**Tentamen Statistiek KW/MBW (deel 2, tweede kans)**

Afdeling: Propedeuse KW/MBW 2019-2020

Examinator: Dr. J.B.M. Melissen

Datum: 13 november 2020, 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. **Lever de antwoorden in op het geprinte antwoordformulier (zet je naam erop), de berekeningen en uitleg op gelinieerd papier.**

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

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

**Opgave 1 (Totaal 20 punten)**

Tijdens militaire oefeningen wordt gebruik gemaakt van standaard NATO 24-uursrantsoenen. In een depot van de logistieke dienst van waaruit deze dozen geleverd worden is van een aantal voorgaande maanden het aantal uitgeleverde dozen bijgehouden: 738, 621, 810, 389, 470, 622, 515. Neem aan dat de uitgeleverde aantallen normaal verdeeld zijn.

**1a. [10pt]** Bereken een 90% betrouwbaarheidsinterval voor op grond van bovengenoemde steekproef, zonder daarbij gebruik te maken van de optie TESTS van de grafische rekenmachine. Rond de grenzen van dit interval af op gehele getallen en wel zodanig dat de 90% betrouwbaarheid gewaarborgd blijft.

De standaarddeviatie is niet bekend en moet worden geschat met behulp van de steekproefstandaarddeviatie die eerst moet worden berekend:

Omdat wordt gewerkt met de steekproefstandaarddeviatie en de steekproefgrootte kleiner is dan 30 moet de -verdeling worden gebruikt met vrijheidsgraden.

Voor het betrouwbaarheidsinterval voor moeten de volgende vergelijkingen worden opgelost:

dit levert als interval . Afronden levert

**1b. [5pt]** Toets: tegen . Bepaal de toetsuitslag via het berekenen van een kritiek gebied. Kies hierbij α = 0,05 en neem bij dit onderdeel aan dat de aantallen normaal verdeeld zijn met standaarddeviatie .

De toetsingsgrootheid is normaal verdeeld met gemiddelde μ = 710 (volgens de worst case situatie van de aanname ) en standaarddeviatie . **2pt**

