

Tentamen Kansrekening en statistiek
wi2105IN—deel 2

22 juni 2009, 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 meerkeuzevragen 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. Gegeven is de volgende informatie over een histogram bij een dataset met zestig punten.

cel	hoogte
$(0, 2]$	0.10
$(2, 3]$	0.30
$(3, 5]$	0.15
$(5, 9]$	0.05

De empirische verdelingsfunctie F_{60} heeft in de punten 5 respectievelijk 10 de waarden

- a. $F_{60}(5) = 0.55$, $F_{60}(10) = 0$
b. $F_{60}(5) = 0.55$, $F_{60}(10) = 1$
c. $F_{60}(5) = 0.55$, $F_{60}(10)$ is niet te bepalen
d. $F_{60}(5) = 0.80$, $F_{60}(10) = 0$
e. $F_{60}(5) = 0.80$, $F_{60}(10) = 1$
f. $F_{60}(5) = 0.80$, $F_{60}(10)$ is niet te bepalen
2. Aan een tentamen met 20 meerkeuzevragen, elk met vier mogelijkheden, doen 40 studenten mee. Alle deelnemende studenten beantwoorden de vragen onafhankelijk van elkaar en loten bij elke vraag uit de vier mogelijkheden met gelijke kans. Zij X het aantal studenten dat minstens 1 vraag goed beantwoord heeft. Dan heeft X een binomiale verdeling met parameters n en p gegeven door
- a. $n = 20$, $p = (\frac{3}{4})^{40}$ b. $n = 20$, $p = 1 - (\frac{1}{4})^{40}$ c. $n = 20$, $p = 1 - (\frac{3}{4})^{40}$
d. $n = 40$, $p = (\frac{3}{4})^{20}$ e. $n = 40$, $p = 1 - (\frac{1}{4})^{20}$ f. $n = 40$, $p = 1 - (\frac{3}{4})^{20}$
3. Om de slaagkans van een experiment te onderzoeken wordt door elk van honderd personen een serie experimenten uitgevoerd tot ze voor het eerst succes hebben. Een model om dit te beschrijven is: het aantal pogingen benodigde N heeft een $Geo(p)$ -verdeling. We willen p schatten met de maximum likelihood methode. De gegevens:

aantal pogingen	1	2	3	≥ 4
frequentie	40	30	15	15

De likelihoodfunctie wordt dan

- a. $p^{85}(1-p)^{15}$ b. $p^{85}(1-p)^{105}$ c. $p^{85}(1-p)^{60}$
d. $p^{100}(1-p)^{105}$ e. $p^{100}(1-p)^{60}$ f. $p^{100}(1-p)^{15}$
4. Gegeven een dataset $x_1, \dots, x_6$ die geïnterpreteerd wordt als een steekproef uit een $N(\mu, \sigma^2)$ verdeling met onbekende μ en σ . Gegeven zijn $\bar{x}_6 = 10.4$ en $s_6 = 2.2$. Een 95% betrouwbaarheidsinterval voor μ wordt dan gegeven door
- a. (7.87, 12.93) b. (8.09, 12.71) c. (8.20, 12.60) d. (8.64, 12.16)
e. (8.78, 12.02) f. (8.92, 11.88)

