Kansrekening en statistiek
wi2105IN–deel 2
27 januari 2010, 14.00–16.00 uur

Bij dit examen is het gebruik van een (evt. grafische) rekenmachine toegestaan. Tevens krijgt u een formuleblad uitgereikt—na afloop inleveren alstublieft.

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. We construeren een histogram met cellen [0,1], (1,3], (3,5], (5,8], (8,11], (11,14], en (14,18]. Gegeven zijn de waarden van de empirische verdelingsfunctie in de randen van de cellen:

t	0	1	3	5	8	11	14	18
$F_n(t)$	0	0.225	0.445	0.615	0.735	0.805	0.910	1.000

Dan is de hoogte van het histogram op de cel $(8, 11]$ gelijk aan

a. 0.0233 **b.** 0.0035 **c.** 0.0700 **d.** 0.2100 **e.** 0.2450 **f.** 0.2683

2. Gegeven data $x_1, x_2, \ldots, x_{15}$ die we opvatten als een steekproef uit een verdeling met kansdichtheid

$$f(x) = \begin{cases} 0 & x < \theta \\ e^{-(x-\theta)} & x \geq \theta \end{cases}.$$

We toetsen de nulhypothese $\theta = 0$ tegen de alternatieve hypothese $\theta > 0$. Als toetsingsgrootheid nemen we $T = \min\{X_1, X_2, \ldots, X_{15}\}$. Grote waarden van T duiden op de alternatieve hypothese. Stel dat de geobserveerde waarde van T gelijk is aan $t = 0.1$. De p-waarde bij deze toets (afgerond op 2 decimalen) is gelijk aan:

a. 0.14 **b.** 0.90 **c.** 0.22 **d.** 0.06 **e.** 0.12 **f.** 0.43

Hint: Als $X_1, X_2, \ldots, X_n$ een steekproef uit een $Exp(\lambda)$ verdeling is, dan heeft

$$\min\{X_1, X_2, \ldots, X_n\}$$

een $Exp(n\lambda)$ verdeling.

3. Uit een verzameling objecten (bijv. tanks) genummerd 1 tot en met K worden *met* teruglegging 20 objecten getrokken. We willen toetsen $H_0 : K = 100000$ tegen $H_1 : K < 100000$, met het hoogste rangnummer M van onze steekproef als toetsingsgrootheid. We vinden als realisatie voor M de waarde 81115. De p-waarde van deze uitkomst is:

a. 0.013 **b.** 0.041 **c.** 0.015 **d.** 0.520 **e.** 0.154 **f.** 0.852

4. Stel X heeft een uniforme verdeling op $[0, \mu]$ waarbij μ onbekend is. De nulhypothese is dat $\mu = 2.5$, en de alternatieve hypothese dat $\mu > 2.5$. 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. Stel dat hij besluit H_0 te verwerpen ten gunste van H_1 als $T \geq 2$. Als de werkelijke waarde van μ gelijk is aan 3 wat is dan (op 2 decimalen nauwkeurig) de kans op een fout van de tweede soort?

a. 0.12 **b.** 0.44 **c.** 0.28 **d.** 0.33 **e.** 0.53 **f.** 0.21

5. Een onderzoeker bestudeert een data set $x_1, x_2, \ldots, x_n$ met behulp van de empirische verdelingsfunctie F_n. Hij merkt dat hij nog een extra data punt x_{n+1} heeft. Hij vraagt zich af hoe F_n vergelijkt met de empirische verdelingsfunctie F_{n+1} van de data set $x_1, x_2, \ldots, x_n, x_{n+1}$. Van de volgende beweringen is er één juist. Welke?

a. $F_{n+1}(x) = F_n(x) - \frac{1}{n+1}$ voor $-\infty < x < \infty$
b. $F_{n+1}(x) = F_n(x) + \frac{1}{n+1}$ voor $-\infty < x < \infty$
c. $F_{n+1}(x_n) \leq F_n(x_n)$ mits $x_n > x_{n+1}$
d. $F_{n+1}(x_2) \leq F_n(x_2)$ mits $x_2 < x_{n+1}$
e. $F_{n+1}(x) = F_n(x)$ mits $x > x_{n+1}$
f. $F_{n+1}(x) = F_n(x)$ mits $x < x_{n+1}$

