

**Tentamen Kansrekening en statistiek**  
**wi2105IN—deel 2**

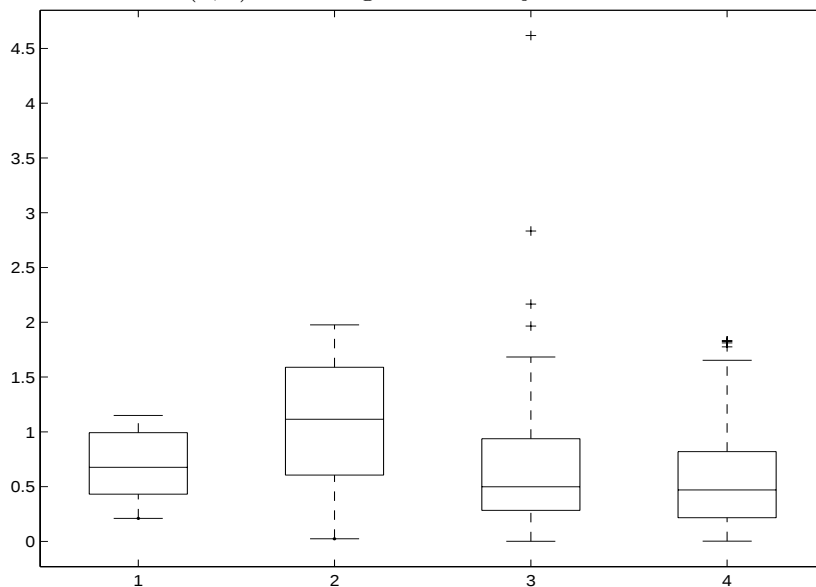
**24 januari 2008, 14.00–16.00 uur, LET OP: TWEE uur**

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 en aan te strepen.

1. Afgebeeld zijn vier boxplots van vier datasets met vijftig elementen. Welke boxplot hoort bij een dataset uit een  $U(0, 2)$ -verdeling en welke bij een dataset uit een  $\text{Exp}(1)$ -verdeling?



- |                                                    |                                                    |
|----------------------------------------------------|----------------------------------------------------|
| a. $U(0, 2)$ : dataset 1, $\text{Exp}(1)$ : set 3; | b. $U(0, 2)$ : dataset 1, $\text{Exp}(1)$ : set 4; |
| c. $U(0, 2)$ : dataset 2, $\text{Exp}(1)$ : set 3; | d. $U(0, 2)$ : dataset 2, $\text{Exp}(1)$ : set 4; |
| e. $U(0, 2)$ : dataset 4, $\text{Exp}(1)$ : set 3; | f. $U(0, 2)$ : dataset 4, $\text{Exp}(1)$ : set 2. |

2. Een steekproef van omvang 5 uit een normale verdeling met onbekende  $\mu$  en onbekende  $\sigma$  resulteert in:

38.0   38.8   41.6   42.2   44.4

Het tweezijdig 90%-betrouwbaarheidsinterval voor  $\mu$  wordt gegeven door:

- |                          |                          |                          |
|--------------------------|--------------------------|--------------------------|
| a. $38.52 < \mu < 43.48$ | b. $39.08 < \mu < 42.92$ | c. $39.22 < \mu < 42.78$ |
| d. $39.50 < \mu < 42.50$ | e. $39.22 < \mu < 44.82$ | f. $39.50 < \mu < 44.82$ |

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

| Cel   | Hoogte |
|-------|--------|
| (0,2] | 0.10   |
| (2,3] | 0.30   |
| (3,4] | 0.30   |
| (4,6] | 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 bivariate dataverzameling  $(x_1, y_1), (x_2, y_2), \dots, (x_{20}, y_{20})$ , beschouwen we een lineair regressie model:  $Y_i = \alpha + \beta x_i + U_i$  voor  $i = 1, 2, \dots, 20$ , waarbij  $U_1, U_2, \dots, U_{20}$  onafhankelijk zijn met verwachting 0 en variantie  $\sigma^2$ . Voor de dataverzameling vinden we

$$\hat{\alpha} = 3, \quad \hat{\beta} = 1.5, \quad s_a = 0.5, \quad s_b = 0.25.$$

