

Tentamen Kansrekening en statistiek
wi2105IN–deel 2

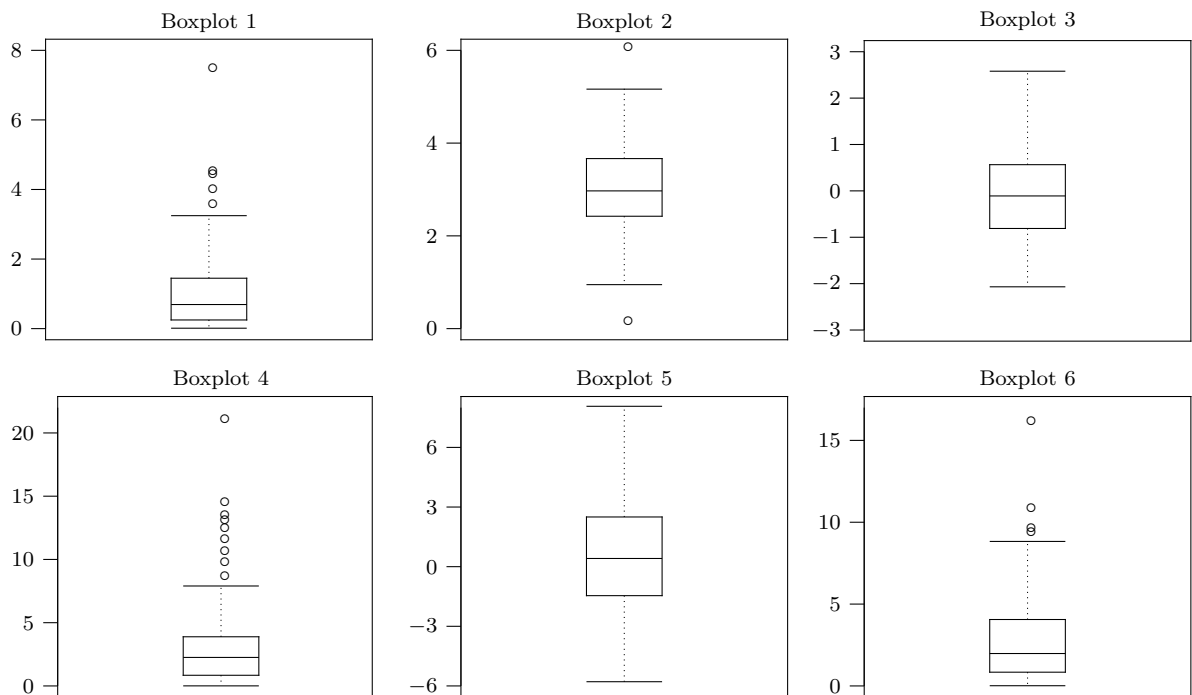
23 juni 2008, 14.00–16.00 uur, LET OP: TWEE uur

Bij dit examen is het gebruik van een (evt. grafische) rekenmachine toegestaan. Tevens krijgt u een formuleblad uitgereikt—na afloop inleveren alstublieft. Normering: De meerkeuze vragen en de open vragen tellen elk voor de helft van het cijfer. Bij de open vragen telt elk (vraag)onderdeel even zwaar.

Meerkeuzevragen

Toelichting: In het algemeen zijn niet altijd vijf van de zes alternatieven 100% fout, het juiste antwoord is het meest volledige antwoord. Maak op het bijgeleverde antwoordformulier het hokje behorende bij het door u gekozen alternatief zwart of blauw. Doorstrepen van een fout antwoord heeft geen zin: u moet het òf uitgummen, òf verwijderen met correctievloeistof òf een nieuw formulier invullen. Vergeet niet uw studienummer in te vullen èn aan te strepen.

1. Onderstaande boxplots zijn allen van datasets van 200 uit een van de volgende verdelingen: de normale verdelingen $N(0, 1)$, $N(0, 9)$, en $N(3, 1)$; en de exponentiële verdelingen $Exp(3)$, $Exp(1)$ en $Exp(1/3)$.



Welke uitspraak over de boxplots is correct?

