

Kansrekening en statistiek
WI2211TI / WI2105IN—deel 2
2 februari 2012, 14.00–16.00 uur

VOOR WI2211TI: Bij dit examen is het gebruik van een (evt. grafische) rekenmachine toegestaan. Een formuleblad is niet toegestaan.

VOOR WI2105IN: Bij dit examen is het gebruik van een (evt. grafische) rekenmachine toegestaan. Een formuleblad wordt uitgereikt.

Meerkeuzevragen

Toelichting: In het algemeen zijn niet altijd twee van de drie 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 een dataverzameling die een realisatie is van een steekproef $X_1, X_2, \dots, X_n$ uit een geometrische verdeling met parameter p . We toetsen de hypothese $H_0 : p = \frac{1}{2}$ tegen $H_1 : p < \frac{1}{2}$ met toetsingsgrootheid $T = \bar{X}_n$. We verwerpen H_0 ten gunste van H_1
 - a. voor waarden van T in de buurt van 2
 - b. voor hele grote waarden van T
 - c. voor waarden van T dichtbij 0
2. Men beschikt over een dataverzameling die een realisatie vormt van een steekproef $X_1, X_2, \dots, X_5$ uit een $N(\mu, 1)$ verdeling. Iemand toetst $H_0 : \mu = 0$ tegen $H_1 : \mu < 0$ met toetsingsgrootheid $\bar{X}_n$. Stel hij besluit H_0 te verwerpen ten gunste van H_1 als voor de data geldt $\bar{x}_5 \leq -0.5$. Bereken in dat geval de kans op een type I fout.
 - a. 0.26
 - b. 0.31
 - c. 0.13
3. Vervolg op de voorgaande opgave. Bereken de type II fout als in werkelijkheid $\mu = -0.8$.
 - a. 0.07
 - b. 0.90
 - c. 0.25
4. Tien monsters van een nieuw product leveren de volgende waarden (in grammen)

15.02	14.78	15.11	15.16	14.88
15.35	15.35	15.10	15.18	15.14

Het gemiddelde hiervan is 15.11 en de (steekproef)standaardafwijking 0.181. De producent vraagt zich af of er sprake is van een significante afwijking ten opzichte van de gewenste waarde 15.0 gram. Toets hiervoor met de daarvoor meest geschikte toetsingsvariabele T de hypothese $H_0 : \mu = 15.0$ tegen $H_1 : \mu > 15.0$. Er geldt: T heeft een Student-verdeling met

- a. 9 vrijheidsgraden en H_0 wordt verworpen bij significantieniveau 0.05
 - b. 10 vrijheidsgraden en H_0 wordt niet verworpen bij significantieniveau 0.05
 - c. 9 vrijheidsgraden en H_0 wordt niet verworpen bij significantieniveau 0.05
5. Men vult machinaal wijnflessen met een hoeveelheid wijn die, gemeten in milliliters, $N(\mu, \sigma^2)$ verdeeld is met $\sigma = 3$ ml. We zijn voornemens voor een aantal wijnflessen de hoeveelheid wijn te bepalen. Stel dat we μ schatten met het steekproefgemiddelde van deze metingen. Als we met 95% zeker willen zijn dat de echte waarde van μ niet meer dan 1 ml afwijkt van onze schatting, hoeveel wijnflessen moeten we dan meten om dit te bereiken?
 - a. 138
 - b. 35
 - c. 75

6. De Pareto-verdeling is een kansverdeling waarbij de cumulatieve verdelingsfunctie wordt gegeven door

$$F(x) = \begin{cases} 0 & \text{als } x < 1 \\ 1 - \frac{1}{x^\alpha} & \text{als } x \geq 1 \end{cases}.$$

Hier is $\alpha > 0$ een (onbekende) parameter. Voor een dataset $x_1, x_2, \dots, x_n$ veronderstellen we een Pareto verdeling. De likelihoodfunctie voor α wordt gegeven door

a. $L(\alpha) = \left(1 - \frac{1}{x_1^\alpha}\right) \left(1 - \frac{1}{x_2^\alpha}\right) \cdots \left(1 - \frac{1}{x_n^\alpha}\right)$

b. $L(\alpha) = \frac{\alpha^n}{x_1^{\alpha+1} x_2^{\alpha+1} \cdots x_n^{\alpha+1}}$

c. $L(\alpha) = \frac{\alpha^n}{x_1^\alpha x_2^\alpha \cdots x_n^\alpha}$

7. Veertig metingen van het kookpunt van een nieuw ontwikkelde legering leveren het 98 % betrouwbaarheidsinterval (924° C, 932° C). Chemiprac besluit nog een serie van veertig bepalingen te doen. Wat valt er te zeggen over het betrouwbaarheidsinterval voor de gehele serie van tachtig metingen.

