# Oefententamen Statistiek KW/MBW deel 2

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 (30, 20, 20, 30 punten). Score = Puntentotaal/10

**Opgave 1 (Totaal 30 punten)**

Tijdens oefeningen wordt gebruik gemaakt van standaard NATO 24-uursrantsoenen. Deze zijn per 12 stuks verpakt in kartonnen dozen. In een depot van de logistieke dienst van waaruit deze dozen geleverd worden is van een aantal voorgaande maanden het aantal uitgeleverde dozen bijgehouden: 810,738, 621, 622, 505, 515, 389. Neem aan dat de aantallen $\underline{x}$ normaal verdeeld zijn.

**1a. [5pt]** Bereken het steekproefgemiddelde en de steekproefstandaarddeviatie van deze steekproef.

$\bar{x} = 605$ **2pt**

$s = 136{,}0882$ **3pt**

**1b. [10pt]** Bereken een 90% betrouwbaarheidsinterval voor $\mu$ 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.

Het betrouwbaarheidsinterval met betrouwbaarheid $\alpha$ voor $\mu$ waarbij $\sigma$ niet bekend is wordt bepaald door $\mu$ te schatten met het steekproefgemiddelde $\bar{x}$ van $n$ meetwaarden en dat als midden te nemen van een interval met linker- en rechteroverschrijdingskansen $\frac{1-\alpha}{2} = 0{,}05$: Bepaal eerst de $t$-waarde bij een rechteroverschrijdingskans van $\frac{1-\alpha}{2}$. Het gevraagde interval is dan $\left(\bar{x} - t\frac{s}{\sqrt{n}}, \bar{x} + t\frac{s}{\sqrt{n}}\right)$, **2pt**

want de geschatte standaarddeviatie van het gemiddelde van $n$ waarden is $\frac{s}{\sqrt{n}}$.

In ons geval is $n = 7$. **2pt**

Bij $\alpha = 0{,}9$ hoort bij een rechteroverschrijdingskans $\frac{1-\alpha}{2} = 0{,}05$ de waarde

$t = \text{invT}(1 - 0.05\,, 7 - 1) = 1{,}9432$, **2pt**

Dus het interval is $\left(\bar{x} - t\frac{s}{\sqrt{n}}, \bar{x} + t\frac{s}{\sqrt{n}}\right) = \left(605 - 1{,}9432\frac{136{,}0882}{\sqrt{7}}, 605 + 1{,}9432\frac{136{,}0882}{\sqrt{7}}\right) = (505{,}0486,\ 704{,}9514)$. **2pt**

Afronden op gehele getallen naar buiten (zodat het interval hooguit groter wordt): (505 , 705). **2pt**

**1c. [5pt]** De commandant van het depot vindt het zojuist bepaalde interval te onnauwkeurig, hij streeft naar een interval met een lengte van hoogstens 150 bij gelijkblijvende betrouwbaarheid. Bereken voor hoeveel maanden de uitgeleverde hoeveelheid pakjes bekend moet zijn om aan deze eis te voldoen. Als het niet mogelijk is, leg dan uit waarom dit niet kan.

De breedte van het betrouwbaarheidsinterval is $2t\frac{s}{\sqrt{n}}$. **1pt**

Door de steekproefgrootte $n$ te vergroten kan de breedte van het interval willekeurig klein worden gemaakt. **1pt**

De breedte van het interval moet zijn $2 \cdot \text{invT}(0.95\,, n-1)\frac{136{,}0882}{\sqrt{n}} = 150$. **1pt**

Let hierbij op dat $t$ ook afhankelijk is van $n$. Hieruit volgt met de solver van de GR dat $n = 10{,}8494$. Naar boven afronden voor alle zekerheid: $n = 11$. **2pt**

**1d. [5pt]** Toets: $H_0$: $\mu \geq 700$ tegen $H_1$: $\mu < 700$. Bepaal de toetsuitslag via het berekenen van een kritiek gebied. Kies hierbij α = 0,05 en ga uit van $\sigma = 130$.