- a. boxplot 1: $Exp(1/3)$; boxplot 2: $N(3, 1)$; boxplot 4: $Exp(1)$.
 - b. boxplot 1: $Exp(1)$; boxplot 2: $N(3, 1)$; boxplot 4: $Exp(1/3)$.
 - c. boxplot 1: $Exp(1)$; boxplot 2: $N(3, 1)$; boxplot 6: $Exp(3)$.
 - d. boxplot 1: $Exp(1)$; boxplot 4: $Exp(3)$; boxplot 6: $Exp(1/3)$.
 - e. boxplot 2: $N(3, 1)$; boxplot 5: $N(0, 9)$; boxplot 6: $Exp(1)$.
 - f. boxplot 2: $N(3, 1)$; boxplot 3: $N(0, 9)$; boxplot 5: $N(0, 9)$.
2. Een steekproef van omvang 4 uit een normale verdeling met onbekende μ en onbekende σ resulteert in:

17.8 15.1 17.0 16.9

Het tweezijdig 95%-betrouwbaarheidsinterval voor μ wordt gegeven door:

- a.** $15.58 < \mu < 17.82$ **b.** $15.12 < \mu < 18.28$ **c.** $14.89 < \mu < 18.51$
d. $15.76 < \mu < 17.64$ **e.** $15.36 < \mu < 18.04$ **f.** $15.79 < \mu < 17.61$

3. Onderstaande tabel legt het histogram van een dataset vast:

Cel	Hoogte
(0,1]	0.30
(1,3]	0.10
(3,5]	0.20
(5,6]	0.10

Bereken de waarde van de empirische verdelingsfunctie $F_n(t)$ in de punten $t = 3$ en $t = 5$.

- a.** $F_n(3) = 0.40$; $F_n(5) = 0.50$ **b.** $F_n(3) = 0.40$; $F_n(5) = 0.55$
c. $F_n(3) = 0.40$; $F_n(5) = 0.75$ **d.** $F_n(3) = 0.30$; $F_n(5)$ is niet te berekenen
e. $F_n(3) = 0.50$; $F_n(5) = 0.90$ **f.** $F_n(3) = 0.50$; $F_n(5)$ is niet te berekenen

4. Voor een toets doen we 10 metingen, die normaal verdeeld zijn, met onbekende verwachting μ en variantie $\sigma^2 = 10$. De nulhypothese stelt: $\mu = 0$; de alternatieve: $\mu < 0$. Als toetsingsgrootheid gebruiken we het gemiddelde van de dataset en als kritieke gebied het interval $(-\infty, -2]$. Als in werkelijkheid geldt $\mu = -1$, dan is de kans op een fout van de tweede soort ongeveer:

- a.** 0.15 **b.** 0.35 **c.** 0.65 **d.** 0.75 **e.** 0.85 **f.** 0.95

5. Voor een toets doen we 10 metingen, die normaal verdeeld zijn, met onbekende verwachting μ en variantie $\sigma^2 = 10$. De nulhypothese stelt: $\mu = 0$; de alternatieve: $\mu < 0$. Als toetsingsgrootheid gebruiken we het gemiddelde van de dataset, we vinden: $\bar{x} = -1.37$. De p -waarde van dit getal is ongeveer:

- a.** 0.04 **b.** 0.05 **c.** 0.09 **d.** 0.20 **e.** 0.45 **f.** 0.94

6. Voor een bivariate dataverzameling $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, beschouwen we een lineair regressie model zonder intercept: $Y_i = \beta x_i + U_i$ voor $i = 1, 2, \dots, n$, waarbij $U_1, U_2, \dots, U_n$ onafhankelijk zijn met verwachting 0 en variantie σ^2 . De kleinste-kwadraten-schatting voor β wordt gegeven door

a. $\frac{\sum y_i}{\sum x_i}$ **b.** $\frac{1}{n} \sum \frac{y_i}{x_i}$ **c.** $\frac{\sum x_i y_i}{\sum x_i^2}$ **d.** $\frac{\sum x_i^2 y_i^2}{\sum x_i^4}$ **e.** $\frac{\sqrt{\sum y_i^2}}{\sqrt{\sum x_i^2}}$ **f.** $\sqrt{\frac{1}{n} \sum \frac{y_i^2}{x_i^2}}$

7. Een studieadvies-bureau hanteert een IQ-test voor eerstejaars studenten. Deze test heeft een variantie van 225. De scores die hiermee verkregen worden, zijn normaal verdeeld. Een steekproef van 25 studenten uit de groep die zich bij dit bureau voor advies heeft aangemeld, scoort op deze test gemiddeld 119. Wat is het 95% betrouwbaarheidsinterval van het populatiegemiddelde?

