Oefentamen Statistiek KW/MBW (deel 2)

Afdeling: Propedeuse KW/MBW 2020-2021
Examinator: Dr. J.B.M. Melissen
Datum: 15 juli 2021, 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)

Tijdens een landmachtoefening wordt een week lang dagelijks bijgehouden hoeveel liter drinkwater wordt verbruikt. Dit leidt tot de volgende waarden: 5215, 5879, 7021, 3983, 3974, 6003, 7972. Neem aan dat de hoeveelheden $\underline{x}$ (kansvariabele die het waterverbruik 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.

$$\overline{x} = 5721{,}0000$$
$$s = 1482{,}6878$$

1b. [7pt] Bereken een 96% betrouwbaarheidsinterval voor het verwachte dagelijkse waterverbruik μ 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 96% betrouwbaarheid gewaarborgd blijft.

Omdat de standaarddeviatie niet bekend is (en bovendien de steekproefgrootte kleiner dan 30) moet de t-verdeling worden gebruikt. De t-waarde bij 96% betrouwbaarheid voor een tweezijdig interval is $t = \text{invT}\left(0.96 + \frac{1-0.96}{2}, 7-1\right) = 2{,}6122.$

De linkeroverschrijdingskans van t is dan 0,96 in het betrouwbaarheidsinterval plus de helft van de overige 0,04 (=0,02) in het interval links daarvan (totaal 0,98).

Het gemiddelde van 7 dagwaarden (steekproefgemiddelde) is normaal verdeeld met gemiddelde μ en standaarddeviatie $\sigma/\sqrt{7}$. We gebruiken hiervoor de schattingen $\overline{x}$ en $s/\sqrt{7}$ en de t-verdeling voor de berekening.

Het betrouwbaarheidsinterval van μ is dan

$$\left[\overline{x} - t\frac{s}{\sqrt{7}}\ , \overline{x} + t\frac{s}{\sqrt{7}}\right] = [4257{,}1142\ ;\ 7184{,}8857].$$

Afronden mag het interval niet kleiner maken, dus afronden naar buiten: $[4257\ ,7185]$.

1c. [8pt] Toets: H_0: $\mu \leq 4630$ tegen H_1: $\mu > 4630$. Bepaal de toetsuitslag door het berekenen van een kritiek gebied op basis van de gegeven steekproef van zeven dagen waterverbruik. Kies in dit geval als onbetrouwbaarheid α = 0,05.
Leg in simpele bewoordingen uit wat de uitslag van deze toets betekent voor het dagelijks waterverbruik.
We zoeken de (kleinste) grens g van het kritieke gebied (g, ∞), zodanig dat de kans op een fout van de eerste soort (je verwerpt H_0, terwijl H_0 toch waar is) kleiner is dan α = 0,05:

$$P(\underline{\overline{x}} > g? \mid H_0) \leq 0{,}05$$

$\underline{\overline{x}}$ is hierin het waterverbruik per dag, gemiddelde over zeven dagen, dat is normaal verdeeld met gemiddelde $\mu \leq 4630$ (want aangenomen wordt dat H_0 geldt) en standaarddeviatie $\sigma/\sqrt{7}$.
We nemen de worst case situatie wanneer H_0 geldt: $\mu = 4630$, want

$$P(\underline{\overline{x}} > g? \mid H_0) = P(\underline{\overline{x}} > g? \mid \mu \leq 4630) \leq P(\underline{x} > g? \mid \mu = 4630),$$

Want als de rechter term $\leq 0{,}05$, dan geldt dat zeker ook voor de linker:

$$P\left(\underline{\overline{x}} > g? \mid \underline{\overline{x}} \sim N(\mu = 4630, \sigma/\sqrt{7})\right) \leq 0{,}05$$

Voor σ gebruiken de schatting s en daarom moeten we gebruik maken van de t-verdeling.

$$P(\underline{x} > g?) = 1 - P(\underline{x} \leq g?) = 1 - P\left(\underline{t} = \frac{\underline{x} - 4630}{(s/\sqrt{7})} \leq \frac{g? - 4630}{(s/\sqrt{7})}\right)$$
$$= 1 - \text{tcdf}\left(t = \frac{g? - 4630}{(s/\sqrt{7})}, 7 - 1\right) \leq 0{,}05$$

