\sigma^2}\,(y - a\mu - b)^2\right)$$

hetgeen wil zeggen dat $\underline{y} = a\underline{x} + b$ een normale verdeling heeft met gemiddelde $a\mu + b$ en variantie $a^2\sigma^2$. Merk op dat $\mu_y = a\mu_x + b$ en $\sigma_y^2 = a^2\sigma_x^2$. Gemiddelde en variantie van $\underline{y}$ kunnen dus berekend worden niet alleen via de verdeling van $\underline{y}$, maar ook rechtstreeks uit het gemiddelde en de variantie van $\underline{x}$. In de volgende paragrafen zullen hierover enige stellingen afgeleid worden.

7.2. De lineaire functie $\underline{y} = a\underline{x} + b$

Stel dat de stochastische variabele $\underline{x}$ de dichtheidsfunctie f(•) bezit, dan is het gemiddelde van $\underline{y} = a\underline{x} + b$ gelijk aan $E\underline{y} = E(a\underline{x} + b)$ en volgens de definitie van de verwachting is

$$E(a\underline{x} + b) = \int_{-\infty}^{\infty} (ax + b)\, f(x)\, dx = a\int_{-\infty}^{\infty} x\, f(x)\, dx + b\int_{-\infty}^{\infty} f(x)\, dx = a\, E\underline{x} + b$$

Dus: $E\underline{y} = a\, E\underline{x} + b$ of $\mu_y = a\mu_x + b$.
De variantie van $\underline{y}$ is $\mathrm{Var}\, \underline{y} = E(\underline{y} - \mu_y)^2$ waarin $\underline{y} - \mu_y = a(\underline{x} - \mu_x)$.
Dat betekent dat

$$E(\underline{y} - \mu_y)^2 = Ea^2(\underline{x} - \mu_x)^2 = \int_{-\infty}^{\infty} a^2(x - \mu_x)^2\, f(x)\, dx = a^2\, \mathrm{Var}\, \underline{x}$$

Dus: $\mathrm{Var}\, \underline{y} = a^2\, \mathrm{Var}\, \underline{x}$ of $\sigma_y^2 = a^2\sigma_x^2$.

Zonder in te gaan op de modificatie van het bewijs in de inleiding voor $a < 0$, vermelden we dat als $\underline{x}$ normaal verdeeld is, de verdeling van $\underline{y} = a\underline{x} + b$ met $a \neq 0$ ook normaal is.

Voorbeeld 7.1
De stochastische variabele $\underline{y}$ is normaal verdeeld met parameters μ en σ^2.
Beschouw $\underline{y} = (\underline{x} - \mu)/\sigma = (1/\sigma)\underline{x} - \mu/\sigma$. Dan is $\mu_y = (1/\sigma)\mu_x - \mu/\sigma = 0$ en $\sigma_y^2 = (1/\sigma^2)\sigma_x^2 = 1$. Bovendien is $\underline{y}$ normaal verdeeld. Met andere woorden: de stochastische variabele $\underline{y}$ heeft een standaard-normale verdeling. □

7.3. De functie $\underline{y} = \varphi(\underline{x})$

Op de verdeling van $\underline{y}$ wordt hier niet ingegaan, echter

$$E\underline{y} = \int_{-\infty}^{\infty} \varphi(x)\, f(x)\, dx$$

laat zich vaak bepalen, zo ook $\mathrm{Var}\, \underline{y}$. Soms kan men approximatief te werk gaan. Veronderstel namelijk dat de stochastische variabele $\underline{x}$ normaal verdeeld is met $\mu_x = \mu$ en $\sigma_x^2 = \sigma^2$. De uitkomsten van $\underline{x}$ liggen dan vrijwel altijd tussen $\mu - 3\sigma$ en $\mu + 3\sigma$ (behoudens een fractie van 0,26%). Stel dat dit interval voldoende klein is om de *continue* functie $\varphi(\bullet)$ over dit interval als een lineaire functie te beschouwen, namelijk

$$\varphi(x) \approx \varphi(\mu) + (x - \mu)\varphi'(\mu)$$

waarin $\varphi'(\mu)$ de eerste afgeleide van $\varphi(x)$ in het punt μ voorstelt. Op grond van het voorgaande volgt dan dat $\underline{y}$ normaal verdeeld is met

$$E\underline{y} \approx E\{\varphi(\mu) + \underline{x} - \mu)\varphi'(\mu)\} = E\{\varphi'(\mu)\underline{x} + \varphi(\mu) - \mu\varphi'(\mu)\} =$$

$$= \varphi'(\mu)\, E\underline{x} + \varphi(\mu) - \mu\varphi'(\mu) = \varphi(\mu)$$

$$\mathrm{Var}\ \underline{y} \approx \{\varphi'(\mu)\}^2\ \mathrm{Var}\ \underline{x} = \{\varphi'(\mu)\}^2\sigma^2$$

Voorbeeld 7.2

Stel $\varphi(x) = x^3$, dan: $E\underline{y} \approx \mu^3$ en $\mathrm{Var}\ \underline{y} \approx 9\mu^4\sigma^2$. □

Voorbeeld 7.3

Stel $\varphi(x) = \ln x$, dan: $E\underline{y} \approx \ln \mu$ en $\mathrm{Var}\ \underline{y} \approx \dfrac{\sigma^2}{\mu^2}$. □

Uiteraard moet in de voorbeelden σ klein zijn; deze voorwaarde lijkt irreëel, echter in paragraaf 8.1 komen we hierop terug.