5. Men beschikt over een dataverzameling die een realisatie vormt van een steekproef $X_1, X_2, \dots, X_{25}$ uit een $N(\mu, 1)$ verdeling. Men toetst $H_0 : \mu = 0$ tegen $H_1 : \mu > 0$ bij significantieniveau 0.05 met toetsingsgrootheid $\bar{X}_n$. Voor de dataverzameling vindt men $\bar{x}_n = 0.2636$.
- de p -waarde is 0.0934, zodat we H_0 verwerpen
 - de p -waarde is 0.0934, zodat we H_0 niet verwerpen
 - de p -waarde is 0.3974, zodat we H_0 verwerpen
 - de p -waarde is 0.3974, zodat we H_0 niet verwerpen
 - de p -waarde is 1.318, zodat we H_0 verwerpen
 - de p -waarde is 1.318, zodat we H_0 niet verwerpen
6. We beschikken over een dataverzameling $x_1, x_2, \dots, x_n$ met gemiddelde $\bar{x}_n$. De dataverzameling is een realisatie van een steekproef $X_1, X_2, \dots, X_n$ uit een $Exp(\lambda)$ verdeling. De steekproefgrootheid $1/\bar{X}_n$ is een schatter voor λ . Men wil de kansverdeling van de stochast $T_n = 1/\bar{X}_n - \lambda$ benaderen door middel van een bootstrap simulatie. Met wat voor soort bootstrap procedure hebben we hier te maken en wat is de bootstrapversie van T_n ?
- een empirische bootstrap met $T_n^* = 1/\bar{X}_n^* - 1/\bar{x}_n$.
 - een empirische bootstrap met $T_n^* = 1/\bar{X}_n^* - \bar{x}_n$.
 - een empirische bootstrap met $T_n^* = \bar{X}_n^* - 1/\bar{x}_n$.
 - een parametrische bootstrap met $T_n^* = 1/\bar{X}_n^* - 1/\bar{x}_n$.
 - een parametrische bootstrap met $T_n^* = 1/\bar{X}_n^* - \bar{x}_n$.
 - een parametrische bootstrap met $T_n^* = \bar{X}_n^* - 1/\bar{x}_n$.
7. 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:
- 0.15
 - 0.35
 - 0.65
 - 0.75
 - 0.85
 - 0.95
8. Beschouw twee steekproeven $X_1, X_2, \dots, X_n$ en $Y_1, Y_2, \dots, Y_m$ uit kansverdelingen met verwachtingen μ_1 en μ_2 , en gelijke varianties. Voor de hypothese $H_0 : \mu_1 = \mu_2$ beschikt men over twee toetsingsgrootheden

$$T_p = \frac{\bar{X}_n - \bar{Y}_m}{S_p} \quad \text{en} \quad T_d = \frac{\bar{X}_n - \bar{Y}_m}{S_d},$$

waarbij S_p^2 en S_d^2 de ‘pooled’ en ‘non-pooled’ variantie zijn. Men beweert:

- van de twee toetsingsgrootheden is T_p het meest geschikt.
 - T_p en T_d hebben beide een t -verdeling onder de nulhypothese.
- I en II zijn beide juist.
 - I is juist en II is onjuist.
 - I is juist; en II is onjuist, tenzij de steekproeven uit normale verdelingen komen.
 - I is onjuist en II is juist.
 - I is onjuist; en II is onjuist, tenzij de steekproeven uit normale verdelingen komen.
 - I en II zijn beide onjuist.
9. Heteroscedasticiteit (Engels: heteroscedasticity) is een term die bij een regressie wordt gebruikt. Het betekent dat
- de residuen niet tot nul optellen;
 - de residuen een parabolisch patroon laten zien;
 - de x_i -waarden niet regelmatig over de as verdeeld zijn;
 - de foutvariantie niet voor alle x -waarden hetzelfde is;
 - het histogram van de residuen niet op de normale verdeling lijkt;
 - het intercept significant van nul verschilt.

10. Een bioloog onderzoekt het effect van een magnesiumrijk dieet op de sekse van het nageslacht van bepaalde zoogdieren. De nulhypothese is $H_0 : p = 0.5$, de alternatieve $H_1 : p > 0.5$, waarbij p de kans op een vrouwtje is bij het volgen van het dieet. In een steekproef van 83 werden 48 vrouwtjes geboren. De p -waarde die hierbij hoort is ongeveer
- a. 0.02 b. 0.05 c. 0.09 d. 0.12 e. 0.15 f. 0.25

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. De hoeken van een gelijkbenige driehoek zijn θ , θ en γ . (Dus $\theta + \theta + \gamma = \pi$ (rad).) Ze worden elk apart gemeten (in radialen); de metingen van de drie hoeken zijn X_1 , X_2 en X_3 . Gegeven is dat X_1 , X_2 en X_3 onafhankelijk zijn en zuiver zijn voor respectievelijk θ , θ en γ , met variantie σ^2 . We definiëren de volgende schatters voor θ :

$$S = \frac{1}{2}(X_1 + X_2) \quad \text{en} \quad T = \frac{1}{6}(2\pi + X_1 + X_2 - 2X_3)$$

- a. Ga voor elk van de schatters na of hij zuiver is voor θ .
- b. Bereken de variantie voor beide schatters en geef aan welke van de twee schatters u zou prefereren.
2. Hieronder is een dataset met levensduren gegeven. We willen een betrouwbaarheidsinterval voor de verwachte levensduur μ maken. Enkele kentallen van de dataset: $\bar{x} = 40.79$, $s = 22.09$, $n = 43$, eerste en derde kwartiel: respectievelijk 27.77 en 51.13.

