Tentamen Kanstat
WI1275TA
27 augustus 2007 14.00 – 17.00 uur

Bij dit examen is het gebruik van een (evt. grafische) rekenmachine toegestaan. Een formuleblad wordt bijgeleverd. Een antwoord alleen is *niet* voldoende: er dient een berekening, toelichting en/of motivatie aanwezig te zijn. Dit alles goed leesbaar en in goed Nederlands (of Engels).

---

1. Arend krijgt als troostprijs voor de verloren finale van de door de winkeliers van Zilverdam georganiseerde Tringowedstrijd tien loten voor de komende maandelijkse loterijen. In deze loterijen worden altijd duizend loten verkocht, en er is één prijs te winnen, namelijk duizend euro. Arend heeft de keus om met alle tien de loten in een en dezelfde loterij mee te spelen, of om in tien verschillende loterijen met één lot mee te doen. De opbrengst bij de eerste keus noemen we X, die bij de tweede keus Y.

a. Geef de kanstabellen van X en van Y

b. Bereken de verwachting van X en van Y.

c. Bereken de variantie van X en van Y.

2. Gegeven is de dataset

$$42 \quad 48 \quad 72 \quad 57 \quad 51 \quad 62 \quad 58 \quad 50$$

Bereken achtereenvolgens

a. De mediaan en het gemiddelde.

b. De MAD en de steekproefstandaardafwijking. Geef aan *hoe* deze gedefinieerd zijn (en laat het rekenwerk evt. over aan je rekenmachine).

c. $F_8(60)$ (De waarde van de empirische verdelingsfunctie in het punt 60.)

3. We trekken een willekeurig punt uit een rechthoek met afmetingen acht bij twaalf. D is de afstand tot de rand van de rechthoek. Maak een schets!

a. Welke waarden kan D aannemen?

b. Toon aan dat voor de in het vorige onderdeel genoemde waarden a geldt: $\mathrm{P}(D \leq a) = (10\,a - a^2)/24$.

c. Wat denk je, op grond van je schets, wat ongeveer de waarde van de verwachting van D is?
Bereken de verwachting van D. Is dit in overeenstemming met je schatting?
Zo nee, wat is er aan de hand?

4. Stel Z_1, Z_2 en Z_3 zijn drie onafhankelijke normale variabelen, met parameters $\mu = 1$ en $\sigma = 3$. Stel $S = X_1 + X_2 + X_3$ en $V = X_1 - X_2 - X_3$.

a. Bereken $\mathrm{P}(Z_1 \leq 4)$.

b. Bereken $\mathrm{P}(S \leq 4)$.

c. Bereken $\mathrm{Cov}(S, V)$.

5. Winkeliers ronden al jaren bedragen af op veelvouden van vijf eurocent. De individuele afrondfouten modelleren we met een stochast X die met kansen $\frac{1}{5}$ een van de waarden -2, -1, 0, 1 of 2 aanneemt.

a. Bereken $\mathrm{Var}(X)$.

b. De totale afrondfout op een dag met 500 klanten is

$$S_{500} = X_1 + X_2 + \ldots + X_{500}.$$

Bereken de verwachting en de variantie van S_{500}. (We nemen aan dat de afrondingen onafhankelijk zijn van elkaar zijn.)

c. Welke bovengrens geeft de ongelijkheid van Chebychev voor de kans $\mathrm{P}(|S_{500}| \geq 50)$?

6. Gegeven is de dataset 1.09 1.29 2.78 1.08 1.61
We nemen aan dat dit een realizatie is van een steekproef is uit een Pareto-verdeling met (onbekende) parameter α. We willen de maximum-likelihood-schatting van α berekenen.

Dichtheid en verdelingsfunctie van de Pareto(α)-verdeling:

$$f(x) = \frac{\alpha}{x^{\alpha+1}}, \quad F(x) = 1 - \frac{1}{x^\alpha} \quad (\text{ voor } x \geq 1\,)$$

a. Geef de likelihood- en de loglikelihoodfunctie.

b. Bereken met een van beide de maximum-likelihoodschatting $\hat{\alpha}$ van α.

c. Ga na in hoeverre de verwachting van de Par($\hat{\alpha}$)-verdeling overeenkomt met het gemiddelde van de dataset.

