**Tentamen Statistiek MBW/KW deel 2 29-07-2022**

Afdeling: Propedeuse MBW/KW 2021-2022

Examinator: Dr. J.B.M. Melissen, Dr. R.J. Nijboer

Datum: vrijdag 29 juli 2022 09:00 – 11:00 (adelborsten 08:15 – 10.15), duur tentamen: 2 uur

**1**. **Alle antwoorden moeten gemotiveerd worden**!

2. Rond eindantwoorden (kommagetallen) af op *vier* decimalen, tenzij anders vermeld.

3. Boeken, reader en aantekeningen mogen worden geraadpleegd.

4. De aanwezigheid van *communicatieapparatuur* is niet toegestaan.

5. Het gebruik van een (grafische) rekenmachine met statistische programmatuur en het

raadplegen van de bijbehorende handleiding is toegestaan. Het *statistische* gebruik van deze

rekenmachine is bij een aantal onderdelen ingeperkt. Let op de aanwijzingen!

6. **De opgaven dienen na afloop van het tentamen ingeleverd te worden.**

Dit tentamen bestaat uit vier opgaven (25, 20, 30, 25 punten). Score = Puntentotaal/9

**Opgave 1 (Totaal 25 punten)**

Tijdens een landmachtoefening wordt een week lang het dagelijks brandstofverbruik bijgehouden. Dit leidt tot de volgende waarden in liters: 4215, 4879, 6021, 2983, 2974, 5003, 6972. Neem aan dat de hoeveelheden (kansvariabele die het brandstofverbruik per dag in liters voorstelt) normaal verdeeld zijn, elke dag met dezelfde verwachtingswaarde en standaarddeviatie.

**1a [5pt]** Bereken van de gemeten waarden het steekproefgemiddelde en de steekproefstandaarddeviatie.

**2pt**

**3pt. (Gedeeld door n i.p.v. n-1, 1372,702 i.p.v. s: 1pt)**

**1b [5pt]** Bereken een 95% betrouwbaarheidsinterval voor het verwachte dagelijkse brandstofverbruik van een dergelijke oefening, op grond van bovengenoemde steekproef, zonder daarbij gebruik te maken van de optie TESTS/Interval van de grafische rekenmachine. Rond de grenzen van dit interval af op gehele liters en wel zodanig dat de betrouwbaarheid gewaarborgd blijft.

Omdat de standaarddeviatie niet bekend is (en bovendien de steekproefgrootte kleiner dan 30) moet de -verdeling worden gebruikt. De -waarde bij 95% betrouwbaarheid voor een tweezijdig interval is **2pt**

Het verwachte dagelijkse brandstofgebruik is het gemiddelde van zeven dagwaarden (steekproefgemiddelde) en dat is normaal verdeeld met gemiddelde en standaarddeviatie . We gebruiken hiervoor de schattingen en en de -verdeling.

Het betrouwbaarheidsinterval van is dan

**3pt ( vergeten: 0pt)**

Afronden mag het interval niet kleiner maken, dus afronden naar buiten: .

**Verkeerd of niet afgerond: -1pt**

**1c [5pt]** Toets: tegen . Bepaal de toetsuitslag door het berekenen van een kritiek gebied op basis van de gegeven steekproef van zeven dagen brandstofverbruik. Kies als onbetrouwbaarheid α = 0,1.

Leg in simpele bewoordingen uit wat de uitslag van deze toets betekent voor het dagelijks waterverbruik.

We zoeken de (kleinste) grens van het kritieke gebied , zodanig dat de kans op een fout van de eerste soort (je verwerpt H0, terwijl H0 toch waar is) kleiner is dan α = 0,1:

is hierin het gemiddelde waterverbruik van één dag, dat is normaal verdeeld met gemiddelde (worst case situatie wanneer H0 geldt) en standaarddeviatie . Voor gebruiken de schatting en de -verdeling.

Er geldt dus , dus . **3pt**

(Let op dat de berekening van de -waarde anders gaat dan in 1b, omdat daar het interval tweezijdig was en hier enkelzijdig).

De steekproefwaarde 4721 is groter dan deze waarde, ligt dus in het kritieke gebied en H0 wordt verworpen.

Dit betekent dat met 90% betrouwbaarheid kan worden gesteld dat het dagelijks waterverbruik meer dan 3000 liter is. **2pt**

**1d [5pt]** Hoeveel liter brandstof moet er **dagelijks** minimaal op voorraad zijn, wil met 98% zekerheid aan de dagelijkse behoefte kunnen worden voldaan? (Antwoord in gehele liters).  
Noem de gezochte minimale dagelijkse hoeveelheid , dan is de kans dat het werkelijke waterverbruik deze hoeveelheid voldoende is minstens 98%, dus

, dus, worst case is .

