Multivariate kansvariabelen

Sandra Van Aert

27 oktober 2011

Univariaat versus multivariaat

- **hoofdstuk 5: univariate kansvariabelen**
  met elke uitkomst ω van een experiment wordt een reëel getal $X(\omega)$ geassocieerd
- **hoofdstuk 9: multivariate kansvariabelen**
  met elke uitkomst ω van een experiment worden k getallen geassocieerd:

$$X_1(\omega), X_2(\omega), \ldots, X_k(\omega)$$

  voor de eenvoud: $X_1, X_2, \ldots, X_k$ of kansvector $\mathbf{X}_k$
  bivariate kansvariabelen: X en Y
  trivariate kansvariabelen: X, Y en Z

Gezamenlijke kansverdeling X en Y

$$\begin{aligned} p_{XY}(x,y) &= P\big[(X=x) \text{ en } (Y=y)\big] \\ &= P\big[(X=x) \cap (Y=y)\big] \end{aligned}$$

Eigenschappen

zoals bij univariate kansvariabelen:

- $0 \leq p_{XY}(x, y) \leq 1$
- $\sum_x \sum_y p_{XY}(x, y) = 1$
  som van alle kansen = 1

Marginale of onvoorwaardelijke kansverdeling

$$p_{XY}(x,y) \begin{array}{l} \nearrow p_X(x) \\ \\ \searrow p_Y(y) \end{array}$$

definitie

$$p_X(x) = \sum_y p_{XY}(x,y)$$

$$p_Y(y) = \sum_x p_{XY}(x,y)$$

voldoen aan voorwaarden voor kansverdeling bij univariate kansvariabelen

Onafhankelijke kansvariabelen

omgekeerd:

$$\begin{array}{ccl} p_X(x) \searrow & & \\ & p_{XY}(x,y)? & \\ p_Y(y) \nearrow & & \end{array}$$

gaat enkel voor onafhankelijke kansvariabelen

voorwaarde: onafhankelijkheid

voor elke waarde $(x, y) : p_{XY}(x, y) = p_X(x) \cdot p_Y(y)$

Voorwaardelijke kansverdeling

$$p_{Y|X}(y \mid x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

zodat

$$p_{XY}(x, y) = p_{Y|X}(y \mid x) \cdot p_X(x)$$

$$p_{X|Y}(x \mid y) = \frac{p_{XY}(x, y)}{p_Y(y)}$$

zodat

$$p_{XY}(x, y) = p_{X|Y}(x \mid y) \cdot p_Y(y)$$

Covariantie, correlatie, variantie van lineaire functie

Sandra Van Aert

27 oktober 2011

Algemeen

- covariantie

$$\sigma_{XY} = \operatorname{cov}(X, Y) = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) p_{XY}(x, y)$$

- correlatie

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- eigenschap

$$-1 \leq \rho_{XY} \leq +1$$

Covariantie: eigenschap

covariantie is speciaal geval van functie:

$$g(X,Y) = (X-\mu_X)(Y-\mu_Y)$$

dus:

$$\begin{aligned}
\sigma_{XY} &= E\left[(X-\mu_X)(Y-\mu_Y)\right] \\
&= E\left[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y\right] \\
&= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X\mu_Y \\
&= E(XY) - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y
\end{aligned}$$

(analogie met $\sigma_X^2 = E(X^2) - [E(X)]^2$)

Onafhankelijke kansvariabelen

indien X en Y onafhankelijk, dan

$$E(XY) = E(X)E(Y) = \mu_X \mu_Y$$

dus

$$\sigma_{XY} = \mu_X \mu_Y - \mu_X \mu_Y = 0$$

let op!

$$\text{onafhankelijk} \Rightarrow \text{covariantie } 0$$

$$\not\Leftarrow$$

Variantie van lineaire functie van meerdere kansvariabelen

algemeen

$$\mathrm{var}(aX+bY+c) = a^2\,\mathrm{var}(X)+b^2\,\mathrm{var}(Y)+2ab\,\mathrm{cov}(X,Y)$$
$$\mathrm{var}(X+Y) = \mathrm{var}(X)+\mathrm{var}(Y)+2\,\mathrm{cov}(X,Y)$$
$$\mathrm{var}(X-Y) = \mathrm{var}(X)+\mathrm{var}(Y)-2\,\mathrm{cov}(X,Y)$$

