Kansrekening en statistiek
wi2105IN–deel 2
16 april 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.
Het multiple-choice deel telt voor 2/3 mee voor het eindcijfer; de open-vragen voor 1/3.

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. Beschouw de volgende twee beweringen over de kernschatter:

(A) Verkleinen van de bandbreedte bij een kernschatter zorgt voor een gladdere kernschatting.

(B) De keuze van de kernfunctie heeft een naar verhouding kleinere invloed dan de keuze van de bandbreedte op de kernschatting.

Welke beweringen zijn waar?

a. Alleen A **b.** Alleen B **c.** A en B **d.** Geen van beide

2. We genereren een bootstrap dataset $(x_1^*, x_2^*, x_3^*, x_4^*)$ uit de empirische verdelingsfunctie van de dataset

$$1 \quad 3 \quad 4 \quad 6.$$

De kans dat precies 2 elementen in de bootstrap dataset kleiner zijn dan 2 is (afgerond op 2 decimalen) gelijk aan

a. 0.42 **b.** 0.06 **c.** 0.04 **d.** 0.05 **e.** 0.21 **f.** 0.25

3. In een rapport staat een 95% betrouwbaarheidsinterval $(1.6\,;\,7.8)$ voor de parameter μ van een $N(\mu, \sigma^2)$-verdeling. Het interval is gebaseerd op 16 waarnemingen, geconstrueerd met behulp van het gestudentizeerd gemiddelde. Het gemiddelde van de dataset is gelijk aan:

a. Voor het beantwoorden van de vraag moeten we de waarde van σ kennen.
b. Voor het beantwoorden van de vraag moeten we de waarde van de s_n kennen.
c. Voor het beantwoorden van de vraag moeten we de waarde van μ kennen.
d. 6.2
e. 5.4
f. 4.7

4. Gegeven onafhankelijke stochasten $X_1, \ldots, X_n$ uit een geometrische verdeling met parameter p. We hebben twee schatters voor p, namelijk

$$S = \frac{1}{\bar{X}_n} \quad \text{en} \quad T = \frac{\#\{i \,:\, X_i = 1\}}{n}.$$

Hierbij gebruiken we de notatie $\#A$ voor het aantal elementen in de verzameling A. Dan geldt

a. S en T zijn beide zuiver voor p
b. S is zuiver en T heeft een negatieve bias
c. T is zuiver en S heeft een negatieve bias
d. S is zuiver en T heeft een positiev bias
e. T is zuiver en S heeft een positieve bias
f. S en T zijn beide onzuiver

5. Stel dat $X_1, \ldots, X_n$ onafhankelijke stochasten zijn met een exponentiële verdeling met parameter λ. Beschouw de schatter $T = nM_n$ voor $1/\lambda$, waarbij $M_n = \min\{X_1, \ldots, X_n\}$. We kunnen laten zien dat de verdeling van M_n exponentieel met parameter $n\lambda$ is. De mean squared error van T (als schatter voor $1/\lambda$) is gelijk aan

a. $\frac{1}{\lambda}$
b. $\frac{n}{\lambda}$
c. $\frac{1}{n\lambda}$
d. $\frac{1}{n^2\lambda^2}$
e. $\frac{1}{\lambda^2}$
f. $\frac{1}{n\lambda^2}$

6. In een zevendaagse studie naar het effect van ozon is een groep van 14 ratten in een ozon-vrije omgeving, en een groep van 13 ratten is een ozon-rijke omgeving geplaatst. Voor iedere rat is de gewichtstoename (in grammen) gemeten. We vragen ons af of ozon een negatieve invloed heeft op gewichtstoename. We onderzoeken dit door de hypothese $H_0 : \mu_1 = \mu_2$ te toetsen tegen de alternatieve hypothese $H_1 : \mu_1 > \mu_2$. Hierbij zijn μ_1 en μ_2 de gewichtstoename voor een rat in de ozon-vrije en ozon-rijke omgeving respectievelijk. In de ozon-vrije omgeving is de gemiddelde gewichtstoename $\bar{x}_{14} = 22.40$; in de ozon-rijke omgeving $\bar{y}_{13} = 11.01$. De gepoolde standaard-deviatie is 4.58. De niet gepoolde standaard-deviatie is 4.14.

