**Tentamen Statistiek MBW/KW deel 2, 2e kans 10-11-2023**

Afdeling: Propedeuse MBW/KW 2022-2023

Examinator: Dr. J.B.M. Melissen, T. Zijlstra MSc.

Datum: vrijdag 10 november 2023, 8:30 – 10:30. **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 (30, 20, 20, 30 punten). Score = Puntentotaal/10.

# Opgave 1 (Totaal 30 punten)

Van 82 cadetten en adelborsten zijn de cijfers van Statistiek deel 1 () en Statistiek deel 1 () uit 2023 uitgezet in een spreidingsdiagram:

De vraag is of er een lineair verband is tussen de cijfers.

De volgende waarden zijn uitgerekend:

**1a [6pt]** Bereken met behulp van de bovenstaande waarden de correlatiecoëfficiënt van Pearson.

Bepaal of er sprake is van een lineaire correlatie tussen de scores van de twee cijfers.

Leg in woorden uit wat de betekenis is van de grootte en het teken van de correlatiecoëfficiënt voor een mogelijke relatie tussen de scores voor de twee vakken.

De correlatiecoëfficiënt van Pearson is een getal tussen -1 en +1 dat aangeeft hoe goed twee variabelen aan een lineair verband voldoen.

De correlatiecoëfficiënt is in dit geval

**4pt**

Hoe dichter bij 1 (of -1), hoe beter de correlatie. In dit geval dus een redelijk sterke 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**

De correlatiecoëfficiënt is positief, dus er is een positieve correlatie (d.w.z. bij een hoger cijfer voor Statistiek 1 wordt een hoger cijfer voor Statistiek 2 voorspeld), het lineaire verband tussen Statistiek 1 en Statistiek 2 is een rechte lijn die stijgend is. **1pt**

**1b [7pt]** Bereken de regressielijn door berekening van de coëfficiënten en met behulp van de tabel (narekenen!). Neem hierbij de score van Analyse als de verklarende variabele .

De regressielijn is met

**3pt**

**3pt**

**1pt**

**1c [2pt]** Bereken met de regressielijn een voorspelling voor de Statistiek 2 score (in vier decimalen) van een cadet die voor Statistiek 1 een 5,5 scoorde.

Vul in en je krijgt een bijbehorende voorspelling van de score: **2pt**

**1d [10pt]** Bereken een **90%** voorspellingsinterval voor de voorspelling uit 1c. Doe de berekeningen in 4 decimalen. Rond de antwoorden af op 2 decimalen

Het voorspellingsinterval is is de -waarde die hoort bij de betrouwbaarheid van 90% met vrijheidsgraden. Bij een betrouwbaarheid van 90% is de linker overschrijdingskans 0,90 + 0,1/2 = 0,95 en

**2pt**

**2pt**

**1pt**

**1pt**

**2pt**

(correct afronden, links naar beneden, recht naar boven) **2pt**

**1e [3pt]** Bereken de Statistiek 1 score waarvoor de regressieformule een Statistiek 2 score voorspelt die 0,2 lager is dan de Statistiek 1 score. Rond deze score af op een decimaal.

(Wanneer je de waarden van en niet hebt kunnen berekenen in 1b neem dan en ).

Uit volgt **3pt**

**1f [2pt]** Stel dat een student als Statistiek 1 score toevallig de gemiddelde Statistiek 1 score van alle cadetten heeft. Laat door berekening zien dat de regressielijn voor deze student een Statistiek 2 score voorspelt die gelijk is aan de gemiddelde Statistiek 2 score van alle studenten.  
Toon vervolgens aan dat dit geen toeval is, namelijk, dat altijd geldt: .  
(Hint: gebruik de formules voor en/of).  
  
 **1pt**  
De formule voor: is omgeschreven: . **1pt**

**1g [1pt]** Verbind de punten in het spreidingsdiagram zodanig dat een ludieke figuur ontstaat.

**Opgave 2 (Totaal 20 punten)**

Een genie-eenheid heeft in het kader van een vredesmissie de opdracht om voormalige oorlogsterreinen te zuiveren van mijnen, boobytraps en niet ontplofte granaten. Tien dagen lang is bijgehouden hoeveel devices onklaar zijn gemaakt: 12, 21, 20, 15, 17, 9, 18, 19, 16, 13.

**2a [6pt]** Bereken van de gemeten waarden het steekproefgemiddelde en de steekproefstandaarddeviatie.

devices/dag **3pt**

devices/dag **3pt**

**(Gedeeld door i.p.v. , 3,6056 i.p.v. s: -2pt**

**Formule met wortel verkeerd: max 1 pt)**

**2b [6pt]** Bereken op grond van de bovengenoemde steekproef een 95% betrouwbaarheids-interval voor het verwachte aantal opgeruimde devices per dag , zonder daarbij gebruik te maken van de optie TESTS/Interval van de grafische rekenmachine. Rond de grenzen van dit interval af op gehele aantallen en wel zodanig dat de betrouwbaarheid gewaarborgd blijft.