6. Stel dat $\bar{X}_n$ en $\bar{Y}_m$ de steekproefgemiddelden zijn van twee onafhankelijke steekproeven van omvang n respektievelijk m. uit dezelfde kansverdeling met verwachting μ en variantie σ^2. We combineren de twee schatters tot een nieuwe schatter

$$T = r\bar{X}_n + (1-r)\bar{Y}_m,$$

waarbij r is een getal is tussen 0 en 1. Dan geldt

a. T is zuiver voor μ met de kleinste MSE voor $r = n/(m+n)$
b. T is zuiver voor μ met de kleinste MSE voor $r = m/(m+n)$
c. T is zuiver voor μ met de kleinste MSE voor $r = 1/2$
d. T is onzuiver voor μ met de kleinste MSE voor $r = n/(m+n)$
e. T is onzuiver voor μ met de kleinste MSE voor $r = m/(m+n)$
f. T is onzuiver voor μ met de kleinste MSE voor $r = 1/2$

7. Gegeven is een dataverzameling die een realisatie is van een steekproef van omvang 16 uit een kansverdeling met verwachting μ. Het gemiddelde van de data is 10 en de steekproefvariantie is 4. Men voert een bootstrapsimulatie uit voor het gestudentiseerde gemiddelde met 1000 herhalingen. Van de 1000 geordende bootstrapwaarden is het volgende bekend:

25-ste	50-ste	100-ste	901-ste	951-ste	976-ste
−3.42	−2.94	−2.01	1.11	1.39	1.62

Het 90% bootstrap betrouwbaarheidsinterval voor μ wordt gegeven door:

a. (8.290, 10.825) **b.** (8.530, 10.695) **c.** (8.995, 10.555)
d. (9.175, 11.710) **e.** (9.305, 11.470) **f.** (9.445, 11.005)

8. Stel dat $X_1, X_2, \ldots, X_n$ een steekproef is uit een verdeling met verwachting μ en variantie σ^2. Beschouw de beweringen

A De steekproefvariantie S_n^2 is een zuivere schatter voor σ^2.

B Het steekproefgemiddelde is een zuivere schatter voor μ.

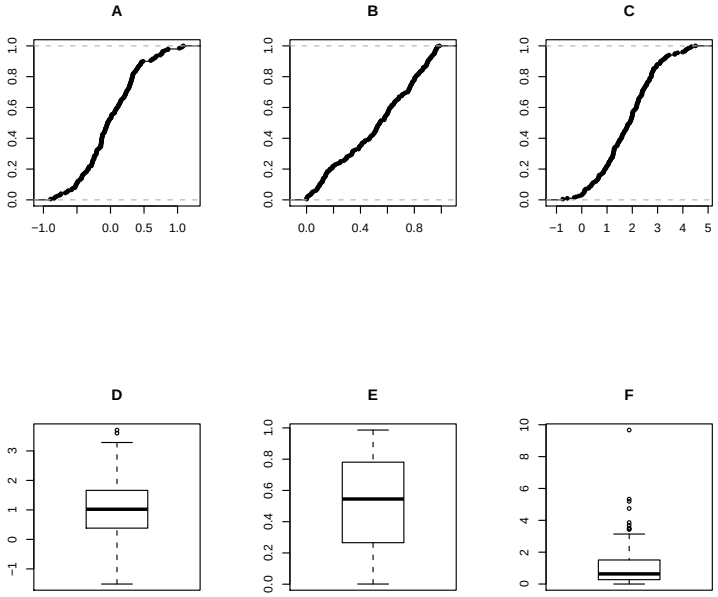
C De steekproefstandaarddeviatie S_n is een zuivere schatter voor σ.

Welke van deze beweringen zijn waar?

a. A en B **b.** geen **c.** B en C
d. A **e.** A, B en C **f.** A en C

9. In onderstaande figuur zijn afbeeldingen A, B en C plots van de empirische verdelingsfunctie van een dataset. Afbeeldingen D, E en F zijn boxplots van een dataset. Alle datasets zijn gebaseerd op 200 waarden, die ofwel uit een normale verdeling, ofwel uit een exponentiële verdeling ofwel uit een uniforme verdeling op $(0, 1)$ getrokken zijn.

Bij welk van deze afbeeldingen is het plausibel dat de data getrokken zijn uit een normale verdeling?



a. A,B,C,D,E
b. A,C,D,E
c. B,F
d. B,E
e. A,C,D
f. A,C,E

10. Om het vermogen tot het schatten van lengte te onderzoeken, is aan 30 willekeurige personen gevraagd de lengte te schatten van een in een rechte lijn liggend stuk touw. De volgende schattingen, in meters en in volgorde van grootte, werden gedaan:

2.2	2.5	2.5	2.7	3.0	3.0
3.0	3.0	3.0	3.2	3.5	3.6
3.6	3.6	3.7	3.7	4.0	4.0
4.0	4.1	4.1	4.2	4.2	4.2
4.3	4.3	4.4	4.5	4.6	4.8

Voor deze dataset geldt $\bar{x}_{30} = 3.65$ en $s_{30} = 0.696$. De werkelijke lengte was 4.0 meter. Neem aan dat de dataset opgevat kan worden als een realisatie van een steekproef $X_1, \ldots, X_{30}$ uit een normale verdeling $N(\mu, \sigma^2)$. Een 90%-betrouwbaarheidsinterval voor μ wordt gegeven door:

a. (3.43, 3.87)
b. $(3.45, 3.85)$
c. $(3.41, 3.89)$
d. $(3.40, 3.90)$

e. $(3.38, 3.92)$

f. Kan niet bepaald worden op grond van deze gegevens.