We toetsen  $H_0 : \beta = 1$  tegen  $H_1 : \beta > 1$  door middel van een  $t$ -toets bij significantieniveau 0.05. Noteer de waarde van de toetsingsgrootheid door  $t$  en de rechter kritieke waarde door  $c_r$ . Dan geldt

- a.**  $t = 2$  en  $c_r = 1.734$       **b.**  $t = 2$  en  $c_r = 2.101$       **c.**  $t = 4$  en  $c_r = 1.645$   
**d.**  $t = 4$  en  $c_r = 2.101$       **e.**  $t = 6$  en  $c_r = 1.645$       **f.**  $t = 6$  en  $c_r = 1.734$

5. Stel  $X$  heeft een uniforme verdeling op  $[0, a]$ , waarbij  $a$  onbekend is. De nulhypothese is dat  $a = 10$ , en de alternatieve hypothese dat  $a < 10$ . Iemand gaat een toets uitvoeren door twee trekkingen  $X_1$  en  $X_2$  te doen en neemt als toetsingsgrootheid  $T$  het maximum van  $X_1$  en  $X_2$ . Hij besluit te  $H_0$  te verwerpen ten gunste van  $H_1$  als  $T \leq c$ . Hoe groot moet hij  $c$  nemen als voor het significantieniveau de waarde  $\alpha = 0.10$  gekozen wordt? Je mag gebruiken dat onder de nulhypothese geldt:  $P(T \leq c) = (c/10)^2$ .

- a.** 1.414      **b.** 2.173      **c.** 3.162      **d.** 4.225      **e.** 6.666      **f.** 8.100

6. Iemand anders voert de toets uit voorgaande vraag uit met  $c = 2$ . Hoe groot is de kans op een fout van de tweede soort als in werkelijkheid  $a = 8$ ?

- a.** 0.06      **b.** 0.25      **c.** 0.32      **d.** 0.45      **e.** 0.94      **f.** 0.98

7. Een toetsingsgrootheid  $T$  heeft onder de nulhypothese de verdelingsfunctie  $F$ , waarvoor geldt  $F(x) = x^7$  voor  $0 < x < 1$ . Het is gegeven dat waarden van de toetsingsgrootheid ver in de rechterstaart leiden tot verwerpen van de nulhypothese. Uit de data volgt  $t = 0.7$ . De bijbehorende  $p$ -waarde is

- a.**  $(0.7)^7$       **b.**  $(0.7)^{1/7}$       **c.**  $1 - (0.7)^7$       **d.**  $(0.3)^7$       **e.**  $1 - (0.3)^7$       **f.**  $1 - 7(0.7)^6$

8. Gegeven zijn onafhankelijke stochasten  $X_1, X_2, \dots, X_n$ , elk met verwachting  $\mu$  en variantie  $\sigma^2$ . Voor het steekproefgemiddelde  $\bar{X}_n$  en de steekproefstandaardafwijking  $S_n$  geldt:

- a.**  $\bar{X}_n$  is zuiver voor  $\mu$  en  $S_n$  is onzuiver voor  $\sigma$  met negatieve bias.  
**b.**  $\bar{X}_n$  is zuiver voor  $\mu$  en  $S_n$  is zuiver voor  $\sigma$ .  
**c.**  $\bar{X}_n$  is zuiver voor  $\mu$  en  $S_n$  is onzuiver voor  $\sigma$  met positieve bias.  
**d.**  $\bar{X}_n$  is onzuiver voor  $\mu$  en  $S_n$  is onzuiver voor  $\sigma$  met negatieve bias.  
**e.**  $\bar{X}_n$  is onzuiver voor  $\mu$  en  $S_n$  is zuiver voor  $\sigma$ .  
**f.**  $\bar{X}_n$  is onzuiver voor  $\mu$  en  $S_n$  is onzuiver voor  $\sigma$  met positieve bias.

9. 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  $\sigma^2$ . De kleinste-kwadraten-schatting voor  $\beta$  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.  $\sqrt{\frac{1}{n} \sum \frac{y_i^2}{x_i^2}}$

10. Neem de (geordende) dataset: 1 2 2 3 3 4 4 4 4 4.

We nemen een empirische bootstrapsteekproef uit deze dataset. Wat is de kans dat deze bootstrapsteekproef precies zes keer een 4 bevat?