7.4. De lineaire functie $\underline{y} = a_1\underline{x}_1 + a_2\underline{x}_2$

7.4.1. Gemiddelde en variantie van $\underline{y}$

Laten $\underline{x}_1$ en $\underline{x}_2$ stochastische variabelen zijn met dichtheidsfuncties $f_1(\bullet)$ en $f_2(\bullet)$, dus:

$$\Pr\{\underline{x}_i \le x_i\} = \int_{-\infty}^{x_i} f_i(t)\, dt \quad \text{voor } i = 1, 2$$

De formules voor i = 1, 2 geven de *marginale* verdelingen van $\underline{x}_1$ en $\underline{x}_2$, maar het zal duidelijk zijn dat hiermee nog niets bekend is over de verdeling van $\underline{x}_1$ en $\underline{x}_2$ *gezamenlijk*, dus over

$$\Pr\{\underline{x}_1 \le x_1 \cap \underline{x}_2 \le x_2\} \overset{\text{def}}{=} \Pr\{\underline{x}_1 \le x_1, \underline{x}_2 \le x_2\}$$

We stellen nu

$$\Pr\{\underline{x}_1 \le x_1, \underline{x}_2 \le x_2\} = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(s,t)\, ds\, dt$$

en hierin is f(•,•) de *simultane* of *gezamenlijke dichtheidsfunctie*[1] van $\underline{x}_1$ en $\underline{x}_2$ die moet voldoen aan

a. $f(x_1,x_2) \geq 0$ voor $-\infty < x_i < \infty$ (i = 1,2)

b. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2)\, dx_1\, dx_2 = 1$

Nu geldt voor alle waarden van x_1, dat $Pr\{\underline{x}_1 \leq x_1\} = Pr\{\underline{x}_1 \leq x_1, \underline{x}_2 < \infty\}$, hetgeen overeenkomt met

$$\int_{-\infty}^{x_1} f_1(s)\, ds = \int_{-\infty}^{x_1}\int_{-\infty}^{\infty} f(s,t)\, ds\, dt$$

Door naar x_1 te differentiëren vinden we

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1,t)\, dt$$

Geheel analoog is

$$f_2(x_2) = \int_{-\infty}^{\infty} f(s,x_2)\, ds$$

Met andere woorden: de marginale dichtheden $f_i(\bullet)$ met = 1, 2 volgen door integratie uit de simultane dichtheidsfunctie f(•,•).
De *verwachting* van een functie g(•,•) van $\underline{x}_1$ en $\underline{x}_2$ wordt nu gegeven door

$$Eg(\underline{x}_1,\underline{x}_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1,x_2)\, f(x_1,x_2)\, dx_1 dx_2$$

Op grond hiervan is het gemiddelde van $\underline{y} = a_1\underline{x}_1 + a_2\underline{x}_2$ gelijk aan

$$E\underline{y} = E(a_1\underline{x}_1 + a_2\underline{x}_2) =$$

---

[1] Het simultane gedrag van twee discrete stochastische variabelen $\underline{x}_1$ en $\underline{x}_2$ wordt beschreven door middel van een simultane kansfunctie $Pr\{\underline{x}_1 = x_1, \underline{x}_2 = x_2\} = p(x_1,x_2)$. Voor p(•,•) gelden dezelfde eisen als voor f(•,•), mits daartoe de integraties door sommaties worden vervangen.

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (a_1x_1 + a_2x_2)\, f(x_1,x_2)\, dx_1\, dx_2 =$$

$$= a_1\int_{-\infty}^{\infty} x_1 \int_{-\infty}^{\infty} f(x_1,x_2)\, dx_2 \cdot dx_1 + a_2\int_{-\infty}^{\infty} x_2 \int_{-\infty}^{\infty} f(x_1,x_2)\, dx_1 \cdot dx_2 =$$

$$= a_1\int_{-\infty}^{\infty} x_1\, f_1(x_1)\, dx_1 + a_2\int_{-\infty}^{\infty} x_2\, f_2(x_2)\, dx_2 = \quad = a_1\, E\underline{x}_1 + a_2\, E\underline{x}_2$$

Stellen we $E\underline{x}_i = \mu_i$ voor i = 1, 2, dan:

$$E\underline{y} = a_1\mu_1 + a_2\mu_2$$

De variantie van $\underline{y}$ wordt nu Var $\underline{y} = E(\underline{y} - \mu_y)^2$, waarin $\underline{y} - \mu_y = a_1(\underline{x}_1 - \mu_1) +$ $+ a_2(\underline{x}_2 - \mu_2)$, dus

$$\mathrm{Var}\,\underline{y} = E\{a_1(\underline{x}_1 - \mu_1) + a_2(\underline{x}_2 - \mu_2)\}^2 =$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{a_1(x_1 - \mu_1) + a_2(x_2 - \mu_2)\}^2 f(x_1,x_2)\, dx_1\, dx_2 =$$

$$= a_1^2\int_{-\infty}^{\infty} (x_1 - \mu_1)^2 f_1(x_1)\, dx_1 + a_2^2\int_{-\infty}^{\infty} (x_2 - \mu_2)^2 f_2(x_2) +$$

$$+ 2a_1a_2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_1 - \mu_1)(x_2 - \mu_2)\, f(x_1,x_2)\, dx_1\, dx_2 =$$

$$= a_1^2\, \mathrm{Var}\,\underline{x}_1 + a_2^2\, \mathrm{Var}\,\underline{x}_2 + 2a_1a_2\, E(\underline{x}_1 - \mu_1)(\underline{x}_2 - \mu_2)$$

De uitdrukking $E(\underline{x}_1 - \mu_1)(\underline{x}_2 - \mu_2)$ wordt de *covariantie* van $\underline{x}_1$ en $\underline{x}_2$ genoemd en aangegeven met $\mathrm{Cov}(\underline{x}_1,\underline{x}_2)$ zodat:

$$\mathrm{Var}\,\underline{y} = a_1^2\, \mathrm{Var}\,\underline{x}_1 + a_2^2\, \mathrm{Var}\,\underline{x}_2 + 2a_1a_2\, \mathrm{Cov}(\underline{x}_1,\underline{x}_2)$$

Merk op dat $\mathrm{Var}\,\underline{y} = a_1^2\, \mathrm{Var}\,\underline{x}_1 + a_2^2\, \mathrm{Var}\,\underline{x}_2$ indien $\mathrm{Cov}(\underline{x}_1,\underline{x}_2) = 0$.