Er geldt dus (los op met … $= 0{,}05$): $t = \text{invT}(0.95\ , 7 - 1) = 1{,}9432$,
dus $g = 4630 + ts/\sqrt{7} = 5718{,}9757$.
(Let op dat de berekening van de t-waarde anders gaat dan in 1b, omdat daar het interval tweezijdig was en hier enkelzijdig).
De steekproefwaarde is groter dan deze waarde, ligt dus in het kritieke gebied en H_0 wordt verworpen.
Dit betekent dat met 95% betrouwbaarheid kan worden gesteld dat het dagelijks waterverbruik minimaal 4630 liter is.

1d. [5pt] Hoeveel liter water 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 g, dan is de kans dat deze hoeveelheid voldoende is (het werkelijke waterverbruik is dan maximaal g) minstens 98%, dus
$P(\underline{x} \leq g?) \geq 0{,}98$, dus, worst case is $P(\underline{x} \leq g?) = 0{,}98$.
Nu is $\underline{x}$, het waterverbruik van één dag, normaal verdeeld met gemiddelde μ en standaarddeviatie σ. We gebruiken hiervoor de schattingen $\overline{x}$ en s en daarom dus de t-verdeling.
We zoeken een linkszijdig interval met kans 0,98. Dat correspondeert met een t-waarde van $t = \text{invT}(0.98\ , 7 - 1) = 2{,}6122$, dus $g = 5721 + ts = 9594{,}0771$.

Rond naar boven af: 9595 liter.

1e. [5pt] Hoeveel liter water 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 $\underline{y}$ het totale waterverbruik van één week, dat is normaal verdeeld met gemiddelde 7μ en standaarddeviatie $\sqrt{7}\sigma$ (want μ en σ gelden per dag). We gebruiken hiervoor de schattingen $7\overline{x}$ en $\sqrt{7}s$ en dus weer de t-verdeling.
Noem de gezochte minimale wekelijkse hoeveelheid w, dan is de kans dat het werkelijke waterverbruik deze hoeveelheid voldoende is minstens 98%, dus
$P(y \leq w?) \geq 0{,}99$, dus, worst case is $P(y \leq w?) = 0{,}98$.
We zoeken een linkszijdig interval met kans 0,98. Dat correspondeert weer met een t-waarde van $t = \text{invT}(0.98\,, 7-1) = 2{,}6122$, dus $w = 7 \times 5721 + t\sqrt{7}s = 50294{,}1987$.
Rond naar boven af: 50295 liter.

De eis is nu veel minder streng dan volgens 1d, want dan zou je elke dag aan een eis moeten voldoen, en wat je dagelijks over houdt zou je in die berekening niet gebruiken voor een volgende dag. Als je de hele weekvoorraad ter beschikking hebt en alleen aan het eind van de week hoeft uit te komen heb je genoeg aan een kleinere voorraad.

Opgave 2 (Totaal 20 punten).

Het aantal acute meldingen per week waarvoor de Explosieven Opruimingsdienst moest uitrukken is gedurende 40 weken geregistreerd (zie tabel).

Meldingen per week	Frequentie
0	10
1	16
2	9
3	5
≥ 4	2
Totaal	42

2a. [4pt] Leg uit waarom het aannemelijk is dat aantal meldingen per week beschreven kan worden door een Poissonverdeling. Welke waarde van μ kies je daarbij?
De Poissonverdeling is een benadering van de binomiale verdeling waarbij n (steekproefgrootte) groot wordt en de succeskans p klein, waarbij tegelijkertijd het verwachte aantal successen $np = \mu$ vast is. De Poissonverdeling beschrijft dus verschijnselen die wel of niet kunnen optreden, waarbij het wel of niet optreden wordt bepaald door zeer veel kleine factoren die allemaal met een zeer kleine kans tot succes (optreden) kunnen leiden en wel zo dat het effect gemiddeld μ keer optreedt. Je kunt je voorstellen dat het vinden van heel veel factoren afhangt: heel veel mensen die in principe de niet ontplofte bom zouden kunnen vinden, maar dan moeten ze wel op het juiste moment op de goede plek zijn, er oog voor hebben, onderkennen waarmee ze te maken hebben en er ook melding van maken, en niet in de tussentijd een hartaanval of verkeersongeval krijgen, of een defecte mobiel, etc. Verder is een belangrijke aanname dat al die effectjes onafhankelijk van elkaar zijn. Dat is niet altijd zo, maar dat zit dan in de benadering die je maakt.
Uit de tabel zie je dat er 42 meldingen waren in 40 weken. Je neemt de week als referentieperiode, dus er zijn gemiddeld $\mu = 42/40 = 1{,}05$ meldingen per week.