U mag aannemen dat de ‘meetfouten’ normaal verdeeld zijn.

- a. Het blijft gecentreerd rond 928° C, en wordt ongeveer gehalveerd.
b. Het midden zal iets verschuiven en het interval wordt ongeveer half zo lang.
c. Het midden zal iets verschuiven en de lengte wordt ongeveer 6.

8. Zij $X_1, X_2, \dots, X_n$ een steekproef van positieve getallen uit een verdeling met verwachting λ en variantie λ , waarbij $\lambda > 0$ een onbekende parameter is. Een schatter voor λ wordt gegeven door $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. De MSE (mean squared error) van deze schatter worden gegeven door:

a. λ^2/n^2

b. λ/n

c. λ/n^2

9. 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

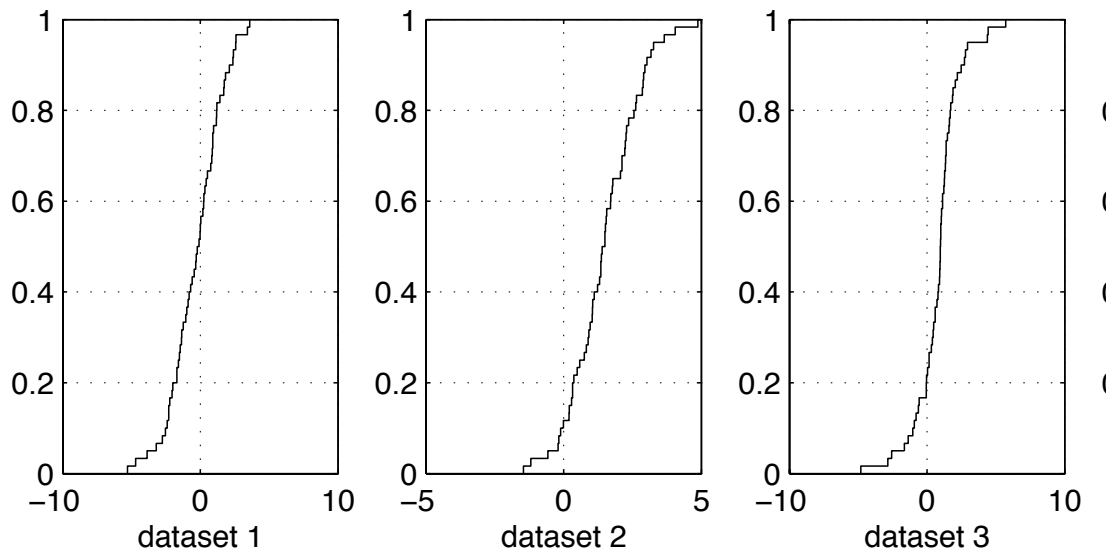
De likelihoodfunctie wordt dan

a. $p^{85}(1-p)^{15}$

b. $p^{85}(1-p)^{105}$

c. $p^{85}(1-p)^{60}$

10. **Voor WI2105IN:** Onderstaand plaatje toont de empirische verdelingsfuncties van drie datasets van 60 elementen. Welk van deze datasets heeft de kleinste mediaan?



a. dataset 1

b. dataset 2

c. dataset 3

10. **Voor WI2211TI:** Beschouw het tweesteeekproevenprobleem, waarbij we aannemen dat de data in beide groepen realisaties zijn van onafhankelijke normaal verdeelde stochastische variabelen. Welke van de onderstaande beweringen zijn waar?

A De gepoolde en ongepoolde variantieschatters geven altijd een verschillende waarde.