7. Om de nauwkeurigheid van een weeginstrument te onderzoeken wordt een serie metingen van een standaardgewicht van 10 gram uitgevoerd. De aannamen zijn dat de metingen normaal verdeeld zijn met (onbekende) verwachting μ, en standaardafwijking $\sigma = 0.0025$ gram.

a. Stel het ijkgewicht wordt tien keer gewogen en het gemiddelde van de tien metingen is 10.016 gram. Geef een 98% betrouwbaarheidsinterval voor μ.

b. Hoeveel metingen moeten er (minimaal) gedaan worden om een 98 % betrouwbaarheidsinterval te krijgen dat korter is dan 0.002 gram?

Antwoorden open vragen

1a Kanstabel voor X:

x	0	1000
$\mathrm{P}(X=x)$	99/100	1/100

$Y/1000$, het aantal keren dat het prijs is, heeft een binomiale verdeling met parameters $n = 10$ en $p = 1/1000$.
Anders gezegd: $Y = 1000Y_1$, waarbij $Y_1 \sim \mathrm{Bin}(10{,}0.001)$.

1b $\mathrm{E}[X] = 0 \cdot 99/100 + 1000 \cdot 1/100 = 10$.
$\mathrm{E}[Y] = 1000\mathrm{E}[Y_1] = 1000\text{"}np\text{"} = 1000 \cdot 10 \cdot 0.001 = 10$.

1c $\mathrm{E}[X^2] = 1000^2 \cdot 1/100 = 10000$, dus $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = 9900$.
$\mathrm{Var}(Y_1) = \text{"}npq\text{"} = 10 \cdot 1/1000 \cdot 999/1000 = 0.00999$,
en $\mathrm{Var}(Y) = 1000^2\mathrm{Var}(Y_1) = 9990$.

2a Het gemiddelde: $\bar{x}_8 = \frac{1}{8}(x_1 + x_2 + \cdots + x_8) = \frac{1}{8} \cdot 440 = 55$.
De dataset geordend: 42 48 50 51 57 58 62 72.
De mediaan: $m = \dfrac{51+57}{2} = 54$.

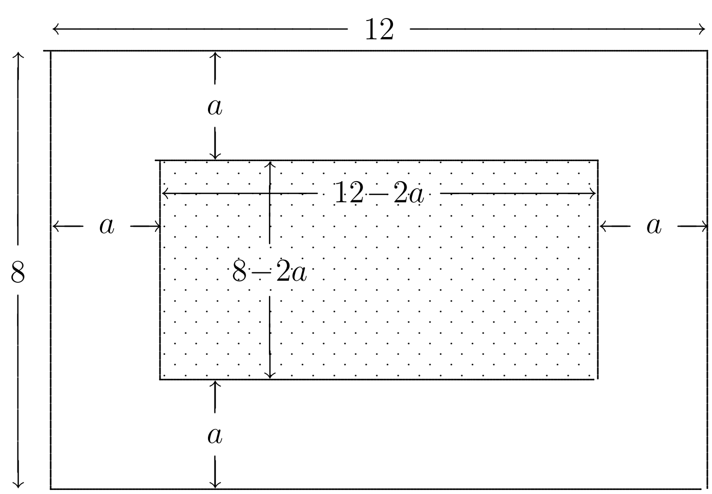
2b $\mathrm{MAD} = \mathrm{Med}\{|x_1 - m|, \ldots, |x_8 - m|\} = \mathrm{Med}\{12, 6, 18, 3, 3, 8, 4, 4\} =$
$= \mathrm{Med}\{3, 3, 4, 4, 6, 8, 12, 18\} = \frac{1}{2}(4+6) = 5$.

$$s_8^2 = \frac{1}{7}\sum_{k=1}^{8}(x_i - 55)^2 = \frac{1}{7}(13^2 + 7^2 + \ldots + 5^2) = 87.1429\ldots \qquad s_8 = \sqrt{87.1429\ldots} \approx 9.335$$

2c

$$F_8(60) = \frac{\text{aantal waarden} \leq 60}{8} = \frac{6}{8} = 0.75$$

3a



Het is evident dat D alleen waarden tussen 0 en 4 aan kan nemen.