2b. [10pt] Toets of het aantal meldingen per week is te beschouwen als een kansvariabele die een Poissonverdeling volgt. Doe deze toetsing door middel van uitrekenen van de p-waarde. Kies als betrouwbaarheid 98%.
Bereken de verwachte frequenties met de Poissonverdeling:

$$42 \times \text{poissonpdf}(\mu = 1.05\,, k = 0) = 14{,}6974, \quad \text{etc.}$$

Meldingen per week	Frequentie observed	Frequentie expected
0	10	14,6974
1	16	15,4323
2	9	8,1019
3	5	2,8357
≥ 4	2	0,9327
Totaal	42	42

De laatste twee entries in de tabel hebben een te kleine verwachte waarde (<5). Dit kun je oplossen door de laatste drie rijen samen te voegen.

Meldingen per week	Frequentie observed	Frequentie expected
0	10	14,6974
1	16	15,4323
≥ 2	16	11,8703
Totaal	42	42

De toetsingsgrootheid is

$$\chi^2 = \frac{(10-14{,}6974)^2}{14{,}6974} + \frac{(16-15{,}4323)^2}{15{,}4323} + \frac{(16-11{,}8703)^2}{11{,}8703}$$
$$= 1{,}5013 + 0{,}0209 + 1{,}4367 = 2{,}9589.$$

De overschrijdingskans van deze waarde is

$$p = \chi^2\text{cdf}(2.9589\,, 10^{10}, \nu = 3-1) = 0{,}2278$$

Deze waarde is niet kleiner dan de gegeven $\alpha = 0{,}02$ dus H_0 kan niet worden verworpen.
Je kunt nu met 98% betrouwbaarheid zeggen dat er geen reden is om eraan te twijfelen dat de geobserveerde frequenties door een Poissonverdeling met $\mu = 1{,}05$ worden beschreven.
Je moet wel voorzichtig zijn met je conclusie, want voor de berekening was het wel nodig om rijen samen te voegen. Het zou dus kunnen zijn dat de originele tabel niet goed wordt beschreven, maar de samengevoegde wel. Dit probleem zou je niet hebben als H_0 was verworpen, want als de samengevoegde tabel niet goed wordt beschreven, dan de originele zeker niet.

2c. [6pt] Bereken een 95% betrouwbaarheidsinterval voor μ.
Je hebt de waarde van μ bepaald door het totaal aantal meldingen te delen door het aantal weken. Volgens het formuleoverzicht kun je het 95% betrouwbaarheidsinterval uitrekenen door oplossen van

$$\text{poissoncdf}(n \cdot \mu_1?, k-1) = 1 - \frac{\alpha}{2} \quad \text{Hierin is } k = \sum_{i=1}^{n} k_i$$
$$\text{poissoncdf}(n \cdot \mu_2?,\ k) = \frac{\alpha}{2}$$

m.a.w.

$$\text{poissoncdf}(40 \cdot \mu_1?, 42-1) = 1 - \frac{0{,}05}{2}$$
$$\text{poissoncdf}(40 \cdot \mu_2?,\ 42) = \frac{0{,}05}{2}$$

Dit levert het 95% betrouwbaarheidsinterval $[0{,}7567\ ;\ 1{,}4192]$

Opgave 3 (Totaal 20 punten).

Uit gegevens van het RIVM is een overzicht gemaakt van de vijftigplussers die tot en met mei 2021 zijn overleden aan corona (drie kolommen: Overleden man / vrouw / totaal), uitgesplitst naar geslacht en leeftijd.
Verder is uit gegevens van het CBS op basis van de overleden totalen per leeftijdscategorie (kolom Overleden totaal) geschat hoe deze totalen normaal gesproken (buiten corona) zouden zijn verdeeld over mannen en vrouwen (zie kolommen Overleden verwacht m / v).