B De toetsingsgrootte heeft een t -verdeling onder de nulhypothese, indien we de gepoolde variantieschatter gebruiken.

a. alleen [A]

b. alleen [B]

c. geen van beide

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. Een bepaalde soort heeft drie mogelijke typen bloed: AA, OO of OA, met kansen: θ^2 , $(1-\theta)^2$ en $2\theta(1-\theta)$ respectievelijk. Een steekproef van 150 exemplaren wordt onderzocht en de aantallen met bloedgroep AA, OO en OA blijken te zijn 15, 70 en 65 respectievelijk. Leid op grond hiervan de maximum-likelihoodschatting voor θ af.
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 type II fout als de echte waarde van θ gelijk is aan 3.5.
 - c. Bepaal het kritieke gebied van de toets (met als toetsingsgrootte de waarneming X), als we eisen dat de kans op een type I fout precies 0.1 moet zijn.
3. **Alleen voor WI2105IN** Gegeven zijn metingen $x_1, x_2, \dots, x_{10}$.
 - a. Geef aan hoe je een bootstrap steekproef kunt genereren volgens de empirische bootstrap.
 - b. We nemen nu aan dat we de metingen kunnen opvatten als realisaties van stochastische variabelen $X_1, X_2, \dots, X_n$ met een exponentiële verdeling met onbekende parameter λ . Geef aan hoe je een bootstrap steekproef kunt genereren volgens de parametrische bootstrap.
3. **Alleen voor WI2211TI** Gegeven is een deel van een ANOVA tabel van een studie die is uitgevoerd volgens randomised block design. Het totaal aantal metingen in de studie bedraagt 12. Op vier plaatsen is de waarde weggelaten. Dit zijn de plaatsen waar A, B, C en D staan.

	Df	Sum sq	Mean Sq
block	4	0.00073	A
factor	3	B	0.35431
Residuals	C	D	0.00142

- a. Hoeveel niveaus (levels) heeft de factor?
- b. Geef de waarden van A, B en C in de tabel,

Puntenverdeling bij open vragen:

Opgave:	1	2a	2b	2c	3a	3b
Punten:	3	2	2	2	1	2

Bepaling van het eindcijfer: Bij de berekening van het eindcijfer wordt een apart cijfer berekend voor het meerkeuze- en open-vragen-deel. Bij het meerkeuzedeel zal een correctie voor gokken worden toegepast. Het eindcijfer is een gewogen gemiddelde van de twee verkregen cijfers, waarbij het meerkeuze-deel voor 2/3 en het open-vragen-deel voor 1/3 meetelt.

Antwoorden multiple choice:

1 b. Aangezien $E[\bar{X}_n] = E[X_1] = 1/p$, verwachten we onder de alternatieve hypothese grote waarden van de toetsingsgrootheid.

2 c. Er geldt $\bar{X}_5 \sim N(\mu, 1/5)$. De gevraagde kans is (Z is een standaard normaal verdeelde stochast)

$$\begin{aligned} P_{\mu=0}(\bar{X}_5 \leq -0.5) &= P_{\mu=0}(\sqrt{5}\bar{X}_5 \leq -0.5\sqrt{5}) \\ &= P(Z \leq -0.5\sqrt{5}) = P(Z \geq 0.5\sqrt{5}) \\ &\approx P(Z \geq 1.12) \approx 0.13. \end{aligned}$$

3 c. Gevraagd is $P_{\mu=-0.8}(\bar{X}_n \geq -0.5)$. Met standaardiseren volgt dat deze kans gelijk is aan

$$\begin{aligned} P_{\mu=-0.8}(\sqrt{5}(\bar{X}_n + 0.8) \geq \sqrt{5}(-0.5 + 0.8)) &= P(Z \geq 0.3\sqrt{5}) \\ &\approx P(Z \geq 0.67) \approx 0.25 \end{aligned}$$

4 Als H_0 waar is, dan heeft $T = \sqrt{n} \frac{\bar{X}_n - 15}{S_n}$ een $t(9)$ -verdeling. De t -waarde van de dataset is gelijk aan 1.9218. De rechter kritieke waarde is $t_{9,0.05} = 1.833$. De gevonden t ligt daar boven, dus H_0 wordt verworpen bij significantieniveau 0.05.