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.

1. Gegeven de dataset x_1, x_2, x_3. We vatten deze dataset op als een realizatie van de onafhankelijke stochasten X_1, X_2 en X_3. We veronderstellen dat X_1 en X_2 een $Exp(\lambda)$ verdeling hebben, en X_3 een $Exp(3\lambda)$ verdeling.

Leid een formule af voor de meest aannemelijke (maximum likelihood) schatter voor λ, gegeven de dataset x_1, x_2, x_3.

2. We beschikken over een dataset $x_1, x_2, \ldots, x_{17}$, die de realisatie is van onafhankelijke stochasten $X_1, X_2, \ldots, X_{17}$ met een Pareto(α) verdeling. Gegeven is dat $\sum_{i=1}^{17} \log x_i = 8.5$.

a. Laat zien dat de mediaan van de Pareto(α) verdeling gelijk is aan $m_\alpha = 2^{1/\alpha}$.

b. We willen weten hoe goed de steekproefmediaan $\text{Med}(X_1, X_2, \ldots, X_{17})$ is als schatter voor de verdelingsmediaan m_α; we willen uitvinden hoe groot de kans is dan hun verschil groter is dan 0.5. Beschrijf zo nauwkeurig mogelijk de parametrische bootstrapsimulatie voor $\text{Med}(X_1, X_2, \ldots, X_{17}) - m_\alpha$ en hoe je hieruit de gevraagde kans schat. Ga er hierbij vanuit dat we α schatten door $\hat{\alpha} = 17/\sum_{i=1}^{17} \log x_i$ (dit is de meest-aannemelijke schatter voor α). Het is niet nodig uit te leggen hoe je een Pareto-verdeelde stochast kunt simuleren.

Antwoorden multiple choice:

1 a. *Zie huiswerk opgave 15.5.* De hoogte op de cel $(8, 11]$ is gelijk aan

$$\begin{aligned}\frac{\text{aantal } x_i \text{ in } (8,11]}{(11-8)n} &= \frac{\text{aantal } x_i \leq 11 - \text{aantal } x_i \leq 8}{3n} \\ &= \frac{\text{aantal } x_i \leq 11}{3n} - \frac{\text{aantal } x_i \leq 8}{3n} \\ &= \frac{1}{3}\left(\frac{\text{aantal } x_i \leq 11}{n} - \frac{\text{aantal } x_i \leq 8}{n}\right) \\ &= \frac{1}{3}\left(F_n(11) - F_n(8)\right) = \frac{1}{3}\left(0.805 - 0.735\right) = 0.0233.\end{aligned}$$

2 c. Onder de nulhypothese heeft iedere X_i een $Exp(1)$ verdeling. De verdeling van T onder de nulhypothese is dus $Exp(15)$. De p-waarde is dus

$$\mathrm{P}(T > 0.1 \mid H_0 \text{ waar }) = e^{-15 \cdot 0.1} \approx 0.22.$$

3 c. De p-waarde is hier de linkerstaartkans van 81115 onder H_0, dus: $\mathrm{P}(M \leq 81115) = (81115/100000)^{20} = 0.0152$.

4 b. Gezocht is $P(T < 2 \mid \mu = 3) = P(X_1 < 2, X_2 < 2 \mid \mu = 3) = (\frac{2}{3})^2 = \frac{4}{9} \approx 0.444$.

5 d. Als $x_2 < x_{n+1}$, dan is $F_n(x_2) = k/n$ en $F_{n+1}(x_2) = k/(n+1) < k/n$, met k het aantal datapunten $\leq x_2$.