De toetsingsgrootheid $\bar{x}$ is normaal verdeeld met gemiddelde μ = 700 (volgens de worst case situatie van de aanname $H_0$) en standaarddeviatie $\frac{130}{\sqrt{7}} = 49{,}1354$. **2pt**

Het kritieke gebied is $Z = (-\infty, g)$ met $g = \text{InvNorm}(0.05, 700, 49.1354) = 619{,}18$. **2pt**

Nu ligt de gemeten waarde $\bar{x} = 605$ in dit kritieke gebied, dus $H_0$ wordt verworpen. **1pt**

**1e. [5pt]** Bereken een 90% betrouwbaarheidsinterval voor $\sigma$ op grond van eerder genoemde steekproef

Schattingsinterval: $[\sigma_1\,, \sigma_2]$ met betrouwbaarheid $100(1-\alpha)\%$. Los op met de GR: $\chi^2\text{cdf}(0\,, g_L,\ n-1) = \frac{\alpha}{2}$ en vind $g_L = 1{,}6354$. Los op $\chi^2\text{cdf}(g_R,\ 10^{10},\ n-1) = \frac{\alpha}{2}$ en vind $g_R = 12{,}5916$. **3pt**

Het 90% schattingsinterval voor $\sigma$ is $\left[s\sqrt{\frac{n-1}{g_R}}\,, s\sqrt{\frac{n-1}{g_L}}\right] = [93{,}941\,, 260{,}6658]$ **2pt**

**Opgave 2 (Totaal 20 punten)**

De Rijksdienst voor het Wegverkeer doet onderzoek naar het aantal geregistreerde auto's zonder WA verzekering.

**2a. [10pt]** Bij een steekproef van 800 auto's bleken 69 auto's niet WA verzekerd te zijn. Bereken een 98% betrouwbaarheidsinterval voor de fractie onverzekerde auto's.

Auto's zijn wel of niet verzekerd met een vaste kans, dus we hebben te maken met een binomiale verdeling. We berekenen (Clopper-Pearson, zie slides) twee extreme waarden $p$, waarvoor

Stap 1: $P(\underline{k} \leq k) = \frac{0{,}02}{2}$ ofwel: $\text{binomcdf}(n = 800, p?, k = 69) = 0{,}01$ **3pt**

Stap 2: $P(\underline{k} \geq k) = \frac{0{,}02}{2}$ ofwel: $\text{binomcdf}(n = 800, p?, k = 69 - 1) = 0{,}99$ **3pt**

Met de GR-solver vinden we (let op dat de startwaarde tussen 0 en 1 ligt)

$$p_1 = 0{,}0646 \quad \text{en} \quad p_2 = 0{,}1121$$ **2pt**

Het 98% betrouwbaarheidsinterval voor de fractie onverzekerde auto's is dus:

$[0{,}0646\ ,\ 0{,}1121]$. **2pt**

**2b. [10pt]** Voor een steekproef van 28 onverzekerde auto's werd de gemiddelde leeftijd bepaald: 8,9 jaar met een steekproefstandaarddeviatie van 1,6 jaar. Bereken een 98% betrouwbaarheidsinterval voor de leeftijd van onverzekerde auto's.

We gaan voor de leeftijd van onverzekerde auto's uit van een normale verdeling, maar omdat de standaarddeviatie geschat is m.b.v. een steekproef en $n = 28 \leq 30$ moeten we de $t$-verdeling gebruiken. Bereken de $t$-waarde door op te lossen

$$\text{tcdf}(t?\ ,\ 10^{10}\ ,\ 28 - 1) = 0{,}01.$$ **3pt**

Dit levert met de GR solver: $t = 2{,}4727$. **3pt**

Het 98% betrouwbaarheidsinterval is dus

$$\left[\bar{x} - \frac{ts}{\sqrt{n}}\ ,\bar{x} + \frac{ts}{\sqrt{n}}\right] = \left[8{,}9 - \frac{2{,}4727 \cdot 1.6}{\sqrt{28}}\ ,\ 8{,}9 + \frac{2{,}4727 \cdot 1.6}{\sqrt{28}}\right] = [8{,}1523\ ,\ 9{,}6477] \text{ jaar}$$ **4pt**

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

| Branden per week | Frequentie |
|---|---|
| 0 | 24 |
| 1 | 40 |
| 2 | 15 |
| 3 | 8 |
| $\geq 4$ | 4 |

**3a. [3pt]** Gedurende hoeveel weken is er geregistreerd en hoeveel branden zijn er totaal waargenomen?

De som van de frequenties is het aantal weken: 24 + 40 + 15 + 8 + 4 = 91 weken.