7.4.2. De variantie van $\underline{y}$ als $\underline{x}_1$ en $\underline{x}_2$ onafhankelijk zijn

De stochastische variabelen $\underline{x}_1$ en $\underline{x}_2$ zijn *onafhankelijk*[2] als voor alle waarden van x_1 en x_2

$$\Pr\{\underline{x}_1 \le x_1, \underline{x}_2 \le x_2\} = \Pr\{\underline{x}_1 \le x_1\} \Pr\{\underline{x}_2 \le x_2\}$$

In dat geval geldt:

$$\int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(s,t)\, ds\, dt = \int_{-\infty}^{x_1} f_1(s)\, ds \int_{-\infty}^{x_2} f_2(t)\, dt$$

Partiële differentiatie naar x_1 en daarna naar x_2 levert nu op dat

$$f(x_1,x_2) = f_1(x_1)\, f_2(x_2)$$

voor alle waarden van x_1 en x_2, en dit houdt in dat

$$\begin{aligned} \mathrm{Cov}(\underline{x}_1,\underline{x}_2) &= E(\underline{x}_1 - \mu_1)(\underline{x}_2 - \mu_2) = \\ &= \int_{-\infty}^{\infty} (x_1 - \mu_1)\, f_1(x_1)\, dx_1 \int_{-\infty}^{\infty} (x_2 - \mu_2)\, f_2(x_2)\, dx_2 = \\ &= (E\underline{x}_1 - \mu_1)(E\underline{x}_2 - \mu_2) = 0 \end{aligned}$$

Voor onafhankelijke stochastische variabelen $\underline{x}_1$ en $\underline{x}_2$ is de variantie van $\underline{y} = a_1\underline{x}_1 + a_2\underline{x}_2$ gelijk aan

$$\mathrm{Var}\, \underline{y} = a_1^2\, \mathrm{Var}\, \underline{x}_1 + a_2^2\, \mathrm{Var}\, \underline{x}_2$$

Opmerking

Zonder verder bewijs wordt vermeld dat als $\underline{x}_1$ en $\underline{x}_2$ bovendien normaal verdeeld zijn, de stochastische variabele $\underline{y} = a_1\underline{x}_1 + a_2\underline{x}_2$ ook normaal verdeeld is.

7.5. De lineaire functie $\underline{y} = \sum_{i=1}^{n} a_i\underline{x}_i$

Stellen we $E\underline{x}_i = \mu_i$ en $\mathrm{Var}\, \underline{x}_i = \sigma_i^2$ voor = 1, 2, ..., n dan kunnen de in de vorige paragraaf afgeleide resultaten gemakkelijk gegeneraliseerd worden tot

a. $E\underline{y} = \sum_{i=1}^{n} a_i\mu_i$

[2] Populair gezien zijn $\underline{x}_1$ en $\underline{x}_2$ onafhankelijk als de uitkomsten van $\underline{x}_1$ niet worden beïnvloed door de uitkomsten van $\underline{x}_2$ en omgekeerd.

b. $\operatorname{Var} \underline{y} = \sum_{i=1}^{n} a_i^2 \sigma_i^2$

als $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$ *onderling onafhankelijke* stochastische variabelen zijn, dat wil zeggen:

$$\Pr\{\underline{x}_1 \le x_1, \underline{x}_2 \le x_2, ..., \underline{x}_n \le x_n\} = \Pr\{\underline{x}_1 \le x_1\} \ ... \ \Pr\{\underline{x}_n \le x_n\}$$

voor alle waarden van $x_1, x_2, ..., x_n$.
Bij continue stochastische variabelen $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$ komt deze onafhankelijkheid neer op de gezamenlijke dichtheidsfunctie

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)\, f_2(x_2) \ldots f_n(x_n)$$

waarin $f_i(\bullet)$ de marginale dichtheidsfunctie van $\underline{x}_i$ voorstelt.

c. Zijn $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$ onderling onafhankelijk en normaal verdeeld, dan bezit de stochastische variabele $\underline{y}$ ook een normale verdeling.

7.6. Bijzondere gevallen; toepassingen

7.6.1. Het verschil van 2 onafhankelijke stochastische variabelen

$\underline{x}_1$ en $\underline{x}_2$ zijn onafhankelijke stochastische variabelen met $E\underline{x}_i = \mu_i$ en $\operatorname{Var} \underline{x}_i = \sigma_i^2$ $(i = 1, 2)$. Het verschil $\underline{y} = \underline{x}_1 - \underline{x}_2$ is dan een stochastische variabele met gemiddelde $\mu_y = \mu_1 - \mu_2$ en variantie $\sigma_y^2 = \sigma_1^2 + \sigma_2^2$. Zijn $\underline{x}_1$ en $\underline{x}_2$ normaal verdeeld, dan is ook $\underline{y}$ normaal verdeeld.