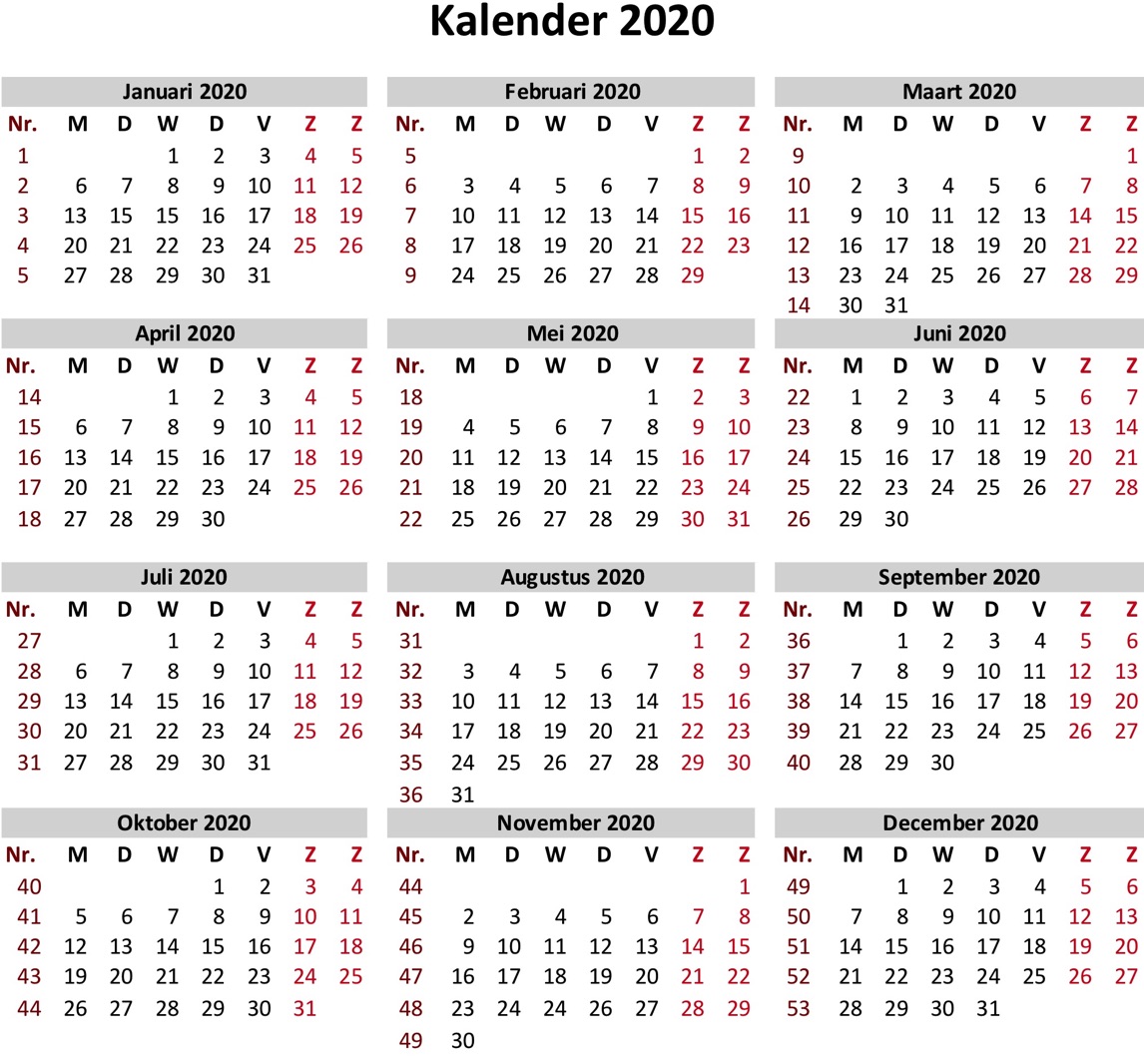
Het kritieke gebied is met . **2pt**

Nu ligt de gemeten waarde in dit kritieke gebied, dus wordt verworpen. **1pt**

**1c. [5pt]** Leg aan de dienstdoend logistiek onderofficier wat dit betekent voor de aan te houden maandelijkse voorraad, zonder daarbij te vervallen in voor hem onbegrijpelijke termen als “toets”, “kritiek gebied”, of “standaarddeviatie”.   
 is verworpen, dat betekent dat met 95% zekerheid

**Opgave 2 (Totaal 20 punten)**

*Triskaidekafobie* is de irrationele angst voor het getal 13 (uit het Grieks: triskadeka = 13, phobos = angst). Het is onderdeel van het Westerse bijgeloof om 13 als een ongeluksgetal te beschouwen. Zo ontbreekt in de meeste Amerikaanse flatgebouwen de 13e verdieping, is er in vliegtuigen vaak geen 13e rij en wordt in de autosport traditioneel geen nummer 13 toegewezen. NASA was zo onverstandig om nr. 13 in de Apollomissies niet over te slaan, waardoor twee dagen na de start een zuurstoftank ontplofte ("Okay Houston, I believe we've had a problem"). Vrijdag de dertiende wordt als een bijzonder ongelukkige dag beschouwd (kruisiging van Jezus op “Goede” vrijdag?).



**2a. [3pt]** Bereken de kans dat een vrijdag in 2020 de dertiende van de maand blijkt te zijn.

Er zijn 39 vrijdagen en twee vrijdag-de-dertien-den, dus deze kans is 2/39 = 0,05128.

**2b.** **[3pt]** Bereken de kans dat een dertiende van een maand in 2020 op een vrijdag valt.

Er zijn 12 dertienden van de maand waarvan twee op een vrijdag vallen, dus de kans is 2/12 = 0,1667.

**2c.** **[3pt]** Bereken de kans dat een dag in 2020 een vrijdag de dertiende is (let op, instinker).

Er zijn 366 dagen in 2020 (schrikkeljaar), waarvan twee vrijdagen de 13e, dus de kans is 2/366 = 0,005464

De kans dat een werkdag in 2020 op een vrijdag de dertiende valt is 0,0078. We nemen nu een steekproef van zes **verschillende** werkdagen, namelijk de zes dagen waarop de tentamens van Statistiek voor KW/MBW vielen (voor Statistiek delen 1 en 2 telkens een eerste, een tweede en een finale kans). Van deze zes dagen viel er één op vrijdag de 13e (vandaag)

**2d.** **[10pt]** Bereken op grond van deze steekproef een 95% schattingsinterval voor de kans dat een werkdag in 2020 op een vrijdag de dertiende valt.