Veronderstel dat de data normaal verdeeld zijn, met gelijke variantie in beide groepen.
De p-waarde van de t-toets is ongeveer gelijk aan: (kies het antwoord dat het beste past)

a. 0.005
b. 0.05
c. 0.025
d. 0.001
e. 0.01
f. geen van de antwoorden bij (a) tot en met (e)

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 steekproef-variantie 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 95% bootstrap betrouwbaarheidsinterval voor μ wordt gegeven door:

a. (8.290, 10.825) **b.** (8.530, 10.695) **c.** (8.995, 10.555)
d. (9.190, 11.710) **e.** (9.305, 11.470) **f.** (9.445, 11.005)

8. Voor een dataverzameling construeert men een histogram met cellen $[0, 1]$, $(1, 3]$, $(3, 5]$, $(5, 8]$, $(8, 11]$, $(11, 14]$ and $(14, 18]$. Gegeven zijn de waarden van de empirische verdelingsfunctie op de grenzen van de cellen:

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

Het 40ste percentiel van de dataverzameling ligt in cel

a. $(1, 3]$ **b.** $(3, 5]$ **c.** $(5, 8]$
d. $(8, 11]$ **e.** $(11, 14]$ **f.** $(14, 18]$

9. Gegeven is de dataverzameling:

11 14 18 20 21 23 31 32 41

43 67 77 80 83 92 134 183 201 215

en de empirische quantielen

p	0.25	0.50	0.75
$q_n(p)$	21	43	92

Het aantal elementen van de dataverzameling dat buiten de snorharen ("whiskers") van de boxplot valt is gelijk aan

a. 0 **b.** 2 **c.** 3 **d.** 4 **e.** 6 **f.** 8

10. We beschikken over een dataverzameling $x_1, x_2, \ldots, x_n$ met gemiddelde $\bar{x}_n$, mediaan m_n en standaardafwijking s_n. De dataverzameling is een realisatie van een steekproef $X_1, X_2, \ldots, X_n$ uit een $N(\mu, \sigma^2)$ verdeling. Om het gedrag te onderzoeken van de mediaan M_n van $X_1, X_2, \ldots, X_n$ als schatter voor μ wil men de kansverdeling van de stochast $T_n = M_n - \mu$ benaderen door middel van een bootstrapsimulatie. Bij het uitvoeren van de simulatie genereert men in elke iteratie een bootstrap dataverzameling $x_1^*, x_2^*, \ldots, x_n^*$ en berekent men t_n^*. Noteer $\bar{x}_n^*$ en m_n^* als het gemiddelde en de mediaan van $x_1^*, x_2^*, \ldots, x_n^*$. Dan

a. is x_i^* getrokken uit een $N(m_n, s_n^2)$ verdeling, en $t^* = m_n^* - m_n$
b. is x_i^* getrokken uit een $N(m_n, s_n^2)$ verdeling, en $t^* = \bar{x}_n^* - \bar{x}_n$
c. is x_i^* getrokken uit een $N(m_n, s_n^2)$ verdeling, en $t^* = m_n^* - \mu$
d. is x_i^* geloot uit $x_1, x_2, \ldots, x_n$ en $t^* = m_n^* - m_n$
e. is x_i^* geloot uit $x_1, x_2, \ldots, x_n$ en $t^* = \bar{x}_n^* - \bar{x}_n$
f. is x_i^* geloot uit $x_1, x_2, \ldots, x_n$ en $t^* = m_n^* - \mu$

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 $x_1, x_2, \ldots, x_n$. We vatten deze dataset op als een realizatie van de onafhankelijke stochasten $X_1, X_2, \ldots, X_n$ uit een Poisson verdeling met parameter μ. Leid een formule af voor de meest aannemelijke (maximum likelihood) schatter voor μ, gegeven de dataset $x_1, x_2, \ldots, x_n$.

2. We genereren een getal x uit een uniforme verdeling op het interval $[0, \theta]$ ($\theta > 0$). We toetsen $H_0 : \theta = 3$ tegen $H_1 : \theta \neq 3$ door H_0 te verwerpen als $x \leq 0.5$ of $x \geq 2.5$.

a. Bereken de kans op het begaan van een type I fout.

b. Bereken de kans op het begaan van een fype II fout als de echte waarde van θ gelijk is aan 3.5.

c. Bepaal het kritieke gebied van de toets (met als toetsingsgrootheid de waarneming X), als we eisen dat de kans op een type I fout precies 0.1 moet zijn.