Voorbeeld 7.4; passingsproblemen

In een motorschild zit een boring waarin een lager moet passen. Dat betekent dat de diameter van het lager kleiner moet zijn dan de diameter van de boring. Het verschil tussen beide diameters wordt speelruimte genoemd. Van een groot aantal boringen en lagers werd de diameter bepaald. Gevonden werd dat de diameter $\underline{x}_1$ van een boring normaal verdeeld is met een standaardafwijking van 2,5 μ (micron) en dat de diameter $\underline{x}_2$ van een lager normaal verdeeld is met een standaardafwijking van 1,65 μ. Bovendien bleek de diameter van een lager gemiddeld 5 μ kleiner te zijn dan de diameter van een boring. Hoe groot is de kans dat een aselect gekozen lager niet in een aselect gekozen boring past?
Het lager past niet in de boring als de speelruimte $\underline{y} = \underline{x}_1 - \underline{x}_2$ negatief is. De stochastische variabele is $\underline{y}$ normaal verdeeld met

$$\mu_y = \mu_1 - \mu_2 = 5 \quad \text{en} \quad \sigma_y = \sqrt{(2{,}5)^2 + (1{,}65)^2} = 3$$

Gevraagd wordt dus $\Pr\{\underline{y} < 0\}$, en deze kans is als $\underline{u}$ in standaard-normaal verdeeld is, gelijk aan

$$\Pr\{\underline{u} < \frac{0-5}{3}\} = \Pr\{\underline{u} > 1{,}67\} = 4{,}75\%$$

Opmerking

Laat gegeven zijn dat reeds bij een speelruimte van minder dan 3 μ moeilijkheden optreden bij het inbrengen van het lager in de boring. Dan is het percentage 'moeilijke passingen' gelijk aan

$$\Pr\{\underline{y} < 3\} = \Pr\left\{\underline{u} < \frac{3-5}{3}\right\} = \Pr\{\underline{u} > 0{,}67\} = 25{,}14\%$$

Als een fabrikant deze lagers koopt en het bovenstaande percentage te hoog vindt, zal hij dit alleen kunnen verlagen door de gemiddelde diameter van de boringen groter te maken. Om slechts 1% 'moeilijke passingen' te krijgen, moet hij er voor zorgen dat

$$\Pr\{\underline{y} < 3\} = 0{,}01 \quad \text{of} \quad \Pr\left\{\underline{u} < \frac{3-\mu_y}{3}\right\} = 0{,}01$$

Dit betekent: $(3 - \mu_y)/3 = -2{,}33$, dus $\mu_y = 10$ μ; het gemiddelde verschil tussen de diameters van de boringen en de diameters van de lagers moet 10 μ bedragen. □

7.6.2. De som van 2 onafhankelijke stochastische variabelen

$\underline{x}_1$ en $\underline{x}_2$ zijn onafhankelijke stochastische variabelen met $E\underline{x}_i = \mu_i$ en $\mathrm{Var}\ \underline{x}_i = \sigma_i^2$ (i = 1, 2). De som $\underline{y} = \underline{x}_1 + \underline{x}_2$ is dan een stochastische variabele met gemiddelde $\mu_y = \mu_1 + \mu_2$ en variantie $\sigma_y^2 = \sigma_1^2 + \sigma_2^2$. Zijn $\underline{x}_1$ en $\underline{x}_2$ normaal verdeeld, dan bezit ook $\underline{y}$ een normale verdeling.

Voorbeeld 7.5; optellen van toleranties

Bij de fabricage van een bepaald soort apparaten worden twee weerstanden van verschillend type (A en B) in serie gemonteerd. De weerstand van type A, aan te geven met $\underline{x}_1$, is normaal verdeeld met een gemiddelde van 1000 Ω en een standaardafwijking van 50 Ω. De weerstand van type B, $\underline{x}_2$, is normaal verdeeld met een gemiddelde van 2000 Ω en een standaardafwijking van 150 Ω. Bij de combinatie van beide weerstanden in een apparaat worden geen bij elkaar passende exemplaren uitgezocht. De totale weerstand $\underline{y} = \underline{x}_1 + \underline{x}_2$ is dus normaal verdeeld met $\mu_y = 3000\ \Omega$ en $\sigma_y = \sqrt{(50)^2 + (150)^2} = 158\ \Omega$.
Uit de tabel van de normale verdeling volgt dan dat deze totale weerstand in

99,9% van de gevallen, dus praktisch altijd, ligt tussen 3000 – 3,29·158 = 2480 Ω en 3000 + 3,29·158 = 3520 Ω.
Bij dit soort problemen wordt vaak op een andere manier geredeneerd: men stelt vast dat de weerstand A in 99,9% van de gevallen zal liggen tussen 835,5 Ω en 1164,5 Ω en dat deze grenzen voor weerstand B gelijk zijn aan 506,5 Ω en 2493,5 Ω en concludeert daaruit dat de weerstand van de combinatie praktisch altijd ligt tussen 835,5 + 506,5 = 2342 Ω en 64,5 + 2493,5 = 3658 Ω. Deze conclusie is natuurlijk juist, maar er zijn ook nauwere grenzen aan te wijzen, waartussen praktisch alle totale weerstanden liggen, de hierboven berekende waarden 2480 Ω en 3520 Ω. De redenering kan gevaarlijk worden als men hem omkeert. Als men bijvoorbeeld aan de totale weerstand de eis stelt dat deze praktisch altijd tussen 2500 Ω en 3500 Ω moet liggen (dat is in dit voorbeeld in 99,84% van de gevallen zo) en de weerstanden van type B — omdat ze ingekocht worden — heeft te accepteren zoals ze zijn, zou men tot de conclusie komen dat de weerstand van type A praktisch altijd zal moeten liggen tussen 993,5 Ω en 1006,5 Ω hetgeen bij een gemiddelde van 1000 Ω en een standaardafwijking van 50 Ω zeker niet het geval is. Er zou dan van deze weerstanden niet minder dan 89,66% uitgesorteerd moeten worden, en deze conclusie is dus onjuist. □