Nu is , het waterverbruik van één dag, normaal verdeeld met gemiddelde en standaarddeviatie . We gebruiken hiervoor de schattingen en en de -verdeling.

We zoeken een linkszijdig interval met kans 0,98. Dat correspondeert met een -waarde van , dus . **4pt**

Afronden naar boven: 8595 liter moet op voorraad zijn. **1pt**

**1e [5pt]** Hoeveel liter brandstof moet er **op weekbasis** minimaal op voorraad zijn wil met 98% zekerheid aan de dagelijkse behoefte kunnen worden voldaan? (Antwoord in gehele liters).

Leg uit waarom deze hoeveelheid niet gelijk is aan zevenmaal de hoeveelheid die in 1d is berekend.

We noemen het waterverbruik van één week, dat is normaal verdeeld met gemiddelde en standaarddeviatie . We gebruiken hiervoor de schattingen en en de -verdeling.

Noem de gezochte minimale wekelijkse hoeveelheid , dan is de kans dat het werkelijke waterverbruik deze hoeveelheid voldoende is minstens 98%, dus

, dus, worst case is .

We zoeken een linkszijdig interval met kans 0,98. Dat correspondeert weer met een -waarde van , dus .

Zevenmaal de waarde van 1d zou veel teveel zijn omdat je nu alleen maar aan het eind van de week goed uit moet komen, i.p.v. elke dag.

**Opgave 2 (Totaal 20 punten).** Bij een groot bedrijf zijn in een jaar (50 weken) 432 sollicitatiegesprekken gevoerd. In de onderstaande tabel is te zien hoeveel gesprekken er per week plaatsvonden.

|  |  |
| --- | --- |
| **Gesprekken**  **per week** | **Frequentie** |
| 0 - 4 | 4 |
| 5 - 9 | 26 |
| 10 - 14 | 18 |
| 15 - 19 | 2 |

**2a [5pt]** Bereken/schat **met behulp van de gegevens uit de tabel** het aantal weken en het aantal gesprekken waarop de tabel is gebaseerd.

De som van de frequenties is het aantal weken: 4 + 26 + 15 + 18 + 2 = 50 weken. **2pt**

Het aantal gesprekken kun je niet exact bepalen, maar is ongeveer gesprekken, we hebben daarbij de gemiddelde waarden van de gegeven ranges gebruikt. **3pt, 432 als antwoord is 0pt**

**2b [10pt]** Toets of het aantal gesprekken per week is te beschouwen als een kansvariabele die een Poissonverdeling volgt, door middel van uitrekenen van een -waarde. Kies als betrouwbaarheid 95% en gebruik in je berekening de verwachte frequenties in één decimaal nauwkeurig.

**2b.** We berekenen eerst de frequenties zoals ze uit de Poissonverdeling zouden volgen. Het gemiddeld aantal gesprekken per week is **1pt**

, **1pt**

**1pt**

, **1pt**

, **1pt**

De kans dat het aantal gesprekken groter dan 19 is, is , die verwaarlozen we.

Omdat de verwachte frequentie te klein is voor toepassing van Chikwadraat (< 5) nemen we de laatste twee categorieën samen tot , waardoor de verwachting daarvan groter dan 5 wordt (zie tabel hieronder).

|  |  |  |
| --- | --- | --- |
| **Gesprekken**  **per week** | **Frequentie**  **Observed** | **Frequentie**  **Expected** |
| 0 - 4 | 4 | 3,4 |
| 5 - 9 | 26 | 28,3 |
| 10 - 19 | 20 | 18,2 |
| Totaal | 50 | 49,9 |

**2pt**

Kijken of Ei en Oi voldoende op elkaar lijken doen we met een aanpassingstoets. De toetsingsgrootheid is

**2pt**

We toetsen hiermee

H0: De waargenomen frequenties kunnen worden verklaard met een Poissonverdeling met .

H1: De waargenomen frequenties kunnen niet zo worden verklaard.

De -waarde (met vrijheidsgraden) is:

**1pt**

Deze kans is niet kleiner dan dus H0 wordt niet verworpen, dus de tabel kan met een betrouwbaarheid van 95% worden verklaard met de Poissonverdeling met .

De verwachte frequentie is eigenlijk ook te klein is dus de eerste twee categorieën moeten eigenlijk ook worden samengenomen tot , waardoor de verwachting daarvan groter dan 5 wordt (zie tabel hieronder).

|  |  |  |
| --- | --- | --- |
| **Gesprekken**  **per week** | **Frequentie**  **Observed** | **Frequentie**  **Expected** |
| 0 - 9 | 30 | 31,7 |
| 10 - 19 | 20 | 18,2 |
| Totaal | 50 | 49,9 |

**2pt**

Kijken of Ei en Oi voldoende op elkaar lijken doen we met een aanpassingstoets. De toetsingsgrootheid is

**2pt**

We toetsen hiermee

H0: De waargenomen frequenties kunnen worden verklaard met een Poissonverdeling met .