We hebben en succesgeval op steekproefgrootte en we lossen met Clopper-Pearson op:

en

Dit levert op: en .

Het 95% schattingsinterval voor de kans dat een werkdag 2020 op een vrijdag valt, op basis van deze steekproef is dus [0,0042 , 0,6412].

Onder complotdenkers bij de cadetten en adelborsten gaat het gerucht dat er sprake is van kwade opzet van de faculteit door één van de zes tentamens op vrijdag de dertiende te plannen, want de kans dat dit gebeurt (één op de zes tentamens) is ruim twintig keer zo groot als de feitelijke kans (0,0078).

**2e**. Kijk nu of je deze complottheorie kunt onderbouwen door de volgende nul-hypothese te verwerpen bij een onbetrouwbaarheidsniveau van , op grond van de steekproef van zes dagen, waarvan er één vrijdag de dertiende is.

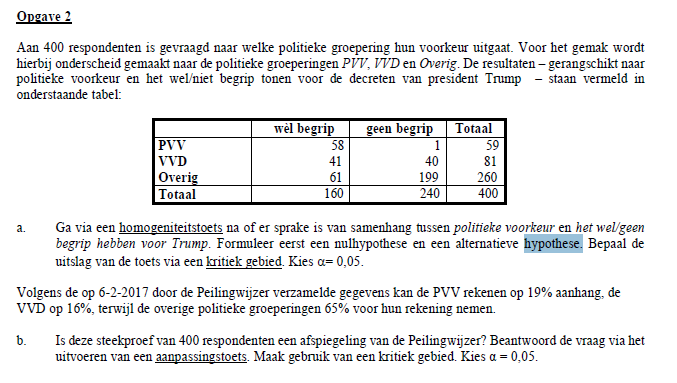
**: De verdeling van dagen die op een vrijdag de dertiende vallen gedraagt zich als een binomiale verdeling met een succeskans van (hoogstens) 0,0078.**

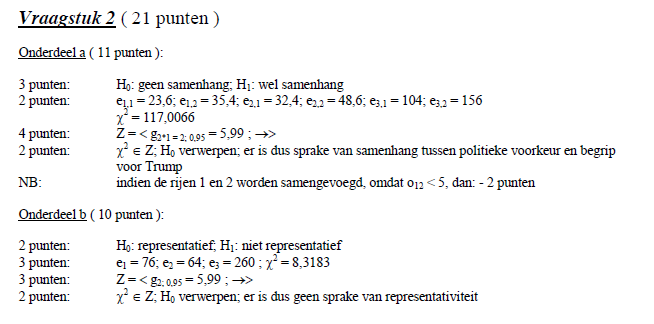
Bereken hiervoor de -waarde.

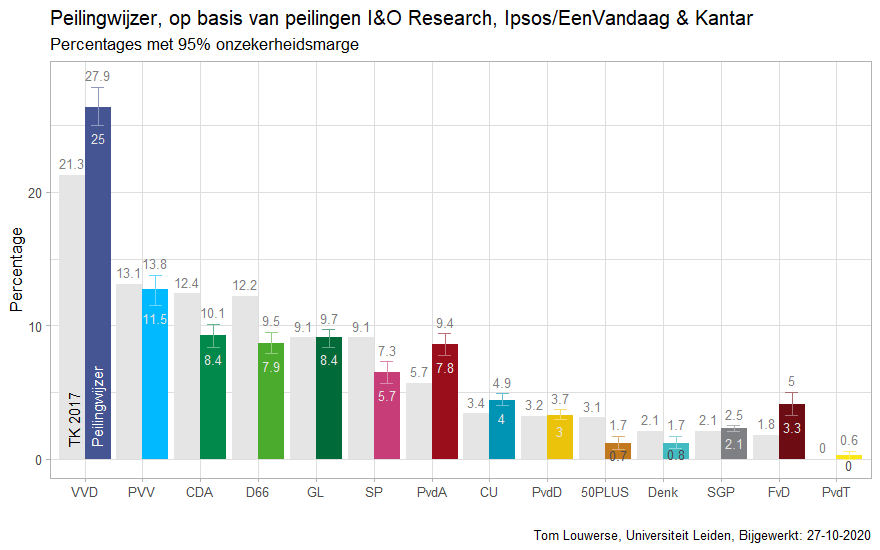
De -waarde is de kans dat onder aanname van de nulhypothese (de kans op een vrijdag de dertiende is 0.0078) de steekproef van