Voorbeeld 7.6

Een textielfabriek ontvangt garen op spincops. Men wil weten hoe sterk het netto-garengewicht op deze cops varieert. Daartoe wordt door wegen de standaardafwijking van het bruto-copsgewicht en die van het gewicht van de lege hulzen bepaald. We noemen deze standaardafwijkingen respectievelijk σ_B en σ_H en vinden de waarden $\sigma_B = 6{,}90$ g en $\sigma_H = 1{,}23$ g. Nu is het bruto-garengewicht $\underline{B}$ de som van het netto-garengewicht $\underline{G}$ en de lege huls $\underline{H}$ zodat als we de standaardafwijking van het netto-garengewicht σ_G noemen, geldt wegens de onafhankelijkheid van $\underline{G}$ en $\underline{H}$:

$$\sigma_B^2 = \sigma_G^2 + \sigma_H^2 \quad \text{of} \quad \sigma_G^2 = \sigma_B^2 - \sigma_H^2$$

Dit betekent dat

$$\sigma_G = \sqrt{(6{,}90)^2 - (1{,}23)^2} = 6{,}79$$

Omdat $\underline{G} = \underline{B} - \underline{H}$ is uiteraard $\sigma_G^2 = \sigma_B^2 + \sigma_H^2 - 2\,\mathrm{Cov}(\underline{B},\underline{H})$. Hierin is $\mathrm{Cov}(\underline{B},\underline{H}) = \mathrm{Cov}(\underline{G} + \underline{H},\underline{H}) = E(\underline{G} + \underline{H} - \mu_G - \mu_H)\,(\underline{H} - \mu_H) =$
$= \mathrm{Cov}(\underline{G},\underline{H}) + \sigma_H^2 = \sigma_H^2$ aangezien $\underline{G}$ en $\underline{H}$ onafhankelijk zijn.

Wederom is dus $\sigma_G^2 = \sigma_B^2 - \sigma_H^2$. □

7.6.3. De som en het gemiddelde van n onderling onafhankelijke en identiek verdeelde stochastische variabelen

Stel $\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$ zijn onderling onafhankelijke stochastische variabelen met dezelfde dichtheidsfunctie f(•). Zij $E\underline{x}_i = \mu$ en Var $\underline{x}_i = \sigma^2$ voor i = 1, 2, ..., n. Dus geldt:

$$\mu = \int_{-\infty}^{\infty} x\, f(x)\, dx \quad \text{en} \quad \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

De som $\underline{y} = \underline{x}_1 + \underline{x}_2 + ... + \underline{x}_n$ is dan een stochastische variabele met gemiddelde $\mu_y = n\mu$ en variantie $\sigma_y^2 = n\sigma^2$.

Gezien het behandelde in paragraaf 3.3 en de in paragraaf 7.5 gegeven definitie van onderling onafhankelijke stochastische variabelen, kunnen we concluderen dat de stochastische variabelen $\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$ een aselecte *steekproef* van n waarnemingen vormen uit een *populatie* met dichtheidsfunctie f(•). Voor deze stochastische variabelen, nogmaals onafhankelijk en identiek verdeeld, geldt dat de simultane dichtheidsfunctie $f(x_1,x_2,\ldots,x_n) = f_1(x_1)f_2(x_2)\ldots f_n(x_n) = \Pi_1^n f(x_i)$ voor alle $x_1,x_2,\ldots,x_n$. De stochastische variabele $\underline{y}$ is dus de *steekproefsom*, μ het *populatiegemiddelde* en σ^2 de *populatievariantie*. Is de steekproef nu afkomstig uit een normaal verdeelde populatie, dan betekent dit dat de waarnemingen $\underline{x}_i$ voor = 1, 2, ..., n normaal verdeeld zijn, hetgeen inhoudt dat de steekproefsom $\underline{y}$ ook een normale verdeling bezit. Met andere woorden:

a. De som van n waarnemingen die een aselecte steekproef vormen uit een populatie met gemiddelde μ en variantie σ^2, heeft een verdeling met gemiddelde (verwachting) $n\mu$ en variantie $n\sigma^2$.

b. Als de populatie een normale verdeling heeft, is ook de verdeling van de steekproefsom normaal.

Beschouwen we nu de stochastische variabele $\underline{\bar{x}} = \frac{1}{n}\underline{y}$ dan is $\underline{\bar{x}}$ het rekenkundig gemiddelde van de stochastische variabelen $\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$; we spreken van het *steekproefgemiddelde*. Hiervan is

$$E\underline{\bar{x}} = \frac{1}{n}\mu_y = \mu \quad \text{en} \quad \text{Var}\,\underline{\bar{x}} = \frac{1}{n^2}\sigma_y^2 = \frac{\sigma^2}{n}$$

Dus geldt:

a. Het rekenkundig gemiddelde van n waarnemingen die een aselecte steekproef vormen uit een populatie met gemiddelde μ en variantie σ^2 heeft een verdeling met gemiddelde (verwachting) μ en variantie σ^2/n.

b. Als de verdeling van de populatie waaruit de steekproef genomen is, normaal is, is de verdeling van het steekproefgemiddelde ook normaal.