- a.** (30.80, 207.00) **b.** (113.12, 124.88) **c.** (101.36, 136.64)
d. (114.08, 123.92) **e.** (53.27, 184.73) **f.** (115.01, 122.99)

8. Gegeven is de dataset

0	1	1	2	3
3	5	6	8	9

De kans dat een bootstrapdataset van 10 getallen uit de empirische verdelingsfunctie van deze dataset precies twee keer het getal 1 bevat is:

- a.** 20% **b.** $\binom{10}{2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^8$ **c.** 80%
d. $\binom{10}{2} \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^8$ **e.** 1 **f.** $2/10!$

9. We beschikken over een dataverzameling $x_1, x_2, \dots, x_n$ die de realisatie is van een steekproef uit een $Exp(\lambda)$ verdeling. De maximum likelihood schatting voor λ is de oplossing van (max betekent: maximaliseer; en min: minimaliseer):

a. $\max \left(1 - e^{-\lambda x_1}\right) \times \dots \times \left(1 - e^{-\lambda x_n}\right).$ b. $\min \left(1 - e^{-\lambda x_1}\right) \times \dots \times \left(1 - e^{-\lambda x_n}\right).$

c. $\max \left(\lambda e^{-\lambda x_1}\right) \times \dots \times \left(\lambda e^{-\lambda x_n}\right).$ d. $\min \left(\lambda e^{-\lambda x_1}\right) \times \dots \times \left(\lambda e^{-\lambda x_n}\right).$

e. $\max \sum_{i=1}^n (1/\lambda - x_i).$

f. $\min \sum_{i=1}^n (1/\lambda - x_i).$

10. Aan de hand van 50 metingen aan een verschijnsel met een normale verdeling met bekende variantie heeft iemand een 90% betrouwbaarheidsinterval gemaakt voor de verwachting: (2.17, 2.42). Men wenst echter een grotere nauwkeurigheid. Om een 99% betrouwbaarheidsinterval te krijgen met lengte hoogstens 0.1, moet het aantal metingen n groter of gelijk zijn aan:

- a. 100 b. 342 c. 747 d. 767 e. 983 f. 2105

Open vragen

Toelichting: 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.

- Bepaal het gemiddelde, de mediaan en de MAD van de dataset: 3 8 1 6 4.
- We beschikken over een dataset $x_1, x_2, \dots, x_{17}$, die de realisatie is van onafhankelijke stochasten $X_1, X_2, \dots, X_{17}$ met een $Exp(\lambda)$ verdeling. Gegeven is dat $\bar{x}_{17} = 4$.
 - Laat zien dat de mediaan van de $Exp(\lambda)$ verdeling gelijk is aan $m_\lambda = \ln 2/\lambda$.
 - We willen weten hoe goed de steekproefmediaan $\text{Med}(X_1, X_2, \dots, X_{17})$ is als schatter voor de verdelingsmediaan m_λ ; we willen uitvinden hoe groot de kans is dan hun verschil groter is dan 0.5. Beschrijf zo nauwkeurig mogelijk de parametrische bootstrapsimulatie voor $\text{Med}(X_1, X_2, \dots, X_{17}) - m_\lambda$ en hoe je hieruit de gevraagde kans schat. Ga er hierbij vanuit dat we λ schatten door $1/\bar{x}_{17}$.
- Een natuurkundige doet onderzoek naar het optreden van een bepaalde fase-overgang. Om technische redenen moet ze de experimenten in paren uitvoeren. Bij elk van de experimenten treedt de fase-overgang op met kans p , onafhankelijk van het andere experiment van het betreffende paar. Ze voert de proef uit voor een n -tal paren, die als onafhankelijk van elkaar te beschouwen zijn. De uitkomsten zijn echter niet per individueel experiment waar te nemen: per paar is het slechts mogelijk vast te stellen hoeveel maal de fase-overgang optrad (dus 0, 1 of 2 keer). Laat de stochast X_i dit aantal weergeven voor het i -de paar.
 - Laat zien dat $T = \frac{1}{2} \bar{X}_n$ een zuivere (Engels: *unbiased*) schatter is voor p .
 - Beschouw $\theta = P(X_i = 1)$. Is $S = 2T(1 - T)$ een zuivere schatter voor θ ? Indien ja: toon dit aan; indien nee: laat zien of de bias positief is of negatief.
- In Ierland is een willekeurig gekozen groep van 100 kiezers gevraagd hoe ze in het referendum over het Europese verdrag gestemd hebben. Noem de fractie voorstemmers in de hele kiezerspopulatie p . Van de 100 ondervraagden hebben er 40 vóór gestemd. Toets met dit gegeven de hypothese $H_0 : p = 0.5$ tegen $H_1 : p < 0.5$ met een significantieniveau van 5% door de p -waarde te bepalen (exact of met de normale benadering).