6 a. *Zie huiswerk opgave 20.8.* Allereerst geldt $\mathrm{E}\left[\bar{X}_n\right] = \mu$ en $\mathrm{E}\left[\bar{Y}_m\right] = \mu$. Vanwege de lineariteit vanverwachting geldt dan

$$\mathrm{E}[T] = \mathrm{E}\left[r\bar{X}_n + (1-r)\bar{Y}_m\right] = r\mathrm{E}\left[\bar{X}_n\right] + (1-r)\mathrm{E}\left[\bar{Y}_m\right] = r\mu + (1-r)\mu = \mu.$$

Dus T is zuiver voor μ. Bovendien is $\mathrm{MSE}(T) = \mathrm{Var}(T)$. Verder is $\mathrm{Var}\big(\bar{X}_n\big) = \sigma^2/n$ en $\mathrm{Var}\big(\bar{Y}_m\big) = \sigma^2/m$. Omdat $\bar{X}_n$ en $\bar{Y}_m$ onafhankelijk zijn, is

$$\mathrm{MSE}(T) = \mathrm{Var}(T) = r^2\mathrm{Var}\big(\bar{X}_n\big) + (1-r)^2\mathrm{Var}\big(\bar{Y}_m\big) = r^2 \times \frac{\sigma^2}{n} + (1-r)^2 \times \frac{\sigma^2}{m}$$

Om de minimale r te vinden, moeten we dit differentiëren naar r en de afgeleide gelijk stellen aan nul:

$$\frac{2r\sigma^2}{n} - \frac{2(1-r)\sigma^2}{m} = 0 \quad \Leftrightarrow \quad 2rm - 2n(1-r) = 0 \quad \Leftrightarrow \quad r = \frac{n}{n+m}.$$

7 e. Voor het 90% bootstrap betrouwbaarheidsinterval moeten we gebruik maken van de 50-ste en 951-ste bootstrap waarden. Het interval is dan

$$\left(10 - 1.39 \cdot \frac{2}{4},\, 10 - (-2.94) \cdot \frac{2}{4}\right) = (9.305,\, 11.470).$$

8 a. Beweringen A en B zijn waar, zie MIPS. Jensen's ongelijkheid toont aan dat C niet waar is.

9 e. De emp. verdelingsfunctie heeft een S-curve voor de normale verdeling: dus figuren A en C. Boxplot F duidt op een scheve verdeling en valt dus af. Voor boxplot E is de uniforme verdeling het meest plausibel.

10 a. Een 90%-betrouwbaarheidsinterval voor μ wordt gegeven door

$$\left(\bar{X}_n - \frac{S_n}{\sqrt{n}} t_{n-1,0.05}, \bar{X}_n + \frac{S_n}{\sqrt{n}} t_{n-1,0.05}\right).$$

Nu geldt $t_{n-1,0.05} \approx 1.699$ en dus is het gezochte interval $(3.43, 3.87)$.

Antwoorden open vragen:

1 Een vergelijkbare opgave was onderdeel van bonustoets 6.
De likelihoodfunctie is

$$L(\lambda) = \lambda e^{-\lambda x_1} \lambda e^{-\lambda x_2} 3\lambda e^{-3\lambda x_1}.$$

De loglikelihoodfunctie is dan

$$\ell(\lambda) = \ln 3 + 3 \ln \lambda - \lambda(x_1 + x_2 + 3x_3).$$

De afgeleide nemen geeft

$$\ell'(\lambda) = \frac{3}{\lambda} - (x_1 + x_2 + 3x_3).$$

Een stationair punt van de likelihoodfunctie volgt uit $\ell'(\lambda) = 0$. Dit geeft $\hat{\lambda} = 3/(x_1 + x_2 + 3x_3)$. Aangezien $\ell''(\lambda) = -3/\lambda^2 < 0$ voor alle $\lambda > 0$ is de likelihoodfunctie concaaf. Dit betekent dat $\hat{\lambda}$ de meest aannemelijke schatter is voor λ.

2 Dit is een bewerking van Exercise 6 van Chapter 18.
a. Los op: $F(m) = 1 - m^{-\alpha} = \frac{1}{2}$. Dit geeft $m_\alpha = 2^{1/\alpha}$.
b. We trekken bootstrapsteekproeven $X_1^*, \ldots, X_{17}^*$ (dus met steekproefgrootte 17) uit de Pareto($\hat{\alpha}$) verdeling, met $\hat{\alpha} = 17/8.5 = 2$. De vraag gaat in feite over de stochast $T^* = \text{Med}(X_1^*, X_2^*, \ldots, X_{17}^*) - m_{\hat{\alpha}}$, namelijk wat is $\text{P}(|T^*| > 0.5)$.
Het simulatierecept: 1. simuleer 17 trekkingen uit een Pareto(2) verdeling; 2. bepaal t^* door de mediaan te nemen van de 17 trekkingen en hiervan $m_{\hat{\alpha}} = 2^{1/2} = \sqrt{2}$ af te trekken, in formule: $t^* = \text{Med}(x_1^*, x_2^*, \ldots, x_{17}^*) - \sqrt{2}$. 3. herhaal dit, zeg, 1000 maal. De gevonden $t_1^*, \ldots, t_{1000}^*$ 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.