**2f.** Welke conclusie kun je met betrekking tot trekken als je het berekende schattingsinterval uit 2d vergelijkt met de kans dat een werkdag in 2020 op vrijdag de dertiende valt? Als je niet zeker bent over je antwoord van 2d, ga dan uit van het (fictieve) schattingsinterval [0,006 , 0,592].

De werkelijke waarde van de kans dat een werkdag een vrijdag de dertiende is (0,0078) ligt in het 95% schattingsinterval dat is gebaseerd op de steekproef van zest tentamendata







**\*Opgave 3 (Totaal 20 punten).** Het aantal brandmeldingen per week in een bepaalde stad is gedurende 100 weken geregistreerd (zie tabel).

|  |  |
| --- | --- |
| **Branden**  **per week** | **Frequentie** |
| 0 | 24 |
| 1 | 40 |
| 2 | 15 |
| 3 | 8 |
| ≥ 4 | 4 |
| Totaal | 91 |

**3a. [10pt]** Toets of het aantal branden per week is te beschouwen als een kansvariabele die een Poissonverdeling volgt met , door middel van uitrekenen van een -waarde. Kies als betrouwbaarheid 98%.

**3a.** We berekenen eerst de frequenties zoals ze uit de Poissonverdeling zouden volgen:

,

,

,

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

|  |  |  |
| --- | --- | --- |
| **Branden**  **per week** | **Frequentie**  **Observed** | **Frequentie**  **Expected** |
| 0 | 24 | 36,6 |
| 1 | 40 | 33,3 |
| 2 | 15 | 15,2 |
| ≥ 3 | 12 | 5,9 |
| Totaal | 91 | 91 |

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

We toetsen hiermee

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

H1: De waargenomen frequenties kunnen niet zo worden verklaard.

Dat kan het snelst door de -waarde uit te rekenen (met vrijheidsgraden):

Dit is kleiner dan dus H0 wordt verworpen, dus de tabel kan worden niet verklaard met de Poissonverdeling met met een betrouwbaarheid van 98%. Er zijn met name veel te veel weken met drie of meer branden.

**3b. [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

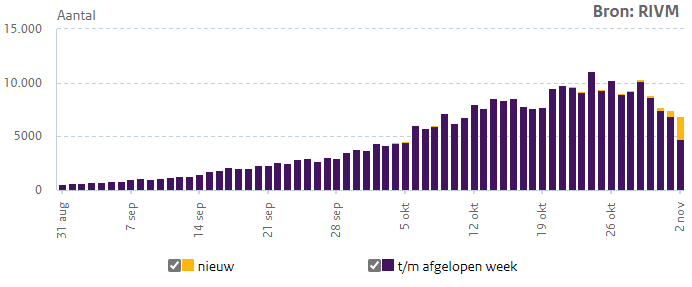
Dat geeft De waarde ligt in het kritieke gebied , dus H0 wordt verworpen.

**3c. [5pt]** Leg uit waarom je zou toetsen of de Poissonverdeling met de metingen verklaart.

**3c.** Een Poissonverdeling beschrijft verschijnselen die niet vaak, maar wel geregeld en onafhankelijk van elkaar optreden per vaste periode. Dat geldt waarschijnlijk/meestal wel voor branden, als er geen pyromanen actief zijn of seizoenseffecten zijn (bos-, schoorsteen- en barbecue-branden), maar dat is een stad minder relevant of kan uitmiddelen. In 100 weken waren er 91 branden, dat is gemiddeld per week, dat verklaart .