Voorbeeld 7.7; betrouwbaarheidsinterval voor het gemiddelde van een normaal verdeelde populatie met bekende variantie.
Gegeven is, dat de efficiëntie (lichtsterkte per eenheid opgenomen vermogen) van TL-lampen in een partij normaal verdeeld is met standaardafwijking van 1,0 lumen per Watt. Van 5 aselect gekozen lampen uit deze partij wordt de efficiëntie bepaald en de resultaten zijn: 69,4; 67,1; 69,2; 68,7; 68,9. Het steekproefgemiddelde $\underline{\bar{x}} = 68{,}66$ lm/W.
Aangezien het hier een steekproef uit de partij betreft, zal het gemiddelde van de partij niet gelijk zijn aan 68,66. Als men dat onbekende gemiddelde μ noemt, heeft het steekproefgemiddelde $\underline{\bar{x}}$ gebaseerd op steekproeven van 5 exemplaren uit deze partij een normale verdeling met gemiddelde μ en standaardafwijking $1{,}0/\sqrt{5} = 0{,}45$. Volgens de tabel van de normale verdeling zal met een kans van 95% het steekproefgemiddelde $\underline{\bar{x}}$ liggen tussen $\mu - 1{,}96 \cdot 0{,}45 = \mu - 0{,}88$ en $\mu + 1{,}96 \cdot 0{,}45 = \mu + 0{,}88$. Dus met een kans van 95% voldoet $\underline{\bar{x}}$ aan

$$\mu - 0{,}88 < \underline{\bar{x}} < \mu + 0{,}88$$

Omkering van deze ongelijkheid geeft dat met kans van 95%:

$$\underline{\bar{x}} - 0{,}88 < \mu < \underline{\bar{x}} + 0{,}88$$

Als men nu hierin de gevonden waarde van $\underline{\bar{x}}$ substitueert, vindt men dat $67{,}78 < \mu < 69{,}54$. Deze uitspraak is juist of niet juist in die zin dat het interval 67,78 – 69,54 de onbekende μ wel of niet bevat. Bij iedere steekproef van 5 stuks zal het op deze wijze berekende *schattingsinterval* voor μ anders zijn aangezien steeds $\underline{\bar{x}}$ een andere uitkomst heeft. Omdat dan in 95% van de gevallen een interval gevonden wordt dat de onbekende μ bevat (dus een juiste uitspraak gedaan wordt) heet een specifiek interval een *95%-betrouwbaarheidsinterval* voor μ. Zou men in plaats van de factor 1,96 bijvoorbeeld 1,645 gebruikt hebben, dan zou men een 90%-betrouwbaarheidsinterval gevonden hebben. □

7.6.4. Momentenschatters

De in voorgaande paragraaf behandelde steekproefsom en steekproefgemiddelde zijn voorbeelden van *steekproeffuncties*; dit zijn functies van de onderling onafhankelijke en identiek verdeelde stochastische variabelen $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$. Op basis hiervan worden parameters van mathematische modellen geschat. Omdat $E\underline{\bar{x}} = \mu$, wordt de schatting van μ via $\underline{\bar{x}}$ zuiver (dat wil zeggen zonder systematische afwijking) genoemd; $\underline{\bar{x}}$ zelf heet dan een *zuivere schatter* voor μ. Deze schatter is onder meer bij de Poisson-verdeling aan de orde geweest. Beschouwen we heel algemeen een model met een parameter θ dan is $\mu = E\underline{x}$ een functie van θ, stel $\varphi(\theta)$. Is de functie $\mu = \varphi(\theta)$ omgekeerd te schrijven als $\theta = \psi(\mu)$, dan wordt de zogenoemde *momentenschatter* $\underline{\tilde{\theta}}$ voor θ gevonden door in de $\psi(\bullet)$ de schatter voor μ in te vullen, dus $\underline{\tilde{\theta}} = \psi(\underline{\bar{x}})$. Ga na dat bij de exponentiële verdeling de momentenschatter voor λ gelijk is aan $\underline{\tilde{\lambda}} = 1/\underline{\bar{x}}$.

Uiteraard is de variantie

$$\underline{s}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{x}_i - \underline{\bar{x}})^2$$

ook een steekproeffunctie, en van deze *steekproefvariantie* kan eenvoudig aangetoond worden dat deze een zuivere schatter voor σ^2 oplevert: $E\underline{s}^2 = \sigma^2$. Hiervan kan gebruik gemaakt worden wanneer een bepaald model twee parameters heeft, θ_1 en θ_2 te noemen. Dan zal in het algemeen gelden dat $\mu = E\underline{x} = \varphi_1(\theta_1,\theta_2)$ en $\sigma^2 = \mathrm{Var}\ \underline{x} = \varphi_2(\theta_1,\theta_2)$. Indien omgekeerd $\theta_1 = \psi_1(\mu,\sigma^2)$ en $\theta_2 = \psi_2(\mu,\sigma^2)$ worden de momentenschatters $\underline{\tilde{\theta}}_1$ en $\underline{\tilde{\theta}}_2$ voor θ_1 en θ_2 gevonden door substitutie van $\underline{\bar{x}}$ als schatter voor μ en s^2 als schatter voor σ^2 in de functies ψ_1 en ψ_2.