3b Uit de figuur volgt: de kans om <u>meer</u> dan a van de rand terecht te komen is de kans om in de gearceerde rechthoek terecht te komen, en die is gelijk aan

$$\mathrm{P}(D > a) = \frac{\text{opp. van gearceerde deel}}{\text{opp. van gehele rechthoek}} = \frac{(8-2a)(12-2a)}{8 \cdot 12} = \frac{a^2 - 10a + 24}{24}$$

Dus $\mathrm{P}(D \leq a) = 1 - (a^2 - 10\,a + 24)/24 = (10\,a - a^2)/24$, voor $0 \leq a \leq 4$.

3c D neemt waarden aan tussen 0 en 4, maar de uitkomsten in de buurt van 0 hebben een groter gewicht dan de uitkomsten dichtbij 5. Bijv., de zone waar $0 \leq D \leq 1$ is veel groter dan de zone waar $3 \leq D \leq 4$ (deze laatste is een rechthoek). Dus zal $0 < \mathrm{E}[D] < 2$, en vanwege het geleidelijke verloop zou ik gokken $1 \leq \mathrm{E}[D] \leq 2$. Het berekenen gaat heel simpel:

$$\mathrm{E}[D] = \int x f_D(x)\,dx = \int x F_D'(x)\,dx = \int_0^4 x(10-2x)/24\,dx = \ldots = \frac{14}{9} \approx 1.556.$$

4a $\mathrm{P}(Z_1 \leq 4) = \mathrm{P}\left(\frac{Z_1 - 1}{3} \leq \frac{4-1}{3}\right) = \mathrm{P}(Z \leq 1)$,
waarbij $Z \sim N(0,1)$.
De laatste kans is gelijk aan $1 - \mathrm{P}(Z \geq 1) \approx 1 - 0.1587 = 0.8413$.

4b $S = X_1 + X_2 + X_3$ heeft een $N(3, 3 \cdot 3^2)$ verdeling.
$\mathrm{P}(S \leq 4) = \mathrm{P}\left(\frac{S-3}{\sqrt{27}} \leq \frac{4-3}{\sqrt{27}}\right) = \mathrm{P}(Z \leq 0.192) \approx 0.576.$

4c Allereerst; $V \sim \mathcal{N}(1-1-1, 3^2+3^2+3^2) = \mathcal{N}(-1, 27)$,
dus $\mathrm{E}[V] = -1$, $\mathrm{Var}(V) = 27$ en $\mathrm{E}[V^2] = \mathrm{Var}(V) + (\mathrm{E}[V])^2 = 28$.
Verder: $\mathrm{E}[S^2] = \mathrm{Var}(S) + (\mathrm{E}[S])^2 = 36$, en
voor $i \neq j$, $\mathrm{E}[Z_i Z_j] = \mathrm{E}[Z_i]\,\mathrm{E}[Z_j] = 1$, terwijl
voor $i = j$, $\mathrm{E}[Z_i Z_j] = \mathrm{E}[Z_i^2] = \mathrm{Var}(Z_i) + (\mathrm{E}[Z_i])^2 = 10$

$$\begin{aligned}
\mathrm{Cov}(S, V) &= \mathrm{E}[SV] - \mathrm{E}[S]\,\mathrm{E}[V] \\
&= \mathrm{E}[(Z_1 + Z_2 + Z_3)(Z_1 - Z_2 - Z_3)] - 3 \cdot (-1) \\
&= \mathrm{E}\left[(Z_1^2 - Z_2^2 - Z_3^2 - 2Z_2Z_3)\right] + 3 \\
&= 10 - 10 - 10 - 2 \cdot 1 + 3 = -9
\end{aligned}$$

5a Op grond van symmetrie geldt $\mathrm{E}[X] = 0$. Dan volgt,

$$\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 = \mathrm{E}\left[X^2\right] = \frac{1}{5}\left((-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2\right) = 2.$$