Het aantal branden kun je niet exact bepalen, maar is minimaal $0 \cdot 24 + 1 \cdot 40 + 2 \cdot 15 + 3 \cdot 8 + 4 \cdot 4 = 110$ branden, het kan alleen zijn dat er in de laatste categorie weken waren met meer dan 4 branden.

**3b. [10pt]** Toets of het aantal branden per week is te beschouwen als een kansvariabele die een Poissonverdeling volgt met $\mu = 1{,}22$ branden per week, door middel van uitrekenen van een $p$-waarde. Kies als betrouwbaarheid 95% en gebruik in je berekening de verwachte frequenties in één decimaal nauwkeurig.

We berekenen eerst de frequenties zoals ze uit de Poissonverdeling zouden volgen:
$P(\underline{k} = 0) = \text{poissonpdf}(1.22\ ,\ 0) = 0{,}2952, \quad E_0 = 0{,}2952 \cdot 91 = 26{,}9$
$P(\underline{k} = 1) = \text{poissonpdf}(1.22\ ,\ 1) = 0{,}3602, \quad E_1 = 0{,}3602 \cdot 91 = 32{,}8$
$P(\underline{k} = 2) = \text{poissonpdf}(1.22\ ,\ 2) = 0{,}2197, \quad E_2 = 0{,}2197 \cdot 91 = 20{,}0$
$P(\underline{k} = 3) = \text{poissonpdf}(1.22\ ,\ 3) = 0{,}08935, \quad E_3 = 0{,}08935 \cdot 91 = 8{,}1$
$P(\underline{k} \geq 4) = 1 - P(\underline{k} \leq 3) = 1 - \text{poisson}\mathbf{cdf}(1.22\ ,\ 3) = 0{,}03553, \ E_4 = 0{,}03553 \cdot 91 = 3{,}2$

Omdat de verwachte frequentie $E_4 = 3{,}2$ te klein is voor toepassing van Chikwadraat (< 5) nemen we de laatste twee categorieën samen tot $\underline{k} \geq 3$, waardoor de verwachting daarvan groter dan 5 wordt (zie tabel hieronder).

| Branden per week | Frequentie Observed | Frequentie Expected |
|---|---|---|
| 0 | 24 | 26,9 |
| 1 | 40 | 32,8 |
| 2 | 15 | 20,0 |
| ≥ 3 | 12 | 11,3 |
| Totaal | 91 | 91 |

Kijken of $E_i$ en $O_i$ voldoende op elkaar lijken doen we met een $\chi^2$ aanpassingstoets. De toetsingsgrootheid is

$$\chi^2 = \sum_{i=0,1,2,3} \frac{(O_i - E_i)^2}{E_i} = 0{,}313 + 1{,}580 + 1{,}250 + 0{,}043 = 3{,}186$$

We toetsen hiermee

$H_0$: De waargenomen frequenties kunnen worden verklaard met een Poissonverdeling met $\mu = 1{,}22$.

$H_1$: De waargenomen frequenties kunnen niet zo worden verklaard.

Dat kan het snelst door de $p$-waarde uit te rekenen (met $4 - 1 = 3$ vrijheidsgraden):

$$p = P\left(\underline{\chi}^2 > 3{,}186\right) = \chi^2\text{cdf}(3.186\ ,\ 10^{10}\ ,\ 3) = 0{,}3638.$$

Deze kans is niet kleiner dan $\alpha = 0{,}05$, dus $H_0$ wordt niet verworpen, dus de tabel kan met een betrouwbaarheid van 95% worden verklaard met de Poissonverdeling met $\mu = 1{,}22$.

**3c. [5pt]** 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

$$\chi^2\text{cdf}(g?\ ,\ 10^{10}\ ,\ 3) = 0{,}02$$

Dat geeft $g = 9{,}8374$. De waarde $\chi^2 = 12{,}0490$ ligt in het kritieke gebied $(9{,}8374\ , \infty)$, dus $H_0$ wordt verworpen.