4.91	6.15	7.75	16.91	23.51	23.55	24.10	24.15	24.61
24.90	27.77	28.54	28.63	28.99	29.58	30.39	30.63	30.76
32.04	33.84	34.26	34.34	38.25	38.82	40.92	41.19	41.61
41.76	44.70	46.17	49.53	50.43	51.13	52.11	52.45	53.84
60.23	61.52	67.58	79.66	89.92	99.08	102.60		

- a. Maak (een schets van) de boxplot van deze dataset. Leg uit hoe de verschillende posities/afmetingen van de boxplot bepaald worden, en geef de bijbehorende numerieke waarden voor de dataset.
- b. Doe even alsof de data afkomstig is van een normale verdeling (beide parameters onbekend) en bepaal het 95% betrouwbaarheidsinterval voor μ .
- c. Voer met behulp van onderstaande bootstrapresultaten de t -toets uit voor $H_0 : \mu = 48$ versus $H_1 : \mu < 48$, bij significantieniveau 0.04. Er is een gestudentiseerde bootstrap met 1000 herhalingen is uitgevoerd; hieronder een deel van de naar grootte geordende bootstrap-uitkomsten.

21–25	–2.367	–2.363	–2.359	–2.327	–2.324
26–30	–2.281	–2.263	–2.216	–2.200	–2.199
31–35	–2.180	–2.173	–2.170	–2.161	–2.133
36–40	–2.114	–2.071	–2.056	–2.043	–2.026
41–45	–2.021	–2.010	–2.002	–1.995	–1.985
46–50	–1.981	–1.979	–1.965	–1.923	–1.903
51–55	–1.899	–1.896	–1.885	–1.863	–1.847
56–60	–1.842	–1.838	–1.836	–1.780	–1.765
941–945	1.406	1.442	1.495	1.507	1.519
946–950	1.527	1.534	1.537	1.538	1.544
951–955	1.564	1.580	1.584	1.599	1.604
956–960	1.615	1.615	1.618	1.620	1.640
961–965	1.657	1.660	1.660	1.666	1.684
966–970	1.690	1.698	1.711	1.720	1.739
971–975	1.746	1.765	1.784	1.786	1.803
976–980	1.822	1.828	1.839	1.892	1.896

Antwoorden multiple choice:

1 e.

2 f. X is het aantal studenten met 'succes', dus $n = 40$. Succes betekent minstens 1 vraag goed, dus is $p = 1 - P(\text{alle 20 vragen fout}) = 1 - (\frac{3}{4})^{20}$.

3 b. $P(N = k) = (1 - p)^{k-1}p$, en $P(X \geq 4) = (1 - p)^3$ (nl. als er de eerste drie keer geen succes is).

$$L(p) = p^{40} \cdot [(1 - p)p]^{30} \cdot [(1 - p)^2 p]^{15} \cdot [(1 - p)^3]^{15} = \dots = p^{85}(1 - p)^{105}$$

4 b. Geheel standaard; σ onbekend, dus t -verdeling gebruiken. Het aantal vrijheidsgraden m is gelijk aan 5. In tabel (B.2): $t_{5,0.025} = 2.571$. Betr. interval:

$$\left[\bar{x}_6 - \frac{s_6}{\sqrt{6}} t_{5,0.025}, \bar{x}_6 + \frac{s_6}{\sqrt{6}} t_{5,0.025} \right] = [8.0909, 12.7091]$$

5 b. De p -waarde is gelijk aan $P(\bar{X}_{25} \geq 0.2636 \mid \mu = 0)$. Volgens de eigenschappen van de normale verdeling heeft $\bar{X}_{25}$ onder $H_0 : \mu = 0$ een $N(0, 1/25)$ verdeling, zodat $5\bar{X}_{25}$ een $N(0, 1)$ verdeling heeft. Zodoende is $P(\bar{X}_{25} \geq 0.2636 \mid \mu = 0) = P(5\bar{X}_{25} \geq 1.32 \mid \mu = 0) = 0.0934$.

6 d. De kansverdeling van de X_i 's is van een parametrisch type: een $Exp(\lambda)$ verdeling. Zodoende hebben we te maken met een parametrische bootstrap. De parameter λ schatten we door $\hat{\lambda} = 1/\bar{x}_n$, en dus is de bootstrap versie van T_n gelijk aan de stochast $T_n^* = 1/\bar{X}_n^* - \hat{\lambda} = 1/\bar{X}_n^* - 1/\bar{x}_n$.