Leeftijd	Overleden man	Overleden vrouw	Overleden totaal	Overleden verwacht m	Overleden verwacht v
50-59	246	124	370	217	153
60-69	820	463	1283	752	531
70-79	2774	1554	4326	2538	1788
80-89	4140	3389	7529	3259	4270
90-99	1335	2246	3581	1550	2031
Totaal	9315	7776	17089	8316	8773

3a. [10pt] Voer een χ^2 homogeniteitsanalyse uit op de kolommen “Overleden man” en “Overleden vrouw”. Bereken daarvoor de waarde van χ^2 en het kritieke gebied bij een betrouwbaarheid van 99%.
We gaan de homogeniteit (onafhankelijkheid van geslacht en leeftijd) in de volgende tabel bekijken:

Leeftijd	Overleden man	Overleden vrouw	Overleden totaal
50-59	246	124	370
60-69	820	463	1283
70-79	2774	1554	4326
80-89	4140	3389	7529
90-99	1335	2246	3581
Totaal	9315	7776	17089

De verwachte waarden bij onafhankelijkheid zijn:

Leeftijd	Overleden man	Overleden vrouw	Overleden totaal
50-59	201,6824	168,3609	370
60-69	699,3472	583,8029	1283
70-79	2358,048	1968,458	4326
80-89	4103,964	3425,917	7529
90-99	1951,958	1629,461	3581
Totaal	9315	7776	17089

De toetsingsgrootheid heeft de volgende waarden:

9,738345 11,68853
20,81525 24,99704
73,37241 87,2639
0,31643 0,397821

195,0029 233,28

Het totaal is $\chi^2 = 656{,}872578$.
Het kritieke gebied is het interval $(g?, \infty)$, waarbij g voldoet aan

$$\chi^2\text{cdf}(0, g?, \nu = (2-1)(5-1) = 4) = 0{,}99$$

Met de GR Solver levert dit $g = 13{,}2767$.
De waarde van de toetsingsgrootheid ligt hier ver boven, dus H_0 moet worden verworpen. Dat betekent dat er met een betrouwbaarheid van 99% een afhankelijkheid tussen leeftijd en geslacht is geconstateerd.
De belangrijkste reden volgt uit de losse waarden van χ^2: in de groep 90-99 zijn naar verhouding veel meer vrouwen overleden dan mannen. Dat kan komen omdat er in die leeftijdsgroep veel meer vrouwen van mannen zijn, vrouwen worden gemiddeld ouder dan mannen.

3b. [10pt] Voer een χ^2 aanpassingsanalyse uit op de kolommen "Overleden man" en "Overleden verwacht man". Bereken daarvoor de waarde van χ^2 en de p-waarde en houd een betrouwbaarheid van 99% aan.

Leeftijd	Overleden man	Overleden verwacht m
50-59	246	217
60-69	820	752
70-79	2774	2538
80-89	4140	3259
90-99	1335	1550
Totaal	9315	8316

$$\chi^2 = \frac{(246-217)^2}{217} + \cdots = 3{,}8756 + 6{,}1489 + 21{,}9448 + 238{,}1593 + 29{,}8226 = 299{,}9512$$

Het kritieke gebied is het interval $(g?, \infty)$, waarbij g voldoet aan

$$\chi^2\text{cdf}(0, g?, \nu = 5) = 0{,}99$$

Het aantal vrijheidsgraden is hier niet 1 minder, omdat de totalen van de twee kolommen niet gelijk zijn (dat is normaal de reden dat je één vrijheidsgraad verliest).
Met de GR Solver levert dit $g = 15{,}0863$.
De berekende toetsingsgrootheid ligt in het kritieke gebied, dus H_0 wordt verworpen. Er is dus bij 50+ mannen geen goede overeenkomst tussen de sterftecijfers in corona met de geschatte sterftecijfers zonder corona. Je kunt wel stellen dat er door corona meer 50+ mannen zijn overleden. De grootste afwijking zit in de groep van 80-89 jarigen waar bijna 900 mannen meer dan normaal overleden.