Antwoorden multiple choice:

1 b. Zie boek: het zijn Boxplot 1 tot en met 6 uit Exercise 17.2, in omgekeerde volgorde gezet.

2 c. $\bar{x}_4 = \frac{x_1 + \dots + x_4}{4} = 16.7$, dus $s_4^2 = \frac{1}{3}((x_1 - 16.7)^2 + \dots + (x_4 - 16.7)^2) = 1.3$ dus $s_4 = 1.14$. $(\bar{X}_4 - \mu)/(S_4/\sqrt{4})$ heeft een $t(3)$ -verdeling. Een tweezijdig 95%-betrouwbaarheidsinterval voor μ wordt gegeven door

$$\left(\bar{x}_4 - \frac{s_4}{\sqrt{4}} t_{3,0.025}, \bar{x}_4 + \frac{s_4}{\sqrt{4}} t_{3,0.025} \right) \approx (14.89, 18.51).$$

3 e. $F_n(t) = \frac{\text{aantal datapunten} \leq t}{n}$. Hoogte van histogram op cel $(a, b]$: $H = \frac{\text{aantal datapunten in } (a, b]}{n \cdot (b - a)}$.

Dus: $\frac{\text{aantal datapunten in } (a, b]}{n} = H \cdot (b - a)$.

$$\begin{aligned} F_n(3) &= \frac{\text{aantal datapunten in } (0, 1] + \text{aantal datapunten in } (1, 3]}{n} \\ &= \frac{\text{aantal datapunten in } (0, 1]}{n} + \frac{\text{aantal datapunten in } (1, 3]}{n} \\ &= 0.30 \cdot (1 - 0) + 0.10 \cdot (3 - 1) = 0.50. \end{aligned}$$

$F_n(5)$ vinden we door de fractie punten in de laatste cel van 1 af te trekken. Dus $F_n(5) = 1 - 0.1 = 0.9$.

4 e. Onder de aanname heeft de toetsingsgrootheid T een $N(-1, 1)$ verdeling ($\sigma^2/10 = 1$). We maken een fout van de tweede soort als we nu de nulhypothese *niet* verwerpen, dus als $T > -2$. $P(T > -2) = P(Z > -1)$ voor een standaardnormale Z , en dus gelijk aan $1 - P(Z > 1) = 1 - 0.1587 = 0.8413$.

5 c. We zoeken $P(T < -1.37)$, voor T met een $N(0, 1)$ verdeling.

6 c. De kleinste kwadraten schatting wordt gevonden door minimaliseren van $S(\beta) = \sum (y_i - \beta x_i)^2$. Het minimum van $S(\beta)$ is oplossing van $S'(\beta) = 0$, hetgeen equivalent is met $\sum (y_i - \beta x_i)x_i = 0$, of te wel $\sum x_i y_i - \beta \sum x_i^2 = 0$. Hieruit volgt $\beta = (\sum x_i y_i) / (\sum x_i^2)$.

7 b. We vinden (σ is bekend! dus kritieke waarde uit de standaardnormale verdeling):

$$\left(\bar{x}_n - z_{0.025} \frac{15}{\sqrt{25}}, \bar{x}_n + z_{0.025} \frac{15}{\sqrt{25}} \right) = (119 - 3 \cdot 1.96, 119 + 3 \cdot 1.96) = (113.12, 124.88)$$

8 b. Dit is natuurlijk de kans op precies 2 successen bij een $\text{Bin}(10, 1/5)$ verdeling.

9 c. Per definitie is de maximum likelihood schatting de oplossing van het maximaliseren van de likelihood. De likelihood wordt gegeven door $L(\lambda) = f(x_1) \cdots f(x_n)$, waarbij $f(x) = \lambda e^{-\lambda x}$ de kansdichtheid is van de $\text{Exp}(\lambda)$ verdeling. Dus

$$L(\lambda) = f(x_1) \cdots f(x_n) = \prod_{i=1}^n (\lambda e^{-\lambda x_i}).$$

10 d Als we het 90% betrouwbaarheidsinterval omzetten naar een 99% betrouwbaarheidsinterval, dan vervangen we $z_{0.05} = 1.645$ door $z_{0.005} = 2.576$. De breedte wordt dan $2.576/1.645$ maal de huidige breedte (0.25), dus: 0.3915; een factor 3.915 te groot. Om de breedte van een betrouwbaarheidsinterval te halveren hebben we *vier* maal zoveel data nodig, dus hier $(3.915)^2 = 15.326$ maal zoveel. Nodig dus minstens $15.326 \times 50 = 766.3$ waarnemingen.