H1: De waargenomen frequenties kunnen niet zo worden verklaard.

De -waarde (met vrijheidsgraden) is:

**1pt**

Deze kans is niet kleiner dan dus H0 wordt niet verworpen, dus de tabel kan met een betrouwbaarheid van 95% worden verklaard met de Poissonverdeling met .

**2c [5pt]** Voer de toets ook uit door berekening van het kritieke gebied.

3b. Je kunt ook met een kritiek gebied en grenswaarde werken, dan moet je met de GR oplossen

Dat geeft De waarde ligt niet in het kritieke gebied , dus H0 wordt niet verworpen.

**Opgave 3 (Totaal 30 punten).** Onlangs namen 85 eerstejaars MBW en KW cadetten en adelborsten deel aan een tentamen Statistiek. In de tabel hieronder is per krijgsmachtdeel te zien hoeveel studenten wel en hoeveel niet slaagden voor dit tentamen.

|  |  |  |  |
| --- | --- | --- | --- |
| **Krijgsmacht-**  **Deel ↓** | **Niet**  **geslaagd** | **Geslaagd** | **Totaal** |
| KL | 16 | 13 | 29 |
| KLU | 22 | 5 | 27 |
| KM | 8 | 12 | 20 |
| KMAR | 6 | 3 | 9 |
| Totaal | 52 | 33 | 85 |

**3a [10pt]** Zoek uit of je met behulp van een homogeniteitstoets de volgende hypothese met betrouwbaarheid van 97% kunt verwerpen:

H0: Het wel of niet slagen voor dit tentamen is onafhankelijk van het krijgsmachtdeel van de studenten.

Houd hierbij geen rekening met het feit dat het verwachte aantal geslaagden bij de KMAR eigenlijk te laag is om deze toets toe te passen.

Als het wel of niet slagen onafhankelijk zou zijn van het krijgsmachtdeel, dan zou de tabel er als volgt uit zien:

|  |  |  |  |
| --- | --- | --- | --- |
| **Krijgsmacht-**  **Deel ↓** | **Niet**  **geslaagd** | **Geslaagd** | **Totaal** |
| KL | 17,74 | 11,26 | 29 |
| KLU | 16,52 | 10,48 | 27 |
| KM | 12,24 | 7,76 | 20 |
| KMAR | 5,51 | 3,49 | 9 |
| Totaal | 52 | 33 | 85 |

De toetsingsgrootheid is

Het aantal vrijheidsgraden is (4-1)\*(2-1) = 3, en de overschrijdingskans is

Deze kans is kleiner dan 0,03. Dit betekent dat de nulhypothese (er is geen afhankelijkheid) wordt verworpen met een hoge betrouwbaarheid (97%).

**3b [2pt]** Welke groep studenten levert de grootste bijdrage aan de conclusie in 3a?

De grootste bijdrage aan komt van KLU, zij scoren slechter dan gemiddeld, en daarna van de KM, zij scoren beter dan gemiddeld.

**3c [2pt]** Zoals al werd vermeld blijkt dat het verwachte aantal geslaagde marechaussees te laag is om de homogeniteitstoets uit te voeren. Leg uit waarom je dit probleem niet kunt oplossen door twee kolommen in de tabel samen te voegen.

Als je de twee kolommen samenvoegt krijg je één kolom met de totale aantallen. Er blijft dan niets meer over om verwachte frequenties voor uit te rekenen (het aantal vrijheidsgraden wordt dan ook 0)

**3d [3pt]** Het probleem is wel op te lossen door twee rijen van de tabel samen te voegen. Kun je de KMAR beter samenvoegen met de marine of met de landmacht? Geef een statistische verklaring voor je antwoord.

Het zal niet veel uitmaken. Je kunt de KMAR waarschijnlijk beter samenvoegen met KL, omdat samenvoegen met KM de verschillen kleiner zal maken.

**Elke uitleg die ergens op slaat krijgt 3 punten, ongeacht de keuze.**

**3e [5pt]** Bereken met de gegevens uit de tabel een 90% betrouwbaarheidsinterval voor het slaagpercentage van de KLU cadetten.

# De KLU studenten hebben een slaagfractie van

Het betrouwbaarheidsinterval is uit te rekenen met Clopper-Pearson, door op te lossen

en **2pt**

**2pt**

Dit levert als 90% betrouwbaarheidsinterval [0,0759 ; 0,3506] voor de slaagfractie, ofwel voor het slaagpercentage: [7,6% ; 35,1%]. **1pt**

**3f [5pt]** Bereken met de gegevens uit de tabel een 90% betrouwbaarheidsinterval voor het slaagpercentage van de adelborsten.