- a. 0.205    b. 0    c. 0.315    d. 0.5    e. 0.324    f. 0.4

## 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.

- Iemand doet twee series (onafhankelijke) metingen van een bepaalde (constante) grootte  $a$ . Serie I bevat acht metingen  $X_1, \dots, X_8$ , volgens methode I, serie II bevat tien metingen  $Y_1, \dots, Y_{10}$  volgens methode II. Ga ervan uit dat  $E[X_i] = E[Y_j] = a$ . Verder is bekend dat  $\text{Var}(X_i) = 0.2$  en  $\text{Var}(Y_j) = 0.5$ .
  - Waar moeten de constanten  $c_1$  en  $c_2$  aan voldoen opdat  $T = c_1 \bar{X}_8 + c_2 \bar{Y}_{10}$  een zuivere schatter is voor  $a$ ?
  - Bereken de variantie van  $T$  voor het geval dat  $c_1 = c_2 = \frac{1}{2}$ .
  - Voor welke keuze van de constanten  $c_1$  en  $c_2$  is  $T$  een *zuivere* schatter voor  $a$  met minimale variantie?
- We beschikken over een dataset  $x_1, x_2, \dots, x_{43}$ , die de realisatie is van onafhankelijke stochasten  $X_1, X_2, \dots, X_{43}$  met een  $\text{Exp}(\lambda)$  verdeling. We schatten  $\lambda$  met schatter  $1/\bar{X}_n$ . Voor de dataset is  $\bar{x}_n = 213.4$ .
  - Beschrijf zo nauwkeurig mogelijk de parametrische bootstrapsimulatie voor de stochast  $T = \lambda \bar{X}_n$ .
  - Hoe bepaal je aan de hand van de bootstrapsimulatie een schatting van de kans dat bovengenoemde schatter een uitkomst geeft die minder dan 10% afwijkt van de werkelijke waarde van  $\lambda$ ?
- Gegeven is een dataset  $v_1, \dots, v_{33}$  van doorslagspanningen van een (electrische) isolator van een nieuw materiaal, in  $kV$ . De doorslagspanning is het voltage waarbij de isolator “doorslaat”, dat wil zeggen, tóch stroom doorlaat; na doorslaan is de isolator onbruikbaar geworden. Men gaat uit van de volgende verdeling van de doorslagspanningen, nl. met verdelingsfunctie  $F(x) = 1 - e^{-\alpha x^2}$  voor  $x > 0$ , waarbij  $\alpha > 0$  een onbekende parameter is. Bepaal de maximum-likelihoodschatting voor  $\alpha$  als gegeven is:  $v_1 + \dots + v_{33} = 124.1$ ,  $v_1^2 + \dots + v_{33}^2 = 581.4$ , en  $\ln v_1 + \dots + \ln v_{33} = 37.89$ .

## Antwoorden multiple choice:

**1 c.** Ten eerste: bij een steekproef uit een  $U(0, 2)$  verdeling zullen alle getallen tussen 0 en 2 liggen, en de kwartielen in de buurt van 0.5 en 1.5. Ten tweede: de exponentiële verdeling is scheef, en voor een  $\text{Exp}(1)$  verdeling is de kans op een waarde groter dan 2 gelijk aan  $e^{-2} = 0.135$ . In een steekproef van 50 zullen gemiddeld  $50 \cdot 0.135 \approx 7$  getallen boven de 2 voorkomen. Bij dataset 3 zijn dit er weliswaar slechts vier, maar bij dataset 4 komen ze helemaal niet voor.