5 b. Volgens het 95% betr.baarheidsinterval heeft de schatting met betrouwbaarheid 95% een afwijking hoogstens $\sigma z_{\alpha/2}/\sqrt{n}$. Dit is gelijk aan $5.88/\sqrt{n}$. Dit moet dus kleiner dan 1 zijn. Hieruit volgt $n \geq 34.57$.

6 b. want de kansdichtheid wordt gegeven door $f(x) = \alpha/x^{\alpha+1}$. Gebruik verder de definitie van de likelihoodfunctie.

7 c. De lengte van het t interval wordt gegeven door $2t_{n-1,\alpha/2}s_n/\sqrt{n}$. Nu is $t_{n-1,\alpha/2}$ vrijwel constant als $n \geq 40$; dit zal dus weinig verschil geven. Volgens de wet van de grote getallen zal $S_n^2 \rightarrow \sigma^2$, als n voldoende groot. Zowel bij $n = 40$ als $n = 80$ zal dus $S_n \approx \sigma$. Door de factor $1/\sqrt{n}$ zal verdubbelen van het aantal waarnemingen tot gevolg hebben dat de lengte van het interval gedeeld wordt door $\sqrt{2}$. De lengte bij 40 waarnemingen is 8, dus bij 80 waarnemingen zal dit ruwweg $8/\sqrt{2} \approx 5.6$ worden. Verder kan het gemiddelde natuurlijk iets verschuiven.

8 b. Bekend is dat $E[\bar{X}_n] = E[X_1] = \lambda$, en $\text{Var}(\bar{X}_n) = \frac{\text{Var}(X_1)}{n} = \lambda/n$. De schatter is zuiver zodat de bias gelijk is aan 0. Omdat de schatter zuiver is, is de MSE gelijk aan de variantie van de schatter. Het goede antwoord is dus

$$\text{bias} = 0, \quad \text{MSE} = \lambda/n.$$

9 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).

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

10 a. Voor WI2105IN: De mediaan van de dataset is het punt waarbij de hoogte van de emp. verdelingsfunctie nagenoeg 0.5 is.

10 b. Voor WI2211TI: [A] is niet waar, bij gelijke steekproefgroottes zijn de gepoolde en ongepoolde altijd gelijk. [B] is waar, dit is een $t(m+n-2)$ verdeling, als m en n de grootten van de groepen zijn.

Antwoorden open vragen:

1 De likelihood functie is gelijk aan (K en K' zijn constanten die dus niet van θ afhangen)

$$\begin{aligned} L(\theta) &= K(\theta^2)^{15} ((1-\theta)^2)^{70} (2\theta(1-\theta))^{65} \\ &= K'\theta^{95}(1-\theta)^{205}. \end{aligned}$$

De loglikelihood is dus gelijk aan

$$\ell(\theta) = \ln(K') + 95 \ln \theta + 205 \ln(1-\theta).$$

De afgeleide is gelijk aan

$$\ell'(\theta) = \frac{95}{\theta} - \frac{205}{1-\theta}.$$

Stel nu de afgeleide gelijk aan nul, dit geeft $\hat{\theta} = 95/300 = 19/60$. Verifieer ten slotte dat het stationaire punt daadwerkelijk een maximum is middels een tekenoverzicht ℓ' .

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 $K = [0; 0.5] \cup [2.5; \infty)$.

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

b. De gevraagde kans is

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

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

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

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 Alleen voor WI2105IN

a. Trek n keer met terugleggen uit $\{x_1, \dots, x_n\}$.

b. Schat de parameter λ op grond van de data, bijv. met maximum likelihood. Stel dat $\hat{\lambda}$ de verkregen schatting is. Trek nu n keer uit een exponentiële verdeling met parameter $\hat{\lambda}$.

3 Alleen voor WI2211TI

a. aantal levels = df+1=4

b. De mean sq voor block, $A = 0.0073/4 = 0.001825$. De sum of sq voor factor is gelijk aan $B = 3 \cdot 0.35431 \approx 1.063$. Df voor residuals $C = (12 - 1) - 3 - 4 = 4$.