Voorbeeld 7.8

Beschouw de uniforme verdeling op het interval $[\theta_1,\theta_2]$ met $\theta_1 < \theta_2$. Dit wil zeggen

$$f(x) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1} & \text{als } \theta_1 \le x \le \theta_2 \\ 0, \text{ elders} \end{cases}$$

Door integratie volgt dat

$$\mu = \tfrac{1}{2}(\theta_1 + \theta_2) \quad \text{en} \quad \sigma^2 = \tfrac{1}{12}(\theta_2 - \theta_1)^2$$

zodat omgekeerd

$$\theta_1 = \mu - \sigma\sqrt{3} \quad \text{en} \quad \theta_2 = \mu + \sigma\sqrt{3}$$

Dus:

$$\underline{\tilde{\theta}}_1 = \underline{\bar{x}} - \underline{s}\sqrt{3} \quad \text{en} \quad \underline{\tilde{\theta}}_2 = \underline{\bar{x}} + \underline{s}\sqrt{3} \qquad \square$$

Opgemerkt dient te worden dat de methode geen zuivere schatters behoeft op te leveren. Zo is in bovenstaand voorbeeld $E\underline{\tilde{\theta}}_i = \mu \mp \sqrt{3}\ E\underline{s}$ $(i = 1, 2)$. Omdat $\text{Var}\ \underline{s} = E\underline{s}^2 - (E\underline{s})^2 > 0$ en $E\underline{s}^2 = \sigma^2$, is blijkbaar $E\underline{s} < \sigma$, waardoor $E\underline{\tilde{\theta}}_i \neq \theta_i$ $(i = 1, 2)$. De statistiek kent vele technieken voor het bepalen van schatters voor parameters!

7.6.5. Meest-aannemelijke schatters

Er wordt in deze paragraaf een schattingstechniek gegeven die in het algemeen voor grote steekproeven optimale resultaten geeft, de zogenaamde M(aximum) L(ikelihood)-methode. Het feit dat een model f(x) een parameter (eventueel een vector) bevat, noteren we expliciet als $f(x;\theta)$. De simultane dichtheid van de waarnemingen $\underline{x}_1,\ldots,\underline{x}_n$ in de steekproef wordt dus gegeven door $\Pi_1^n f(x_i;\theta)$, zijnde een functie van $x_1,\ldots,x_n$ en θ. Worden voor $x_1,\ldots,x_n$ de steekproefresultaten gesubstitueerd, dan resteert een functie uitsluitend in θ; deze functie heet de *Likelihood-functie* of *aannemelijkheidsfunctie* $L(\theta) = \Pi_1^n f(x_i;\theta)$. Een meest aannemelijke schatting voor θ is nu die waarde van θ die $L(\theta)$ maximaal maakt. Deze schatting wordt bepaald door de waarnemingsuitkomsten $x_1,\ldots,x_n$ en is daarom een steekproeffunctie; deze wordt met $\hat{\theta}$ aangegeven. Bij de exponentiële verdeling met $\theta = \lambda$ is bijvoorbeeld $f(x;\lambda) = \lambda e^{-\lambda x}$ als $\lambda > 0$ en $x > 0$. Dan is $L(\lambda) = \lambda^n e^{-\lambda\Sigma x_i}$, dus

$$\frac{dL}{d\lambda} = e^{-\lambda\Sigma x_i}(n\lambda^{n-1} - \lambda^n\Sigma x_i) = 0, \quad \text{als } \hat{\lambda} = \frac{n}{\Sigma x_i} = \frac{1}{\bar{x}}$$

Eenvoudig is in te zien dat $L(\hat{\lambda}) \geq L(\lambda)$ voor alle $\lambda > 0$. De ML-schatter voor λ is dus $\underline{\hat{\lambda}} = 1/\underline{\bar{x}}$.

Bij vrije extremen zijn $L(\theta)$ en $\ln L(\theta)$ maximaal voor dezelfde θ-waarde, en hiervan moet men gebruik maken omdat het maximaliseren van ln L eenvoudiger is. Zo is voor de exponentiële verdeling $\ln L = n \ln \lambda - \lambda\Sigma x_i$, dus

$$\frac{d \ln L}{d\lambda} = \frac{n}{\lambda} - \Sigma x_i = 0, \quad \text{als } \hat{\lambda} = \frac{1}{\bar{x}}$$

Voorbeeld 7.9

Beschouw de normale verdeling met $\theta = (\mu,\sigma^2)$. Dan

$$L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\ \Sigma(x_i - \mu)^2\right)$$

$$\ln L = -\frac{1}{2\sigma^2} \Sigma\ (x_i - \mu)^2 - \frac{n}{2} \ln \sigma^2 - \frac{n}{2} \ln 2\pi$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \Sigma\ (x_i - \mu) = 0, \quad \text{als } \widehat{\mu} = \overline{x}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \Sigma\ (x_i - \overline{x})^2 = 0, \quad \text{als } \widehat{\sigma}^2 = \frac{1}{n} \Sigma\ (x_i - \overline{x})^2$$

Ga na dat in $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma}^2)$ de functie $\ln L$ inderdaad maximaal is; $\underline{\overline{x}}$ en $\frac{1}{n}\,\Sigma(\underline{x}_i - \underline{\overline{x}})^2$ zijn dus de ML-schatters voor de parameters μ en σ^2 van de normale verdeling.

7.7. Opgaven

1. Een machine vult dozen, waarvan het gewicht $\underline{y}$ pond bedraagt, met $\underline{x}$ pond zout; $\underline{x}$ en $\underline{y}$ zijn onafhankelijk en normaal verdeeld met $\mu_x = 100$, $\sigma_x = 1$, $\mu_y = 5$, $\sigma_y = 0{,}5$. Welk percentage van de gevulde dozen zal een gewicht hebben tussen 104 en 106 pond?

2. De stochastische variabelen $\underline{x}_1$ en $\underline{x}_2$ zijn onafhankelijk en standaard-normaal verdeeld. Bereken de kans dat zowel $\underline{x}_1$ als $\underline{x}_2 > 0{,}5$ zijn; respectievelijk de kans dat minstens één van de twee $> 0{,}5$ is.