3. Stel dat we data $x_1, \ldots, x_n$ hebben die we opvatten als een realisatie van een steekproef $X_1, \ldots, X_n$ uit een normale verdeling met onbekende verwachting μ en bekende variantie σ^2. In dit geval wordt een $100(1-\alpha)\%$ betrouwbaarheidsinterval voor μ gegeven door

$$\left(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \qquad (*)$$

a. Stel dat $\sigma = 2$ en $\alpha = 0.05$. Hoe groot moet n zijn opdat de lengte van het betrouwbaarheidsinterval maximaal 1 is?

b. Stel dat we berekend hebben dat het betrouwbaarheidsinterval voor μ gelijk is aan $\left(1\frac{1}{2}, 3\right)$. Geef een betrouwbaarheidsinterval voor $1 - \mu$ (bij dezelfde data en dezelfde betrouwbaarheid).

Antwoorden multiple choice:

1 b. Zie hoofdstuk 15 MIPS.

2 e. De gevraagde kans is de kans op 2 keer een 1, daarom is de kans

$$\binom{4}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^2 \approx 0.21.$$

3 f. MIPS, opgave 23.10a. Het betrouwbaarheidsinterval is symmetrisch rond het gemiddelde. Het gemiddelde is dus gelijk aan $(1.6 + 7.8)/2 = 4.7$.

4 e. nT heeft een binomiale verdeling met parameter p. De verwachting van nT is dus np, waaruit volgt dat T zuiver is.
S is niet zuiver. Omdat de functie $g(x) = 1/x$ convex is voor $x > 0$ geldt dat

$$\mathrm{E}[S] = \mathrm{E}\left[g(\bar{X}_n)\right] \geq g(\mathrm{E}\left[\bar{X}_n\right]) = g(\mathrm{E}[X_1]) = g(1/p) = p.$$

Schatter S heeft dus een positieve bias.

5 e. Aangezien $\mathrm{E}[T] = n\mathrm{E}[M_n] = n/(n\lambda) = 1/\lambda$ geldt dat T zuiver is. De MSE is dan dus gelijk aan de variantie van de schatter. Nu geldt

$$\mathrm{Var}(T) = \mathrm{Var}(nM_n) = n^2\mathrm{Var}(M_n) = \frac{n^2}{(n\lambda)^2} = \frac{1}{\lambda^2}.$$

6 e. Variatie op MIPS opgave 28.3(a). Aangezien de varianties gelijk verondersteld worden, gebruiken we als toetsingsgrootheid

$$S = \frac{\bar{X}_{23} - \bar{Y}_{22}}{S_p},$$

waarbij S_p de gepoolde standaard-deviatie is. We verwerpen de nulhypothese voor grote waarden van S. Onder H_0 heeft S een t-verdeling met $14+13-2 = 25$ vrijheidsgraden. De geobserveerde waarde van de toetsingsgrootheid bedraagt $(22.4 - 11.01)/4.58 \approx 2.487$. De p-waarde wordt dus gegeven door

$$\mathrm{P}(S \geq 2.487) \approx 0.01.$$

De laatste kans kan gevonden worden door terug te zoeken in de tabel met rechter kritieke waarden van de t verdeing met 25 vrijheidsgraden.

7 d. Voor het 90% bootstrap betrouwbaarheidsinterval moeten we gebruik maken van de 25-ste en 976-ste bootstrap waarden. Het interval is dan

$$\left(10 - 1.62 \cdot \frac{2}{4},\ 10 - (-3.42) \cdot \frac{2}{4}\right) = (9.190,\ 11.710).$$

8 a. Het 40-ste percentiel $q_n(0.40)$ is dat getal zodanig dat (ongeveer) 40% van de data $q_n(0.40)$. Uit de gegevens blijkt dat 22.5% van de data is ≤ 1 en 44.5% van de data is ≤ 3, dus ligt $q_n(0.40)$ ergens tussen 1 en 3.

9 b. De IQR is $92 - 21 = 71$, dus $21 - 1.5 \times IQR = -85.5$ en $92 + 1.5 \times IQR = 198.5$. Zodoende loopt de linker snorhaar van 21 naar 11 en de rechter snorhaar van 92 tot 183. De twee datapunten 201 en 215 liggen voorbij de rechter snorhaar.