De standaarddeviatie is niet bekend, (en steekproefgrootte is kleiner dan 30) dus de **-verdeling** moet worden gebruikt. De -waarde voor 95% tweezijdige betrouwbaarheid is **2pt**

Het aantal opgeruimde devices per dag is normaal verdeeld met een geschat gemiddelde van en een standaarddeviatie van devices per dag.

Het 95% schattingsinterval voor , de verwachtingswaarde van het aantal opgeruimde devices per dag, is dan

**3pt**

Afronden: . **1pt**

**2c [8pt]** Bereken op grond van bovengenoemde steekproef en met behulp van de -verdeling een 95% betrouwbaarheidsinterval (zie formuleblad) **voor de standaarddeviatie** van het aantal opgeruimde devices per dag kan worden gevlogen, zonder daarbij gebruik te maken van de optie TESTS/Interval van de grafische rekenmachine. Rond de grenzen van dit interval af één decimaal.

met betrouwbaarheid . ,

Los op met de GR: en vind . **2pt**

Los op en vind . **2pt**

Schattingsinterval is **2pt**

**Afgerond: [2,6 , 7,0] 2pt**

**Opgave 3 (Totaal 20 punten).**

**3a [7pt]** In 2003 waren er volgens het CBS 191.032 gezinnen met één zoon en 180.974 gezinnen met één dochter. Er wordt vaak gedacht dat de kans op een jongen of een meisje bij geboorte gelijk is. Voer een aanpassingstoets uit om aan de hand van deze steekproef de volgende hypothese te toetsen met een betrouwbaarheidsniveau van 99,9% en leg in woorden uit wat je resultaat betekent:

**H0:** Bij geboorte is de kans op een jongen of een meisje gelijk.

Er zijn in totaal 191.032 + 180.974 = 372.006 kinderen in de genoemde gezinnen. Onder de aanname dat H0 waar is zouden er 372.006 / 2 = 186.003 jongens en 186.003 meisjes moeten zijn. We krijgen dan de volgende tabel:

|  |  |  |
| --- | --- | --- |
| Geslacht | Observed | Expected |
| meisje | 180.974 | 186.003 |
| jongen | 191.032 | 186.003 |

**2pt**

**2pt**

De -waarde (met vrijheidsgraad) is:

**2pt**

Deze kans is kleiner dan dus H0 wordt verworpen, de toets geeft aan dat bij geboorte de kans op een meisje of een jongen niet gelijk is. **1pt**

**3b [6pt]** Voer de toets ook uit door berekening van het kritieke gebied.

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

**3pt**

Dat geeft **2pt**

De waarde ligt in het kritieke gebied , dus H0 wordt verworpen. **1pt**

**3c [7pt]** Bereken op grond van de gegevens van het CBS een schatting voor de kans op een jongen.

Bereken vervolgens met behulp van Clopper-Pearson een 99% betrouwbaarheidsinterval voor deze schatting.

De schatting voor de kans op een jongen is . **2pt**

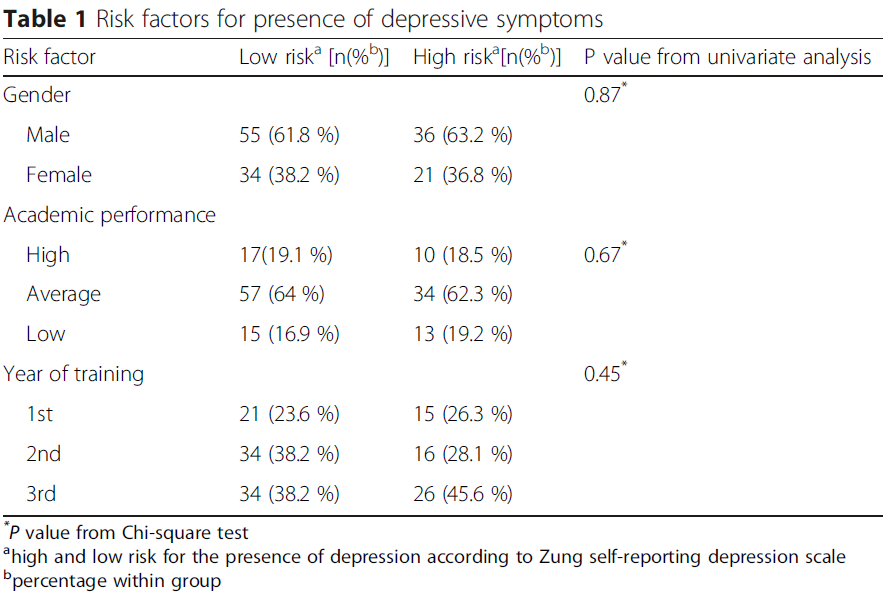
Voor het betrouwbaarheidsinterval lossen we met Clopper-Pearson op:

en **2pt**

**2pt**

Dit levert het interval **1pt**

**Opgave 4 (Totaal 30 punten).** In het wetenschappelijk artikel Nasioudis, D., Palaiodimos, L., Dagiasis, M., Katsarou, A., & Ntouros, E. (2015). Depression in military medicine cadets: a cross-sectional study. *Military Medical Research*, *2*, 1-5. werd onderzoek gedaan onder 55 vrouwelijke en 91 mannelijke cadetten in de bacheloropleiding Geneeskundige Dienst van de Griekse Hellenic Military School of Combat Support Officers gedaan naar depressie. Depressie werd gemeten met behulp van de *Zung score* die kan worden bepaald met een vierpunts scorelijst van 20 items die door de cadetten werd ingevuld. Dit levert een Zung depressiescore op. Een score onder 50 is normaal, tussen 50 en 59 is er sprake van een milde depressie, tussen 60 en 69 een behoorlijke depressie en boven de 70 is de depressie extreem. De cadetten werden onder andere ingedeeld op *Low risk* (depressiescore <45) en *High risk* (score ≥45) en vervolgens ingedeeld op geslacht, academische resultaten (*Average* betekent dat de gemiddelde score over alle vakken binnen één standaarddeviatie van het gemiddelde ligt) en studiejaar.



**4a [10pt]** In de tabel wordt steeds een gegeven op grond van een Chi-Square waarde en een gegeven. Reken de gegeven waarde voor de *Year of training* tabel na in minstens drie decimalen nauwkeurig, d.w.z. maak de tabel met *Expected* waarden, bereken de waarde van en vervolgens de overschrijdingskans .

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Observed | | | |  | Expected | | | |
| Year of training | Low risk | High risk | Totaal |  | Year of training | Low risk | High risk | Totaal |
| 1 | 21 | 15 | 36 |  | 1 | **21,9452** | **14,0548** | 36 |
| 2 | 34 | 16 | 50 |  | 2 | **30,4795** | **19,5205** | 50 |
| 3 | 34 | 26 | 60 |  | 3 | **36,5753** | **23,4247** | 60 |
| Totaal | 89 | 57 | 146 |  | Totaal | 89 | 57 | 146 |

**4pt**

**3pt**

**3pt**

**4b [5pt]** Welke conclusies kun je trekken uit de p-waarden van Table 1 uit het artikel?

Geen van de waarden zijn klein (niet kleiner dan bv. 0,05), **2pt**

dat betekent (er wordt een homogeniteitstest gedaan) dat de twee variabelen niet afhankelijk zijn, d.w.z., depressierisico is gelijk verdeeld over de geslachten, de studieprestaties en de jaren, zijn daar niet van afhankelijk **3pt**

**4c [5pt]** In het artikel wordt een ander artikel aangehaald waarin bij niet-militaire medische studenten een high risk percentage werd geconstateerd van 30,4%. Kun je nu met een 99% betrouwbaarheid stellen dat militaire medische studenten een hoger risico op depressie hebben?

Per medische student is er een kans van 0,304 op high risk depressie. Dit voldoet aan een binomiale verdeling. Voor de Griekse studenten geldt nu dat de kans dat er onder de 146 studenten er 57 of meer een high risk hebben gelijk is aan

Deze kans is niet kleiner dan 0,01, dus de bewering is niet waar met een betrouwbaarheid van 99%.

**4d [10pt]** Bereken de kans dat een cadet een *Average* studiescore heeft. Hoeveel van de cadetten zouden er eigenlijk *Average* hebben moeten scoren?

Bereken de kans dat van de 146 cadetten er hoogstens 91 een gemiddelde studiescore hebben. Wat kun je hieruit concluderen als je een betrouwbaarheid van 95% aanhoudt?

Een gemiddelde studiescore geldt voor studenten die minder dan één standaarddeviatie van het gemiddelde af scoren. De kans daarop is , **3pt**

dus van de 146 cadetten zouden er gemiddeld *average* moeten scoren. **3pt**

De kans dat dat er 91 of minder zijn is . **3pt**

Deze kans is niet kleiner dan 0,05, dus niet genoeg afwijking. Nu is de conclusie dat de studieresultaten niet afwijken van een normale verdeling. **1pt**