**2 a**  $\bar{x}_5 = \frac{x_1 + \dots + x_5}{5} = 41.0$ , dus  $s_5^2 = \frac{1}{4} ((x_1 - 41.0)^2 + \dots + (x_5 - 41.0)^2) = 6.8$ .  $\frac{\bar{X}_5 - \mu}{S_5/\sqrt{5}}$  heeft een  $t(4)$ -verdeling. Een tweezijdig 90%-betrouwbaarheidsinterval voor  $\mu$  wordt gegeven door

$$\left( \bar{x}_5 - \frac{s_5}{\sqrt{5}} t_{4,0.05}, \bar{x}_5 + \frac{s_5}{\sqrt{5}} t_{4,0.05} \right) \approx (38.52, 43.48).$$

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

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

$$\begin{aligned} F_n(3) &= \frac{\text{aantal datapunten in } (0, 2] + \text{aantal datapunten in } (2, 3]}{n} \\ &= \frac{\text{aantal datapunten in } (0, 2]}{n} + \frac{\text{aantal datapunten in } (2, 3]}{n} \\ &= 0.10 \cdot (2 - 0) + 0.30 \cdot (3 - 2) = 0.50 \end{aligned}$$

$F_n(5)$  is niet te berekenen op grond van de gegevens, doordat 5 midden in een cel ligt.

**4 a.** De waarde van de toetsingsgrootheid is  $t_b = (1.5 - 1)/0.25 = 2$ . Onder de nulhypothese heeft de toetsingsgrootheid een  $t(18)$  verdeling. We verwerpen alleen voor grote waarden van  $T_b$ , dus de rechter kritieke waarde is  $t_{18,0.05} = 1.734$ .

**5 c.** Er moet gelden:  $P(T \leq c | a = 10) = P(X_1 \leq c \text{ en } X_2 \leq c | a = 10) = (c/10)^2 \leq 0.1$ . De grootste  $c$  waarvoor dit het geval is voldoet aan  $c/10 = \sqrt{0.1}$ , en dat is  $c = 10\sqrt{0.1} \approx 3.162$ .

**6 e.** Gevraagd is  $P(\text{niet verwerpen} | a = 8) = 1 - P(T < 2 | a = 8) = 1 - (2/8)^2 = 15/16 \approx 0.94$ .

**7 c.** De  $p$ -waarde is  $P(T > 0.7) = 1 - F(0.7)$ .

**8 a.** We weten  $E[\bar{X}_n] = \mu$  en  $E[S_n^2] = \sigma^2$ . Dus is  $\bar{X}_n$  zuiver voor  $\mu$ , en volgens de ongelijkheid van Jensen is  $\sigma^2 = E[S_n^2] > (E[S_n])^2$ , en dus ook  $E[S_n] < \sigma$ . Dus  $S_n$  is onzuiver voor  $\sigma$  en de bias  $E[S_n] - \sigma$  is negatief.

**9 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  $\beta = (\sum x_i y_i) / (\sum x_i^2)$ .

**10 a.** Deze kans is  $\binom{10}{6} (1/2)^{10} = 210/1024 = 0.205$ .

## Antwoorden open vragen:

**1a**  $E[X_i] = E[Y_j] = a$ , dan ook  $E[\bar{X}_8] = a$  en  $E[\bar{Y}_{10}] = a$ . Dus  $E[c_1\bar{X}_8 + c_2\bar{Y}_{10}] = c_1a + c_2a = (c_1 + c_2)a$ . Conclusie:  $E[c_1\bar{X}_8 + c_2\bar{Y}_{10}] = a$  precies als  $c_1 + c_2 = 1$ .

**1b** Aangezien de metingen onafhankelijk zijn geldt:  $\text{Var}(c_1\bar{X}_8 + c_2\bar{Y}_{10}) = \text{Var}(c_1\bar{X}_8) + \text{Var}(c_2\bar{Y}_{10}) = c_1^2 \text{Var}(\bar{X}_8) + c_2^2 \text{Var}(\bar{Y}_{10}) = c_1^2 \frac{\text{Var}(X_i)}{8} + c_2^2 \frac{\text{Var}(Y_i)}{10}$ . Voor  $c_1 = c_2 = \frac{1}{2}$  levert dit:  $(\frac{1}{2})^2 \frac{0.2}{8} + (\frac{1}{2})^2 \frac{0.5}{10} = \frac{3}{160} = 0.01875$ .