Antwoorden open vragen:

1 $\bar{x}_5 = (3 + 8 + 1 + 6 + 4)/5 = 22/5 = 4.4$. De data op volgorde: 1, 3, 4, 6, 8; dus de mediaan is 4. De absolute verschillen ten opzichte van de mediaan zijn: 3, 1, 0, 2, 4, op volgorde: 0, 1, 2, 3, 4; dus de MAD is 2.

2 Dit is Exercise 6 van Chapter 18.

2b We trekken bootstrapsteekproeven $X_1^*, \dots, X_{17}^*$ (dus met steekproefgrootte 17) uit de $Exp(\hat{\lambda})$ verdeling, met $\hat{\lambda} = 1/4 = 0.25$; we nemen dus *die* exponentiële verdeling die bij onze data past, want $1/\bar{x}_{17}$ is de maximum likelihoodschatting voor λ . De vraag gaat in feite over de stochast $T^* = \text{Med}(X_1^*, X_2^*, \dots, X_{17}^*) - m_{\hat{\lambda}}$, namelijk wat is $P(|T^*| > 0.5)$.

Het simulatierecept: 1. simuleer 17 trekkingen uit een $Exp(0.25)$ verdeling; 2. bepaal t^* door de mediaan te nemen van de 17 trekkingen en hiervan $m_{\hat{\lambda}} = 4 \ln 2$ af te trekken, in formule: $t^* = \text{Med}(x_1^*, x_2^*, \dots, x_{17}^*) - 4 \ln 2$. 3. herhaal dit, zeg, 1000 maal. De gevonden $t_1^*, \dots, t_{1000}^*$ kunnen gebruikt worden om verdelingskenmerken van T^* te bepalen. In ons geval is de fractie t^* 's die in absolute waarde groter is dan 0.5 de schatting voor de gevraagde kans.

3a We moeten laten zien: $E[T] = p$, ongeacht de waarde van p . Er geldt $X_i \sim \text{Bin}(2, p)$, want we tellen het aantal successen in 2 Bernoulli experimenten. Er volgt: $E[\bar{X}_n] = E[X_1] = 2p$, en dus $E[T] = \frac{1}{2}E[\bar{X}_n] = p$.

3b Wegens $X_i \sim \text{Bin}(2, p)$ geldt $P(X_i = 1) = 2p(1 - p)$. Omdat de afbeelding $p \mapsto 2p(1 - p)$ strikt concaaf is op $[0, 1]$ volgt uit de ongelijkheid van Jensen: $E[S] = E[2T(1 - T)] < 2E[T](1 - E[T]) = 2p(1 - p)$. Met andere woorden: S onderschat $P(X_i = 1)$.

Ook zonder de ongelijkheid van Jensen is dit te doen, omdat $\text{Var}(T) > 0$ volgt $E[T^2] > (E[T])^2$, waaruit de negatieve bias volgt: $E[S] = 2(E[T] - E[T^2]) < 2(E[T] - (E[T])^2) = 2(p - p^2) = 2p(1 - p)$.

N.B. Je kunt $E[T^2]$ en dus $E[S]$ zelfs exact uitrekenen. Want $\text{Var}(T) = \frac{1}{4}\text{Var}(\bar{X}_n) = \frac{2p(1-p)}{4n}$ en dus $E[T^2] = (E[T])^2 + \text{Var}(T) = p^2 + \frac{p(1-p)}{2n}$, zodat $E[S] = 2(E[T] - E[T^2]) = (2 - 1/n)p(1 - p)$; de bias is dus $-p(1 - p)/n$. Dit werd niet van je verwacht.

4 Er wordt gevraagd om $P(X \leq 40)$ voor $X \sim \text{Bin}(100, 50)$. Deze kans is ongeveer 0.0284; met de normale benadering wordt via $z = -2$ de waarde 0.0228 gevonden. Beide waarden leiden tot verwerpen van de nulhypothese, omdat we éénzijdig toetsen met $\alpha = 0.05$.