X **en** Y **onafhankelijk** $\Rightarrow \mathrm{cov}(X,Y)=0$

$$\mathrm{var}(aX+bY+c) = a^2\,\mathrm{var}(X)+b^2\,\mathrm{var}(Y)$$
$$\mathrm{var}(X+Y) = \mathrm{var}(X)+\mathrm{var}(Y)$$
$$\mathrm{var}(X-Y) = \mathrm{var}(X)+\mathrm{var}(Y)$$

Het schatten van populatieparameters

Sandra Van Aert

27 oktober 2011

Populatieparameters schatten

- populatiegemiddelde μ
  - gemiddelde bevolkingsdichtheid
  - gemiddelde zwaveldioxidepotentieel
- populatievariantie σ^2
  - variantie bevolkingsdichtheid
- populatieproportie π
  - percentage defecte producten

- doel: uitspraken doen over onbekende populatieparameters
- hoe? steekproefgegevens verzamelen $\rightarrow$ populatieparameters schatten

Schatting?

- functie van steekproefgegevens
- steekproefgemiddelde

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

- steekproefvariantie

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

Schatting?

- steekproefproportie

$$\hat{p} = \sum_{i=1}^{n} \frac{x_i}{n}$$

waarbij

$$\begin{cases} x_i = 1, & \text{indien succes} \\ x_i = 0, & \text{indien faling} \end{cases}$$

Schatter

- steekproef $x_1, x_2, \ldots, x_n$
- steekproefgemiddelde $\overline{x}$
- elke onderzoeker bekomt andere steekproefgegevens

- reden:
  trekken van steekproef, verzamelen van steekproefgegevens = *kansexperiment*

Schatter = kansvariabele

- steekproefwaarnemingen $X_1, X_2, \ldots, X_n$
- steekproefgemiddelde $\overline{X}$

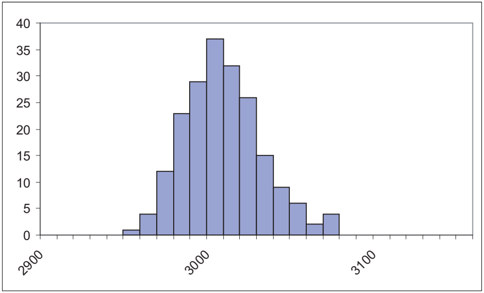
zijn kansvariabelen met

- een verwachte waarde
- een variantie
- een kansverdeling of -dichtheid

Voorbeeld

- bestuderen van normaal verdeelde populatie $N(3000, 100^2)$
- 200 studenten
- elk 20 metingen
- doel: centrale ligging schatten
  $\rightarrow$ populatiegemiddelde $\mu = 3000$
  $\rightarrow$ populatiemediaan $\gamma_{0.5} = 3000$
- hoe?
  $\rightarrow$ steekproefgemiddelde $\overline{X}$
  $\rightarrow$ steekproefmediaan Me
- Java-applet
  http://www.kuleuven.ac.be/ucs/java/
  (onder Basics / Distribution of mean / Continuous)

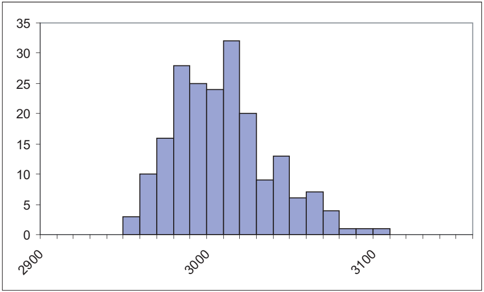
Vervolg voorbeeld



histogram van 200 steekproefgemiddeldes

- gemiddelde 200 steekproefgemiddeldes = 2999.93
- standaarddeviatie = 23.83

Vervolg voorbeeld



histogram van 200 medianen

- gemiddelde van 200 medianen = 2999.73
- standaarddeviatie van 200 medianen = 28.99
- steekproefgemiddelde en steekproefmediaan zijn zuivere of onvertekende schatters

Definitie zuivere schatter

als $\hat{\theta}$ een schatter is van θ en $E(\hat{\theta}) = \theta$, dan is $\hat{\theta}$ een zuivere of onvertekende schatter.