**3d. [2pt]** Leg uit waarom je zou toetsen of de Poissonverdeling met $\mu = 1{,}22$ de metingen verklaart.

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 91 weken waren er minstens 110 branden, maar waarschijnlijk niet erg veel meer, want de kans op 4 branden is groter dan de kans op 5 branden. Dit levert gemiddeld $\mu = 1{,}21$ per week. De waarde is ietsje groter genomen om af en toe iets meer dan 4 branden te verklaren.

**Opgave 4 (Totaal 30 punten)**

In het kader van het verminderen van milieubelasting en het verbeteren van efficiëntie werd binnen Defensie de actie Paper Tiger gehouden. Tijdens deze actie werd gewerkt aan bewustwording bij het personeel en werden maatregelen genomen om het gebruik van papier in de bedrijfsvoering terug te dringen. Om het effect van de actie te evalueren werd bij zes depots de uitgifte van papier vlak voor en direct na de actie gemeten. In de tabel hieronder is het aantal pallets papier weergegeven dat in een maand door elk depot werd verstrekt.

| **Depot** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| **Uitgifte voor Paper Tiger** | 20 | 26 | 18 | 29 | 28 | 23 |
| **Uitgifte na Paper Tiger** | 13 | 22 | 6 | 20 | 17 | 12 |

**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 aantallen pallets papier dat door deze depots per maand vóór de actie werd verstrekt en het aantal daarna. Leg uit hoe daarbij het teken en de grootte van de berekende coëfficiënt een rol spelen.

De uitgifte vóór Paper Tiger is de verklarende variabele, die kies je als X, de uitgifte erna is het effect dat is Y.

| **Depot** | $\underline{X}$ | $\underline{Y}$ | $\underline{XY}$ | $\underline{X^2}$ | $\underline{Y^2}$ |
|---|---|---|---|---|---|
| **1** | **20** | **13** | 260 | 400 | 169 |
| **2** | **26** | **22** | 572 | 676 | 484 |
| **3** | **18** | **6** | 108 | 324 | 36 |
| **4** | **29** | **20** | 580 | 841 | 400 |
| **5** | **28** | **17** | 476 | 784 | 289 |
| **6** | **23** | **12** | 276 | 529 | 144 |

| | | | | | |
|---|---|---|---|---|---|
| **Gem.** | 24 | 15 | 378,6667 | 592,3333 | 253,6667 |

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 $X$ = uitgifte van papier vóór het PT en $Y$ = de uitgifte erna.
De correlatiecoëfficiënt is

$$r = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\sqrt{\left(\overline{X^2} - \overline{X}^2\right)\left(\overline{Y^2} - \overline{Y}^2\right)}} = 0{,}8627$$

want $\overline{X} = 24, \overline{Y} = 15, \overline{XY} = 378{,}6667, \overline{X^2} = 592{,}3333, \overline{Y^2} = 253{,}6667$
De correlatiecoëfficiënt is positief, dus er is een positieve correlatie (d.w.z. bij een grotere uitgifte vóór PT hoort een grotere uitgifte erna, het lineaire verband tussen uitgifte voor en na is een rechte lijn die stijgend is.
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.

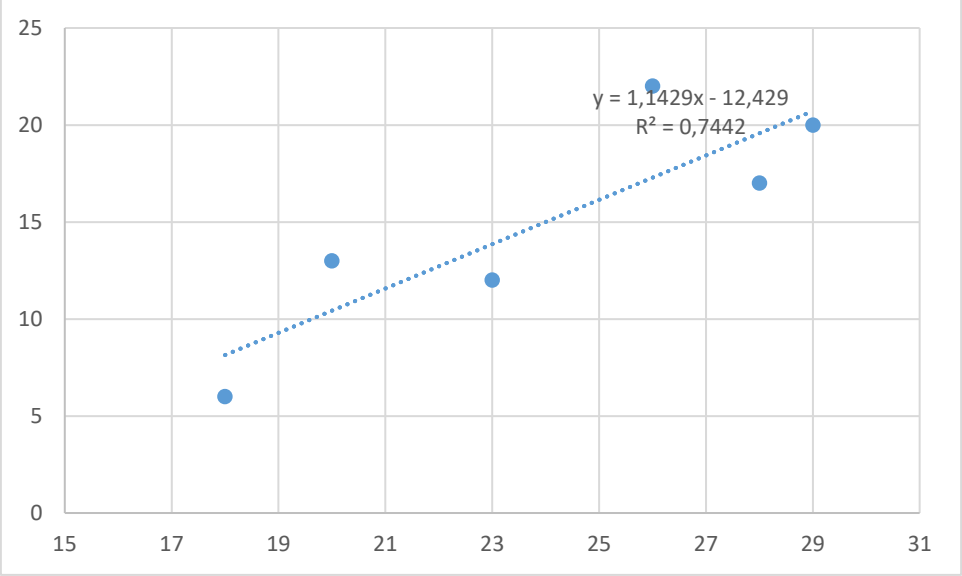
**4b [10pt]** In een zevende depot zijn in de maand vóór de actie 26 pallets papier uitgegeven. Bereken de regressielijn en bepaal hiermee een statistisch verantwoorde voorspelling van het aantal verstrekte pallets papier in dit depot in de maand na Paper Tiger. Rond je antwoord af op gehele pallets.

