

Statistiek voor AI

Uitwerkingen Inleveropgave 23 Oktober 2012

Jeroen Goudsmit

30 oktober 2012

- 1 Men wil met behulp van een kleine steekproef bepalen voor hoeveel procent van de volwassenen in een bepaalde stad het gemiddelde bruto inkomen per maand lager dan modaal (2.546 - CPB 2012) is. Neem aan dat dat in werkelijkheid 40% is.

Een onderzoeksbureau laat willekeurige telefoonnummers in het dorp genereren (eenzelfde telefoonnummer mag dus meerdere keren voorkomen) en laat zo'n nummer door een telefoniste bellen die bij diegene die opneemt informeert of diens bruto maandinkomen lager dan modaal is (we nemen voor het gemak aan dat er alleen volwassenen opnemen).

Zij X de stochast bij het éénmaal bellen ($X = 1$ als het inkomen onder modaal is en $X = 0$ anders). Stel dat er drie telefoontjes worden gepleegd.

- 1a Geef de verdeling van stochast $\bar{X}_3$.

Antwoord De stochast $\bar{X}_3$ neemt de waarden $\frac{0}{3} = 0$, $\frac{1}{3}$, $\frac{2}{3}$ en $\frac{3}{3} = 1$ aan. De verdeling is als volgt, voor $k = 0, 1, 2, 3$.

$$P\left(\bar{X}_3 = \frac{k}{3}\right) = \binom{3}{k} \frac{2^k \cdot 3^{3-k}}{5^3}$$

Dit volgt direct uit de omschrijving op slide 24 van het college op 16 oktober. De precieze waarden zijn als volgt:

$$\begin{aligned} P(\bar{X}_3 = 0) &= \frac{27}{125} = 0.216 \\ P(\bar{X}_3 = \frac{1}{3}) &= \frac{54}{125} = 0.432 \\ P(\bar{X}_3 = \frac{2}{3}) &= \frac{36}{125} = 0.288 \\ P(\bar{X}_3 = 1) &= \frac{8}{125} = 0.064 \end{aligned}$$

- 1b Wat is het gemiddelde van $\bar{X}_3$ en wat de variantie? Wordt het gemiddelde van $\bar{X}_3$ als waarde aangenomen in de verdeling?

Antwoord Het gemiddelde van $\bar{X}_3$ is te berekenen als volgt:

$$E(\bar{X}_3) = \sum_{k=0}^3 P(\bar{X}_3) \cdot \frac{k}{3} = \frac{1}{3125} \cdot (54 + 2 \cdot 36 + 3 \cdot 8) = \frac{150}{375} = \frac{2}{5}$$

Deze waarde wordt nooit aangenomen door de verdeling. De variantie is de kwadratische afwijking van het gemiddelde, d.w.z.

$$\begin{aligned} \text{Var}(\bar{X}_3) &= E\left((\bar{X}_3 - E(\bar{X}_3))^2\right) \\ &= \frac{1}{4} \cdot \left(0.216 \cdot (0 - 1.2)^2 + 0.432 \cdot \left(\frac{1}{3} - 1.2\right)^2 + 0.288 \cdot \left(\frac{2}{3} - 1.2\right)^2 + 0.064 \cdot (1 - 1.2)^2\right) \\ &= 0.08 \end{aligned}$$

In plaats van deze complexe berekeningen hadden we de waarden ook kunnen aflezen van de formules uit slide 24 van het college op 16 oktober, waar we weten $p = \frac{2}{5}$ en $n = 3$.

$$E(\bar{X}_3) = p = \frac{2}{5}$$

$$\text{Var}(\bar{X}_3) = \frac{p \cdot (1-p)}{n} = \frac{\frac{2}{5} \cdot \frac{3}{5}}{3} = \frac{6}{25 \cdot 3} = \frac{2}{25} = 0.08$$

- 2 Over een 10-zijdige dobbelsteen wordt je verteld dat op elke zijde het getal 1 of het getal 2 staat. Je mag vijf maal met de dobbelsteen gooien en op grond daarvan raden wat de kans is dat je met deze dobbelsteen 2 gooit. Stel dat in werkelijkheid op precies acht zijdes het getal 2 staat en dat elke zijde gelijke kans heeft. Zij X de stochast met als mogelijke waarden de uitkomsten bij het éénmaal gooien van de dobbelsteen.

- 2a Wat zijn de waarden en de verdeling van X en van $\bar{X}_5$?