# De adelborsten hebben een slaagfractie van .

Het betrouwbaarheidsinterval is uit te rekenen met Clopper-Pearson, door op te lossen

en

Dit levert als 90% betrouwbaarheidsinterval [0,3936 ; 0,7829] voor de slaagfractie, ofwel voor het slaagpercentage: [39,4% ; 78,3%].

**3g [3pt]** Als het goed is zul je zien dat de twee intervallen uit 3e en 3f elkaar niet overlappen. Welke conclusie kun je uit dat gegeven trekken? Met welke betrouwbaarheid?

De betrouwbaarheidsintervallen gaan over verschillende waarden, namelijk de slaagfractie van de KLU cadetten en die van de adelborsten. En dan niet de percentages die voor dit tentamen gelden, maar de waarde van p in de binomiale verdeling die slagen of niet slagen bepaalt voor de twee groepen. Je kunt dus concluderen dat de werkelijke scoorpercentages waarschijnlijk significant verschillend zijn, want als de intervallen zouden overlappen zou daar de gemeenschappelijke waarde van beide groepen in kunnen liggen.

Wat is nu de kans dat de slaagkansen voor de twee groepen eigenlijk gelijk zijn? De gemeenschappelijke waarde ligt dan in het interval van KLU, maar links van dat van KM (kans ), of tussen de twee intervallen (kans ), of in het interval van KM, maar rechts van dat van KLU (kans ). De totale kans is dus

# De betrouwbaarheid (kans) van de uitspraak dat dit niet klopt is dus , dus minstens 90%.

**3 punten als erover is nagedacht (ongeacht het resultaat)**

# Opgave 4 (Totaal 25 punten)

Na het uitvoeren van een oefening buiten Defensieterrein in Nederland komen bij Defensie altijd schadeclaims binnen van particulieren, boeren en gemeentes over (vermeende) beschadigingen aan infrastructuur, gewassen, en over aangetroffen vervuilingen. In de tabel hieronder zijn voor zes oefeningen de geclaimde kosten in kaart gebracht

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** |
| **Duur oefening (dagen)** | 2 | 8 | 4 | 2 | 4 | 10 |
| **Kosten (1000 euros)** | 20 | 58 | 32 | 16 | 42 | 80 |

**4a [10pt]** Bereken met behulp van een tabel de correlatiecoëfficiënt van Pearson. Bepaal of er sprake is van een lineaire correlatie tussen de duur van de oefening en de veroorzaakte kosten. Leg uit hoe daarbij het teken en de grootte van de berekende coëfficiënt een rol spelen.

De duur van de oefening is de verklarende variabele, die kies je als X, de kosten zijn het effect dat is Y.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| **1** | 2 | 20 | 40 | 4 | 400 |
| **2** | 8 | 58 | 464 | 64 | 3364 |
| **3** | 4 | 32 | 128 | 16 | 1024 |
| **4** | 2 | 16 | 32 | 4 | 256 |
| **5** | 4 | 42 | 168 | 16 | 1764 |
| **6** | 10 | 80 | 800 | 100 | 6400 |
| **Gem.** | **5** | **41,33333** | **272** | **34** | **2201,333** |

**5pt**

De correlatiecoëfficiënt van Pearson is een getal tussen -1 en +1 dat aangeeft hoe goed twee variabelen aan een lineair verband voldoen. In dit geval is dat = duur van de oefening in dagen en = de kosten in 1000 euro.

De correlatiecoëfficiënt is

**3pt**

want

De correlatiecoëfficiënt is positief, dus er is een positieve correlatie (d.w.z. bij een langere oefening horen meer kosten), het lineaire verband tussen duur oefening en kosten is een rechte lijn die stijgend is. **1pt**

Hoe dichter bij 1 (of -1), hoe beter de correlatie. In dit geval dus een behoorlijk goede correlatie. Dat betekent dat er een behoorlijk goed lineair verband zal zijn tussen X en Y, dus het is verantwoord om lineaire regressie toe te passen. **1pt**

**4b [8pt]** Bereken de regressielijn door berekening van de coëfficiënten en met behulp van een tabel.

De regressielijn is met

**4pt**

**3pt**

**1pt**

**4c [2pt]** Bereken met de regressielijn een statistisch verantwoorde voorspelling voor de te verwachten kosten voor een oefening die zes dagen gaat duren.

Vul in en je krijgt een bijbehorende voorspelling van de kosten: k€. **2pt**

**4d [5pt]** Bereken een 95% voorspellingsinterval voor de waarde die in 4c is berekend.

Het voorspellingsinterval is

is de -waarde die hoort bij de betrouwbaarheid van 95% met vrijheidsgraden. Bij een betrouwbaarheid van 95% is de linker overschrijdingskans 0,95 + 0,05/2 = 0,975 en

**1pt**

**1pt**

**1pt**

**1pt**

**1pt**