De regressielijn is $Y = aX + b$ met

$$a = \frac{\overline{X \cdot Y} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2} = 1{,}14286$$

$$b = \overline{Y} - a\overline{X} = -12{,}42857$$

Vul verkoop vóór WK in: $X_0 = 26$ in en je krijgt een bijbehorende voorspelling na WK van $Y_0 = 17{,}2857$.



**4c [5pt]** Bereken een 96% voorspellingsinterval voor de waarde die in 4b is berekend.

Het voorspellingsinterval is $[Y_0 - ts_f\ , Y_0 + ts_f]$
$t$ is de $t$-waarde die hoort bij de opgegeven betrouwbaarheid met $\nu = \boldsymbol{n} - \mathbf{2}$ vrijheidsgraden.
Bij een betrouwbaarheid van 96% is de linker overschrijdingskans 0,96 + 0,04/2 = 0,98 en

$$t = \text{invT}(0.98\ , 4) = 3{,}0000$$

$$s_\varepsilon = \sqrt{\frac{n}{n-2}\left(\overline{Y^2} - a\overline{XY} - b\overline{Y}\right)} = 3{,}316$$

$$u = \frac{1}{n}\left(1 + \frac{\left(X_0 - \overline{X}\right)^2}{\overline{X^2} - \overline{X}^2}\right) = 0{,}177$$

$$s_f = s_\varepsilon\sqrt{u+1} = 3{,}598$$

$$[Y_0 - ts_f\ , Y_0 + ts_f] = [6.492\ , 28{,}080\ ]$$

**4d [5pt]** Voer een homogeniteitstoets met betrouwbaarheid 95% uit om te bepalen of er onafhankelijkheid bestaat tussen de variabelen uitgifte voor/na Paper Tiger en het depot, door berekening van de $p$-waarde.

Bereken de marginale totalen en hiermee de frequenties in de tabel op grond van onafhankelijkheid, bv. 143*33/234 = 20,1667, etc.

| **Expected** | **1** | **2** | **3** | **4** | **5** | **6** | **Tot** |
|---|---|---|---|---|---|---|---|
| **Voor PT** | 20,1667 | 29,3333 | 14,6667 | 29,9444 | 28,1111 | 20,7778 | 143 |
| **Na PT** | 12,8333 | 18,6667 | 9,3333 | 19,0556 | 17,8889 | 13,2222 | 91 |
| **Totaal** | 33 | 48 | 24 | 49 | 46 | 34 | 234 |

Bereken de $\chi^2$-waarde uit de twee tabellen:
$\chi^2 =$ 0,00138 + 0,37878 + 0,75756 + 0,02978 + 0,00044 + 0,07189 +
0,00217 + 0,59523 + 1,19046 + 0,04680 + 0,00069 + 0,11297 = 3,1882
Bereken de $p$-waarde:

$$p = P\left(\underline{\chi}^2 > 3{,}1882\right) = \chi^2(3.1882\ , 10^{10}\ , 5) = 0{,}67099$$

De $p$-waarde is niet kleiner dan 0,05, dus de $H_0$ wordt niet verworpen, dus de variabelen zijn met 95% betrouwbaarheid onafhankelijk.