**1c** In het eerste onderdeel is gevonden: om zuiverheid te krijgen moet  $c_1 = 1 - c_2$ . Uit **b** volgt:  $\text{Var}((1 - c_2)\bar{X}_8 + c_2\bar{Y}_{10}) = \frac{2}{8}(1 - c_2)^2 + \frac{5}{10}c_2^2 = \frac{1}{4} - \frac{1}{2}c_2 + (\frac{2}{8} + \frac{5}{10})c_2^2 = \frac{1}{4} - \frac{1}{2}c_2 + \frac{3}{4}c_2^2 = g(c_2)$ . Om het minimum van  $g$  te vinden:  $g'(c) = -\frac{1}{2} + \frac{3}{2}c$ , dus  $g'(c) = 0$  voor  $c = \frac{1}{3}$ . Een tekenschema van  $g'$  leert dat  $g$  in  $c = \frac{1}{3}$  inderdaad een minimum heeft.

Antwoord:  $T = (1 - \frac{1}{3})\bar{X}_8 + \frac{1}{3}\bar{Y}_{10} = \frac{2}{3}\bar{X}_8 + \frac{1}{3}\bar{Y}_{10}$ .

**2a** We trekken bootstrapsteekproeven  $X_1^*, \dots, X_{43}^*$  (dus met steekproefgrootte 43) uit de  $\text{Exp}(\hat{\lambda})$ , met  $\hat{\lambda} = 1/213.4$ ; we nemen dus *die* exponentiele verdeling die bij onze data past, want  $1/\bar{x}_n$  is de maximum likelihoodschatting voor  $\lambda$ . De stochast  $T^* = \hat{\lambda}\bar{X}_{43}^*$  is de bootstrap versie van  $T$ . Het simulatierecept: 1. simuleer 43 trekkingen uit een  $\text{Exp}(1/213.4)$  verdeling; 2. bepaal  $t^*$  door het gemiddelde te nemen van de 43 trekkingen en te delen door 213.4, in formule:  $t^* = \bar{x}_{43}^*/213.4$ ; 3. herhaal dit, bijvoorbeeld 1000 maal; de gevonden  $t_1^*, \dots, t_{1000}^*$  kunnen gebruikt worden om verdelingskenmerken van  $T^*$  te bepalen. Deze zijn op hun beurt schattingen van de corresponderende kenmerken van  $T$ .

**2b** We willen de kans bepalen dat  $1/\bar{X}_{43}$  minder dan 10% afwijkt van  $\lambda$ :

$$P\left(\left|\frac{1}{\bar{X}_{43}} - \lambda\right| < 0.1\lambda\right) = P\left(\left|\frac{1}{\lambda\bar{X}_{43}} - 1\right| < 0.1\right) = P\left(0.9 < \frac{1}{T} < 1.1\right) \approx P\left(0.9 < \frac{1}{T^*} < 1.1\right),$$

waarbij we in de laatste stap het bootstrappincipe toepassen. Omschrijven geeft dat we zoeken:  $P(0.909091 < T^* < 1.111111)$ . Deze kans kunnen we schatten met de fractie van onze bootstrapuitkomsten  $t_1^*, \dots, t_{1000}^*$  die tussen 0.909091 en 1.111111 liggen.

**N.B.** We vragen niet om een betrouwbaarheidsinterval; de 10% slaat op de *afwijking*, niet op een of andere onbetrouwbaarheid.

**3** Door differentiëren (kettingregel) vinden we  $f(x) = 2\alpha x e^{-\alpha x^2}$ , voor  $x > 0$ . We willen de loglikelihood opschrijven, dus nemen we gelijk de logaritme:  $\ln f(x) = \ln 2 + \ln \alpha + \ln x - \alpha x^2$ . We vinden:

$$\ell(\alpha) = 33 \ln 2 + 33 \ln \alpha + \sum_{i=1}^{33} \ln v_i - \alpha \sum_{i=1}^{33} v_i^2.$$

Differentiëren geeft:

$$\ell'(\alpha) = \frac{33}{\alpha} - \sum_{i=1}^{33} v_i^2,$$

en we zien direct dat dit een dalende functie is van  $\alpha$ , een stationair punt is dus een maximum. De oplossing van de vergelijking  $\ell'(\alpha) = 0$  is  $\alpha = 33 / \sum_{i=1}^{33} v_i^2$  dus  $\hat{\alpha} = 33/581.4 = 0.0568$ .