7 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 $1 - P(Z > 1) = 1 - 0.1587 = 0.8413$.

8 b. We hebben te maken met twee steekproeven uit kansverdelingen met gelijke varianties. In dat geval is T_p de meest geschikte toetsingsgrootheid, dus bewering I is juist. In het geval dat we te maken hebben met steekproeven uit normale verdelingen met gelijke varianties, heeft alleen T_p een t -verdeling. Dus bewering II is onjuist, ook bij steekproeven uit normale verdelingen.

9 d. Zie §22.2.

10 c. Te bepalen $P(X \geq 48)$ voor X met een $Bin(83, 0.5)$ verdeling. Het exacte antwoord is 0.0937, met de normale benadering: 0.0768.

Antwoorden open vragen:

1a Omdat X_1 en X_2 beide zuiver zijn voor θ , is $E[X_1] = \theta$ en $E[X_2] = \theta$. Dit betekent dat

$$E[S] = \frac{1}{2}(E[X_1] + E[X_2]) = \frac{1}{2}(\theta + \theta) = \theta.$$

Dus S is zuiver voor θ . Op een zelfde manier, gebruikmakend van het feit dat $\pi = 2\theta + \gamma$, is

$$\begin{aligned} E[T] &= \frac{1}{6}(2\pi + E[X_1] + E[X_2] - 2E[X_3]) \\ &= \frac{1}{6}(2\pi + \theta + \theta - 2\gamma) \\ &= \frac{1}{6}(2(2\theta + \gamma) + \theta + \theta - 2\gamma) \\ &= \theta. \end{aligned}$$

Dus ook T is zuiver voor θ .

1b Voor S krijgen we (vanwege de onafhankelijkheid van X_1 en X_2)

$$\begin{aligned}\text{MSE}(S) &= \text{Var}(S) + (\text{E}[S] - \theta)^2 \\ &= \text{Var}(S) = \text{Var}\left(\frac{1}{2}(X_1 + X_2)\right) \\ &= \frac{1}{4}[\text{Var}(X_1) + \text{Var}(X_2)] = \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{\sigma^2}{2}.\end{aligned}$$

Voor T krijgen we

$$\begin{aligned}\text{MSE}(T) &= \text{Var}(T) + (\text{E}[T] - \theta)^2 \\ &= \text{Var}(T) = \text{Var}\left(\frac{1}{6}(2\pi + X_1 + X_2 - 2X_3)\right) \\ &= \frac{1}{36}[\text{Var}(X_1) + \text{Var}(X_2) + 4\text{Var}(X_3)] \\ &= \frac{1}{36}(\sigma^2 + \sigma^2 + 4\sigma^2) = \frac{\sigma^2}{6}.\end{aligned}$$

We zien dat de mean squared error van T het kleinst is en derhalve is T te prefereren boven S .

2a De box loopt van 27.77 (eerste kwartiel) tot 51.13 (derde kwartiel). Bij 34.34 loopt de streep door de box die de mediaan markeert. De linker snorhaars eindigt bij 4.91, de kleinste data-waarde. De rechter snorhaars zou tot 86.2 mogen lopen en eindigt dus bij 79.66 (de grootste waarde daaronder), de drie grootste waarden worden dus apart gemarkeerd. Zie verder Chapter 16.

2b Omdat ook de variantie van de normale verdeling onbekend is, moet die geschat worden en maken we een betrouwbaarheidsinterval op basis van de t -verdeling (van het gestudentiseerde gemiddelde). Het aantal vrijheidsgraden is 42, we gebruiken dus $t_{42,0.025} \approx 2.019$ (door interpolatie in de tabel). We vinden dus $40.79 \pm 2.019 \cdot 22.09/\sqrt{43} = 40.79 \pm 2.019 \cdot 3.369 = 40.79 \pm 6.80 = (33.99, 47.59)$.

2c We vinden $t = (40.79 - 48)/(22.09/\sqrt{43}) = -2.1403$, hetgeen tussen de 34ste en 35ste bootstrapwaarde in ligt. De p -waarde is dus ongeveer 0.034. We verwerpen, al zou ik de bootstrapsimulatie nog eens doen met, zeg, 10 000 herhalingen, voor een nauwkeuriger bepaling.