5b $\mathrm{E}[S_{500}] = \sum \mathrm{E}[X_i] = 0$, en vanwege de onafhankelijkheid

$$\mathrm{Var}(S_{500}) = \sum_{i=1}^{500} \mathrm{Var}(X_i) = 500 \cdot 2 = 1000.$$

5c Chebyshev: $\mathrm{P}(|X-\mu| > a) \le \dfrac{\mathrm{Var}(X)}{a^2}$. Hier: $X = S_{500}$ met verwcahting $\mu = 0$ en variantie 1000.

$$\mathrm{P}(|S_{500}| > 50) = \mathrm{P}(|S_{500} - 0| > 50) \le \frac{1000}{50^2} = 0.25.$$

6a

$$\begin{aligned} L(\alpha) &= f_\alpha(x_1) \cdot f_\alpha(x_2) \cdots f_\alpha(x_5) \\ &= (\alpha x_1^{-\alpha-1}) \cdot (\alpha x_2^{-\alpha-1}) \cdots (\alpha x_5^{-\alpha-1}) \\ &= \alpha^5 (x_1 x_2 x_3 x_4 x_5)^{-\alpha-1} \\ &= \alpha^5 (1.09 \cdot 1.29 \cdot 2.78 \cdot 1.08 \cdot 1.61)^{-\alpha-1} = \alpha^5 (6.7969)^{-\alpha-1} \\ \\ \ell(\alpha) &= \ln(L(\alpha)) = 5\ln\alpha - (\alpha+1)\ln 6.7969 \end{aligned}$$

6b Aanpak: afgeleide van $\ell(\alpha)$ berekenen en gelijk aan nul stellen:

$$\ell'(\alpha) = \frac{5}{\alpha} - \ln 6.7969 = 0 \Leftrightarrow \alpha = \hat{\alpha} = \frac{5}{\ln 6.7969} = 2.609$$

$\ell'(\alpha)$ is, voor $\alpha > 1$, evident een dalende functie dus gaat in 2.609 van positieve naar negatieve waarden, en heeft dus in het punt $\hat{\alpha}$ een maximum.

6c Het gemiddelde: $\bar{x}_5 = 1.57$. Een Pareto(α)-variabele heeft verwachting

$$\int_1^\infty x \, \frac{\alpha}{x^{\alpha+1}} \, dx = \int_1^\infty \alpha x^{-\alpha} \, dx = \left[\alpha \frac{x^{1-\alpha}}{1-\alpha} \right]_1^\infty = \frac{\alpha}{\alpha - 1} \quad \text{als} \quad \alpha > 1$$

Voor $\alpha = \hat{\alpha}$ wordt dit $\dfrac{2.609}{1.609} = 1.621$, zeer dicht in de buurt van het gemiddelde. Anders gezegd: als we α geschat hadden via de formule $\dfrac{\alpha}{\alpha-1} = 1.57$.

(Bij de zogenaamde momenten-schatter hadden we $\hat{\alpha}$ berekend uit $\dfrac{\alpha}{\alpha-1} = 1.57$, wat in dit geval $\hat{\alpha} = \dfrac{1.57}{0.57} = 2.75$ had opgeleverd.)

7a Het gemiddelde van 10 metingen heeft een $\mathcal{N}(\mu, \sigma^2/10)$ verdeling. Een 98 % betrouwbaarheidsinterval heeft de grenzen

$$\bar{x}_{10} \pm z_{0.01} \cdot \frac{0.0025}{\sqrt{10}}, \quad \text{waarbij} \ z_{0.01} = 2.326 \ \text{(afgerond: 2.33)}$$

dus dat wordt $[10.0142, 10.0178]$.

7b Het interval wordt bij n metingen kleiner dan 0.002 gram als

$$2z_{0.01}\frac{0.0025}{\sqrt{n}} \leq 0.002$$

dus als

$$\sqrt{n} \geq \frac{2z_{0.01}0.0025}{0.002} \approx 5.82$$

Oftewel, als $n \geq 5.82^2 = 33.8...$, dus, aangezien n geheel moet zijn: $n \geq 34$.