Antwoord De waarden die $\bar{X}_5$ aan kan nemen zijn de volgende waarden, voor $k = 0, \dots, 5$:

$$\frac{1 \cdot k + 2 \cdot (5 - k)}{5} = \frac{1 \cdot k + 2 \cdot 5 - 2 \cdot k}{5} = 2 - \frac{k}{5}.$$

Oftewel, de waarden 5, 5.2, 5.4, 5.6, 5.8, 6. De verdeling is weer binomiaal, en deze keer geldt

$$P(\bar{X}_5 = 5 + k) = \binom{5}{k} \frac{1^k \cdot 4^{5-k}}{5^5}.$$

Uitkomsten hiervan zijn

$$\begin{aligned} P(\bar{X}_5 = 5) &= 0.00032 \\ P(\bar{X}_5 = 5.2) &= 0.00640 \\ P(\bar{X}_5 = 5.4) &= 0.05120 \\ P(\bar{X}_5 = 5.6) &= 0.20480 \\ P(\bar{X}_5 = 5.8) &= 0.40960 \\ P(\bar{X}_5 = 6) &= 0.32768. \end{aligned}$$

- 2b Leg uit waarom $\bar{X}_n$ geen goede benadering is van de kans dat je met deze dobbelsteen 2 gooit, zelf niet als n groot wordt.

Antwoord De waarden die $\bar{X}_n$ aanneemt vallen geheel buiten het interval $[0, 1]$.

- 2c Voor welke stochast Y is het zo dat $\bar{Y}_n$ een redelijke benadering is van de kans dat je met deze dobbelsteen 2 gooit bij grote n . En waarom is dat zo?

Antwoord De stochast Y die waarde 1 aanneemt indien je 2 gooit, en 0 indien je 1 gooit. Dit komt door de “Centrale Limietstelling”, als uitgelegd op slide 20 van het hoorcollege op 18 oktober.

- 3 De steekproef in het onderzoek van Floor, Minke en Annemarie had grootte 25. Het gemiddelde aantal uren dat mensen in de steekproef sporten per week bleek 4.12, met een variantie van 12.03. Zij μ het ware gemiddelde aantal uren dat CKI-ers in Utrecht (de populatie) sporten per week.

Hoe groot is de kans minstens dat bij een willekeurige steekproef ter grootte 25 het steekproefgemiddelde $\bar{X}_{25}$ niet meer dan 4 afwijkt van μ ? Je mag aannemen dat de variantie in de populatie gelijk is aan die in de steekproef, 12.03.

Antwoord Hier willen we dus een ondergrens weten van $P(|\bar{X}_{25} - 4.12| \leq 4)$. Dit volgt uit de “Wet van de Grote Getallen”, als uitgelegd op [slide 12 van het hoorcollege op 18 oktober](#), geeft ons:

$$P(|\bar{X}_n - E(\bar{X}_n)| < \epsilon) \geq 1 - \frac{\text{Var}(X)}{n\epsilon^2}.$$

Als we nu het bovenstaande lezen met $\epsilon = 4$, $n = 25$, $\text{Var}(X) = 12.03$ en $E(\bar{X}_n) = \mu$ dan verkrijgen we:

$$P(|\bar{X}_{25} - \mu| < 4) \geq 1 - \frac{12.03}{25 \cdot 4^2} \approx 0.97.$$

De kans is dus minstens 0.97.

- 4 In de steekproef van Renate en Marleen is het gemiddelde aantal uren studie per week inclusief colleges 23.56 uur. Stel dat de variantie in de populatie (CKI-ers in Utrecht) 5 uur is en μ het ware gemiddelde.

De steekproef in het onderzoek van Renate en Marleen had grootte 25. Is bij een willekeurige steekproef van die grootte de kans dat het steekproefgemiddelde meer dan 3 uur afwijkt van het ware gemiddelde niet groter dan 0.1? En dezelfde vraag voor 0.01 in plaats van 0.1.

Antwoord Ook hier gebruiken we de wet van de grote getallen, die ons geeft dat

$$P(|\bar{X}_n - E(\bar{X}_n)| \geq \epsilon) \leq \frac{\text{Var}(X)}{n\epsilon^2}.$$

We vullen in voor $\epsilon = 3$, $n = 25$, $\text{Var}(X) = 5$ en $E(\bar{X}_n) = \mu$. Dit geeft

$$P(|\bar{X}_{25} - \mu| \geq 3) \leq \frac{5}{25 \cdot 3^2} \approx .022.$$

De kans op een afwijking van het ware gemiddelde die groter is dan 3 is dus hooguit 0.022, wat zeker kleiner is dan 0.1. Dus bij een willekeurige steekproef van grootte 25 is de kans dat het steekproefgemiddelde meer dan 3 uur afwijkt van het ware gemiddelde niet groter dan 0.1. We weten helaas niets over of hetzelfde zo is met 0.01 in plaats van 0.1, omdat 0.01 onder de berekende bovengrens ligt.