3. In een magazijn worden dozen op elkaar gezet. Deze zijn gemiddeld 10 cm hoog. De hoogte van de dozen is normaal verdeeld met een standaardafwijking van 1 cm. De beschikbare ruimte bedraagt 106 cm.
Wanneer men de dozen in aselecte volgorde opstapelt, bepaal dan de kansen
a) dat er voor een stapel van 10 dozen onvoldoende ruimte is.
b) dat er voor een stapel van 11 dozen wel voldoende ruimte is.
c) dat er voor een stapel van 10 dozen wel en voor de 11-de doos niet voldoende ruimte is.

4. Een bepaald soort projektielamp heeft een gemiddelde levensduur van 32,5 branduren en een standaardafwijking van 4 branduren.
De levensduren van de lampen zijn normaal verdeeld en onderling onafhankelijk. De bioscoop-exploitanten eisen dat de lampen minstens 30 uur zullen branden.
a) Hoeveel procent van de door de fabrikant afgeleverde lampen zal niet aan de eis voldoen?
b) 0p welke gemiddelde levensduur moet de fabrikant zijn fabricageproces instellen opdat 97,5% van de afgeleverde lampen aan de eis van de bioscoopexploitanten voldoet?

Deze herinstelling van het fabricageproces kost veel geld. De fabrikant tracht dit op te vangen door pakketten van n lampen af te leveren met de garantie dat elk pakket goed is voor minstens 30n branduren.
c) Hoeveel procent van de afgeleverde pakketten zal, indien elk pakket 4 lampen bevat. niet aan deze garantie voldoen?
d) 0p welke gemiddelde levensduur moet de fabrikant zich richten opdat bovengenoemd percentage 2,5% zal bedragen?
e) Hoe groot moet n zijn opdat 0,62% van de afgeleverde pakketten niet aan de garantie voldoen zonder dat de fabrikant zich op een andere dan de oorspronkelijke gemiddelde levensduur behoeft in te stellen.

5. In een weverij verwerkt men garen, dat op spoelen gewikkeld ontvangen wordt. Het nominale gewicht van het garen per spoel is 300 gram. Men weegt een zeer groot aantal spoelen met garen en vindt een gemiddeld gewicht van 405,2 gram en een standaardafwijking van 12,8 gram. Na het weven weegt men de lege spoelen. Daarbij vindt men een gemiddeld gewicht van 99,7 gram en een standaardafwijking van 11,9 gram. Men neemt aan, dat het gewicht van het garen op de spoel normaal verdeeld en onafhankelijk is van het gewicht van de spoelen, waar het op gewikkeld is.
a) Welk percentage der spoelen bevat minder garen dan het nominale gewicht aangeeft?
b) Hoe groot is de kans, dat het gemiddelde garengewicht van een steekproef van 5 spoelen kleiner zal zijn dan het nominale gewicht?

6. Van een grote partij assen is de diameter normaal verdeeld met gemiddelde μ_1 = 14,82 mm en standaardafwijking σ_1 = 0,03 mm. Van een even grote partij boringen is de diameter eveneens normaal verdeeld en wel met gemiddelde μ_2 = 14,89 mm en standaardafwijking σ_2 = 0,04 mm. Een as past in een boring als zijn diameter minstens 0,05 mm en hoogstens 0,15 mm kleiner is dan die van de boring.

a) Bereken welk percentage assen in boringen zal passen als as en boring aselect aan elkaar worden toegevoegd.

Men kan wél de gemiddelden van as en boring door verandering van instelling der machines wijzigen, maar niet de standaardafwijkingen.

b) Welke waarde kan voor het onder a genoemde percentage maximaal bereikt worden?

7. Een jamfabrikant vult zijn potten door een lege pot op een weegschaal te zetten en vervolgens jam in die pot te doen tot het gewicht van pot en jam samen 500 g is.
De fabrikant weet dat de gewichten van de lege potten normaal verdeeld zijn met gemiddelde 50 g en standaardafwijking 7,2 g en dat bij het wegen onafhankelijke en normaal verdeelde meetfouten optreden die een standaardafwijking van 12 g hebben en een gemiddelde 0.

a) Wat is de standaardafwijking van de netto-jamgewichten?

Aan de fabrikant is de eis gesteld dat bij hoogstens 5% van al de door hem gevulde potten het netto-jamgewicht minder dan 435 g bedraagt. Hij wil dat bereiken door de nauwkeurigheid van de meetapparatuur te vergroten.

b) Hoe groot mag de standaardafwijking van de meetfouten hoogstens zijn als aan de gestelde eis voldaan moet worden?

8. Bij de fabricage van thermostaten voor wasmachines worden beurtelings 4 aluminium strippen en 4 koperen plaatjes op elkaar gelast. De dikte van de aluminium strippen is normaal verdeeld met een gemiddelde van 2,0 mm en een standaardafwijking van 0,3 mm. Voor de dikte van de gelaste samenstelling gelden de tolerantiegrenzen 21,43 ± 1,30 mm. Deze dikte bleek bij onderzoek van een grote partij gelaste samenstellingen normaal verdeeld te zijn met een gemiddelde van 21,60 mm. Bovendien bezat 7,08% van de samenstellingen een te kleine dikte.

a) Bereken de standaardafwijking van de dikte van de gelaste samenstelling.

b) Hoeveel procent van de gelaste samenstellingen bezit een te grote dikte?

c) Bepaal het gemiddelde en de standaardafwijking van de dikte van de afzonderlijke koperen plaatjes.