10 Omdat we veronderstellen dat $X_1, X_2, \ldots, X_n$ een steekproef vormt uit een $N(\mu, \sigma^2)$ verdeling hebben we te maken met een parametrische bootstrap. In de bootstrapsimulatie moeten we $x_1^*, x_2^*, \ldots, x_n^*$ genereren uit een normale verdeling met geschatte parameters: schat μ door m_n en σ^2 door s_n^2. Aangezien we de bootstrapsimulatie uitvoeren voor $T_n = M_n - \mu$, moeten we in elke iteratie $X_1, X_2, \ldots, X_n$ vervangen door $x_1^*, x_2^*, \ldots, x_n^*$, en dus de mediaan M_n van $X_1, X_2, \ldots, X_n$ vervangen door de mediaan m_n^* van $x_1^*, x_2^*, \ldots, x_n^*$. En verder moeten we parameter van de $N(\mu, \sigma^2)$ verdeling vervangen door de parameter $\mu^* = m_n$ van de geschatte $N(m_n, \sigma^2)$ verdeling.

Antwoorden open vragen:

1 De likelihoodfunctie is

$$L(\mu) = \prod_{i=1}^{n} e^{-\mu} \frac{\mu^{x_i}}{x_i!} = ce^{-n\mu}\mu^{\sum_{i=1}^n x_i}$$

met $c = (\prod_{i=1}^n x_i!)^{-1}$ een constante die niet van μ afhangt. De loglikelihoodfunctie is dan

$$\ell(\mu) = -n\mu + \log\mu \sum_{i=1}^{n} x_i + \log c.$$

De afgeleide nemen geeft

$$\ell'(\mu) = -n + \mu^{-1} \sum_{i=1}^{n} x_i.$$

Een stationair punt van de likelihoodfunctie volgt uit $\ell'(\hat{\mu}) = 0$. Dit geeft $\hat{\mu} = \bar{x}_n$. Aangezien $\ell''(\lambda) < 0$ voor alle $\mu > 0$ is de likelihoodfunctie concaaf. Dit betekent dat $\hat{\mu}$ de meest aannemelijke schatting is voor μ. De meest aannemelijke schatter wordt dus gegeven door $\bar{X}_n$.

2 Dit is een bewerking van Exercise 26.3 uit het boek.
a. De toetsingsgrootheid is de observatie X. Onder H_0 heeft X een uniforme verdeling op $[0, 3]$. Het kritieke gebied wordt gegeven door $K = [0; 0.5] \cup [2.5; \infty)$.

$$\mathrm{P}(\text{type 1 fout}) = \mathrm{P}(X \in K \mid H_0) = \frac{1}{3}0.5 + \frac{1}{3}0.5 = \frac{1}{3}.$$

b. De gevraagde kans is

$$\mathrm{P}(X \notin K \mid \theta = 3.5) = \mathrm{P}(X \in (0.5; 2.5) \mid \theta = 3.5) = \frac{2}{3.5} = \frac{4}{7}.$$

c. Er geldt $K = [0; c_\ell] \cup [c_u; \infty)$, waarbij we c_ℓ en c_u moeten bepalen. Aangezien de type I fout 0.1 moet zijn, geldt

$$\mathrm{P}(X \leq c_\ell \mid H_0) + \mathrm{P}(X \geq c_u \mid H_0) = 0.1.$$

Het ligt voor de hand beide kansen gelijk aan 0.05 te nemen (alhoewel dit niet hoeft), in dit geval moet gelden

$$\frac{c_\ell}{3} = 0.05 \quad \text{en} \quad 1 - \frac{c_u}{3} = 0.05.$$

Hieruit volgt $c_\ell = 0.15$ en $c_u = 2.85$.

3 Zie bonustoets 6 van afgelopen jaar.
a. De lengte is $2z_{\alpha/2}\sigma/\sqrt{n} = 7.84/\sqrt{n}$. Dit moet ≤ 1 zijn. Dus $n \geq (7.84)^2 \approx 61.5$. Neem dus $n \geq 62$.
b. Als $P(L_n \leq \mu \leq U_n) = \gamma$, dan ook $P(1 - U_n \leq 1 - \mu \leq 1 - L_n) = \gamma$. Dit impliceert dat het BI voor $1 - \mu$ gelijk is aan $(-2, -\frac{1}{2})$.