# Opgave 4 (Totaal 30 punten)

De tweede coronagolf van 2020 begon in september en leek vanaf 31 oktober over zijn hoogtepunt heen te zijn, er leek sprake te zijn van een daling van het aantal nieuwe besmettingen per dag.



In de onderstaande tabel vind je de waarden voor het aantal nieuwe besmettingen per dag vanaf 29 oktober (dag 0).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dag vanaf 29 oktober 2020** | **0** | **1** | **2** | **3** | **4** |
| **Aantal coronabesmettingen** | 10326 | 8805 | 7694 | 7445 | 6861 |

**4a [10pt]** Bereken met behulp van de tabel op het antwoordformulier de correlatiecoëfficiënt van Pearson. Bepaal of er sprake is van een lineaire correlatie tussen de dag van de meting en het aantal geconstateerde nieuwe besmettingen. Leg uit hoe daarbij het teken en de grootte van de berekende coëfficiënt een rol spelen.

4a.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **D** |  |  |  |  |  |
| **1** | **0** | **10326** | 0 | 0 | 10666276 |
| **2** | **1** | **8805** | 8805 | 1 | 77528025 |
| **3** | **2** | **7694** | 15388 | 4 | 59197636 |
| **4** | **3** | **7445** | 22335 | 9 | 55428025 |
| **5** | **4** | **6861** | 27444 | 16 | 47073321 |
| **Gem.** | 2 | **8226,2** | 147994,4 | 6 | 69170656,6 |

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 = dag van meting en = het aantal besmettingen op die dag.

De correlatiecoëfficiënt is

De correlatiecoëfficiënt is negatief, dus er is een negatieve correlatie, een dalende trend in het aantal nieuwe besmettingen.

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

**4b [10pt]** Bereken de lineaire regressielijn en bepaal hiermee een statistisch verantwoordde schatting voor 3 november (dag 5).

**4b.** De regressielijn is met

Vul dag 5 in: en je krijgt een bijbehorende voorspelling voor 3 november van .

**4c [10pt]** Op 3 november (dag 5) worden er 7826 nieuwe coronabesmettingen geconstateerd. In de pers wordt gesproken van een trendbreuk, een omslag van een daling in een sterke stijging. Voer een -aanpassingstoets met betrouwbaarheid 95% uit om te bepalen of er overeenkomst bestaat tussen geconstateerde nieuwe besmettingen sinds 29 oktober en de waarden die worden voorspeld door het lineaire model dat in 4b is bepaald. Voer deze toets uit door berekening van de -waarde en trek op grond hiervan een conclusie of de hoge waarde van 3 november kan duiden op een breuk met de eerder geconstateerde lineaire daling, of dat de waarde inderdaad kan wijzen op een trendbreuk. Doe deze berekening tweemaal, een keer voor de dagen 0 t/m 4 en een keer voor dagen 0 t/m 5. Leg uit wat je uit deze twee p-waarden betekenen. Maak voor je berekeningen gebruik van het antwoordformulier en vul de ontbrekende waarden aan.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dag (i)** | **0** | **1** | **2** | **3** | **4** | **5** | **Tot** |
| **Nieuwe besmettingen** | 10326 | 8805 | 7694 | 7445 | 6861 | 7826 | Niet invullen |
| **Voorspelde besmettingen** | 9884,2 | 9055,2 | 8226,2 | 7397,2 | 6568,2 | 758,77 | Niet invullen |
|  | 19,7474 | 6,9131 | 34,4311 | 0,3089 | 13,0526 | Niet invullen | 74,453 |
|  | 19,7474 | 6,9131 | 34,4311 | 0,3089 | 13,0526 | 758,7703 | 833,2233 |

Voor dag 1 t/m 4 geldt dat het aantal vrijheidgraden 5-1=4 is en

Voor dag 1 t/m 5 geldt dat het aantal vrijheidgraden 6-1=5 is en

De p-waarde is kleiner dan 0,05, dus de H0 wordt verworpen, dus de variabelen voldoen met 95% betrouwbaarheid niet aan het voorspelde verband. Aan de berekening zie je dat de laatste waarde een extreem grote bijdrage levert