Opgave 4 (Totaal 30 punten)

In de tabel hieronder is van zes studenten het eindexamencijfer wiskunde en het eindcijfer Statistiek vermeld.

Student →	1	2	3	4	5	6
Eindexamencijfer wiskunde	8	6	7	6	9	8
Cijfer Statistiek	7,6	5,7	7,5	5,8	8,8	7,2

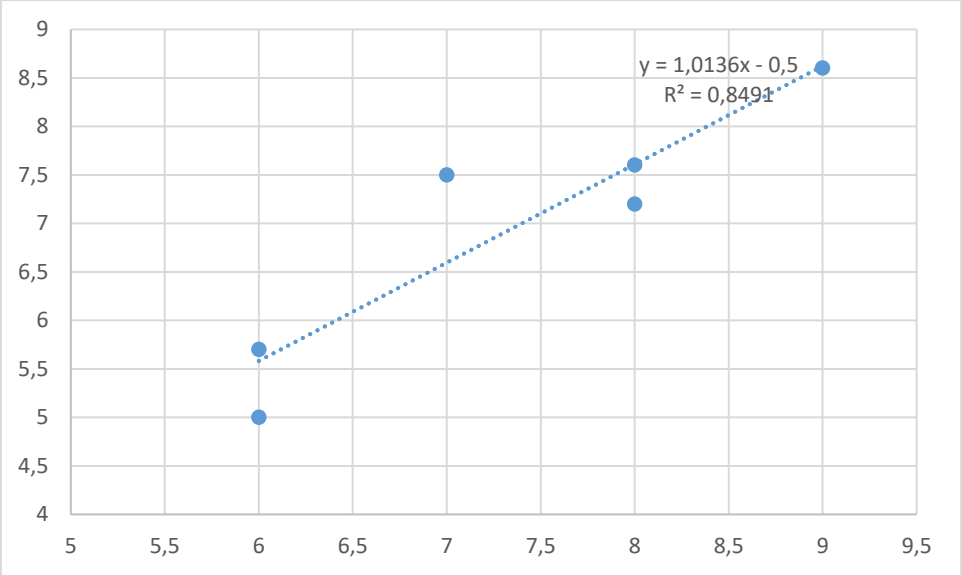
4a [8pt] Bereken handmatig de correlatiecoëfficiënt van Pearson. Bepaal of er sprake is van een lineaire correlatie tussen de twee cijfers.

x	y	x^2	y^2	xy
8	7,6	64	57,76	60,8
6	5,7	36	32,49	34,2
7	7,5	49	56,25	52,5
6	5	36	25	30
9	8,6	81	73,96	77,4
8	7,2	64	51,84	57,6
7,3333	6,9333	55,0000	49,5500	52,0833

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sqrt{\left(\overline{X^2} - \bar{X}^2\right)\left(\overline{Y^2} - \bar{Y}^2\right)}} = 0{,}9215$$

4b [12pt] Bereken de regressielijn, maak een schets van de puntenwolk en en bepaal hiermee een voorspelling van het verwachte Statistiekcijfer dat een student haalt als hij op zijn eindexamen een 5 zou hebben gehaald.

$$a = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2} = 1{,}0136$$
$$b = \bar{Y} - a\bar{X} = -0{,}5000$$



$$X_0 = 5$$
$$Y_0 = aY_0 + b = 4{,}5680$$

4c [10pt] Bereken een 95% voorspellingsinterval voor het cijfer Statistiek van de student die een 5 op zijn eindexamen haalde voor wiskunde.

$$t = \text{invT}(0.975\ , 6-2) = 2{,}7762$$

$$u = \frac{1}{n}\left(1 + \frac{\left(X_0 - \overline{X}\right)^2}{\overline{X^2} - \overline{X}^2}\right) = 0{,}9088$$

$$s_\varepsilon = \sqrt{\frac{n}{n-2}\left(\overline{Y^2} - a\overline{XY} - b\overline{Y}\right)} = 0{,}5810$$

$$s_f = s_\varepsilon\sqrt{u+1} = 0{,}8027$$

Het 95% voorspellingsinterval is nu

$$\left[Y_0 - ts_f\ , Y_0 + ts_f\right] = [2{,}3395\ ;\ 6{,}7965]$$

======== XXXXXXXX ========