voorbeeld: $E(\overline{X}) = \mu$

bemerk: $\hat{\theta}$ is klassieke notatie voor schatter van onbekende populatieparameter θ

Efficiënte schatter

- wat zien we nog?
- steekproefgemiddelde zit vaakst in de buurt van 3000
- histogram van steekproefmediaan valt breder uit
- gevolg: steekproefgemiddelde heeft kleinere variantie dan steekproefmediaan
- met andere woorden: steekproefgemiddelde biedt preciezere informatie over centrale ligging dan steekproefmediaan
- daarom: $\overline{X}$ is een efficiëntere schatter dan Me

Gemiddelde gekwadrateerde afwijking (GGA)

- keuze tussen …
  …vertekende efficiënte schatter
  …onvertekende inefficiënte schatter
- kies schatter die

$$\text{GGA} = \text{var}(\hat{\theta}) + \underbrace{\left[E(\hat{\theta}) - \theta\right]^2}_{\text{vertekening}}$$

minimaliseert

I. Steekproefgemiddelde $\overline{X}$

- $\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$
- $E(\overline{X}) = \mu$

  onvertekende schatter van μ
- $\text{var}(\overline{X}) = \frac{\sigma^2}{n}$

  meest precieze lineaire onvertekende schatter (best linear unbiased estimator, BLUE)

Kansverdeling $\overline{X}$

- geval 1: normaal verdeelde populatie

  als $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$, dan kan aangetoond worden dat

  $$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

  ongeacht het aantal waarnemingen

- geval 2: niet-normaal verdeelde populatie (vb. uniform, exponentieel, binomiaal)

  als $X_1, X_2, \ldots, X_n \sim \cancel{N}(\mu, \sigma^2)$, dan is het niet meteen duidelijk welke kansdichtheid $\overline{X}$ heeft.

Kansverdeling $\overline{X}$: niet normaal verdeelde populatie

- kleine steekproeven

  geen algemeen antwoord

- grote steekproeven:

  centrale limietstelling $\Rightarrow \overline{X} \overset{\text{BEN.}}{\sim} N(\mu, \frac{\sigma^2}{n})$

Centrale limietstelling

als $X_1, X_2, \ldots, X_n$ onafhankelijke kansvariabelen met verwachte waarde μ en variantie σ^2,

dan is

$$\overline{X} = \frac{Y}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

benaderend normaal
met verwachte waarde $\mu_{\overline{X}} = \dfrac{\mu_Y}{n} = \dfrac{n\mu}{n} = \mu$

en variantie $\sigma^2_{\overline{X}} = \dfrac{\sigma^2_Y}{n^2} = \dfrac{n\sigma^2}{n^2} = \dfrac{\sigma^2}{n}$

(Stelling 11.3)

Kansverdeling $\overline{X}$: niet normaal verdeelde populatie

- kleine steekproeven

  geen algemeen antwoord

- grote steekproeven:

$$\text{centrale limietstelling} \Rightarrow \overline{X} \overset{\text{BEN.}}{\sim} N(\mu, \frac{\sigma^2}{n})$$

- wanneer is steekproef groot genoeg?
  - afhankelijk van oorspronkelijke kansverdeling of kansdichtheid
  - $n \geq 30$ is meestal voldoende

II. Steekproefproportie $\hat{P}$

- $\hat{P}$ = aantal "successen" in steekproef gedeeld door n
- $\hat{P} = \sum_{i=1}^{n} \frac{X_i}{n}$

  waarbij $X_i = \begin{cases} 1, & \text{indien succes} \\ 0, & \text{indien faling} \end{cases}$

  en dus X_i Bernoulli verdeeld met parameter π
- $\hat{P}$ is speciaal geval van steekproefgemidd. $\overline{X}$
- $E(\hat{P}) = \pi$
- $\text{var}(\hat{P}) = \frac{\pi(1-\pi)}{n}$

Kansverdeling of -dichtheid $\hat{P}$

- n groot: centrale limietstelling bij grote n

$$\begin{cases} n\pi \geq 5 \\ n(1-\pi) \geq 5 \end{cases}$$

$$\Rightarrow \hat{P} \overset{\text{BEN.}}{\sim} N(\pi, \frac{\pi(1-\pi)}{n})$$