Ministerie van Defensie

Faculteit Militaire Wetenschappen

Gegevens student	
Naam:	
Peoplesoftnummer:	
Klas:	
Handtekening:	

Algemeen			
Vak:	Statistiek (deel 2)	**Vakcode:**	STA#2
Datum:	25 juli 2025	**Tijdsduur:**	9:00-12:00
Examinator:	Dr. ir. D.A.M.P. Blom	**Aantal pagina's:**	4
Peer-review:	Dr. M.P. Roeling	**Aantal opgaven:**	4

Algemene instructies

- Alle antwoorden dienen gemotiveerd te worden. Indien u een deelopgave niet kunt oplossen en het antwoord in vervolgvragen nodig hebt, probeer uit te gaan van een redelijke fictieve waarde.
- Rond je antwoorden waar nodig af op vier decimalen.
- U mag een grafische rekenmachine gebruiken zonder CAS (Computer Algebra Systeem).
- Antwoorden, in welke vorm dan ook, mogen de zaal niet verlaten.
- Vermeld op elk antwoordvel je naam, Peoplesoft-nummer en maak een nummering van je antwoordvellen.
- Iedere vorm van mobiele (potentiële) datadragers (telefoon, smartwatch, etc.) of andere vormen om te frauderen (bv. communicatieapparatuur) zijn niet toegestaan gedurende de gehele duur van het tentamen en mogen ook niet in het lokaal meegebracht worden of zijn uitgeschakeld en ingeleverd.
- Schrijf leesbaar ter voorkoming van misverstanden bij de beoordeling van uw werk. Indien uw antwoord niet leesbaar is, wordt uw antwoord fout gerekend.
- Toiletbezoek tijdens het tentamen vindt enkel plaats na toestemming van de examinator.
- Lever bij het verlaten van de zaal, kladpapier, tentamenopgaven en andere tentamen-gerelateerde documenten in bij de examinator.

Cijferberekening / cesuur
- Het eindcijfer voor het vak Statistiek wordt voor 50% bepaald door dit tentamen. - Het tentamen is opgebouwd uit 4 open vragen. Bij iedere (sub)vraag is het aantal te behalen punten tussen haakjes aangegeven. In totaal kunt u 100 punten verdienen. - Het tentamencijfer wordt bepaald door het totaal aantal punten te delen door 10. Het tentamencijfer moet minimaal een 5,0 zijn om de cursus Statistiek met succes af te ronden.

Procedure na het tentamen
- De cijfers van dit tentamenonderdeel worden in principe binnen 10 werkdagen na de afname bekend gemaakt. - Met vragen over de beoordeling kunt u tot 10 werkdagen na bekendmaking van de cijfers terecht bij de cursuscoördinator.

Veel succes!

Formuleblad Statistiek (2024-2025)

Statistiek deel 1

Steekproefgemiddelde (gegeven een steekproef met n uitkomsten $x_1, x_2, \ldots, x_n$)

$$\overline{x} = \frac{\sum_i x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Steekproefvariantie en steekproefstandaardafwijking:

$$s^2 = \frac{\sum_i (x_i - \overline{x})^2}{n-1} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}$$

$$s = \sqrt{s^2} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1}}$$

Rekenregels kansrekening:

$$P(A \text{ of } B) = P(A) + P(B) - P(A \text{ en } B) \qquad \text{(optelregel)}$$

$$P(B) = 1 - P(\text{niet } B) \qquad \text{(complementregel)}$$

$$P(A \mid B) = \frac{P(A \text{ en } B)}{P(B)} \qquad \text{(conditionele kansen)}$$

Discrete en continue kansverdelingen:

	Discrete kansvariabelen	**Continue kansvariabelen**
Uitkomstenruimte:	Eindig / aftelbaar oneindig	Overaftelbaar oneindig
Toepassingen:	Tellen / categoriseren	Meten
Kansbegrip:	Kansfunctie $p(k) = P(X = k)$	Kansdichtheidsfunctie $f(x)$
CDF:	$F(k) = P(X \leq k) = \sum_{\ell:\ell \leq k} p(\ell)$	$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy$
Verwachtingswaarde:	$E[X] = \sum_k k \cdot P(X = k)$	$E[X] = \int x \cdot f(x)\, dx$
Variantie:	$\text{Var}(X) = \sum_k (k - E[X])^2 \cdot P(X = k)$	$\text{Var}(X) = \int (x - E[X])^2 \cdot f(x)\, dx$
Standaardafwijking:	$\sigma(X) = \sqrt{\text{Var}(X)}$	$\sigma(X) = \sqrt{\text{Var}(X)}$

Speciale kansverdelingen:

- $X \sim$ Binomiaal(n, p): tellen van aantal successen bij onafhankelijke kansexperimenten met twee uitkomsten (Bernoulli-experimenten): succes / mislukking.

  Parameters: het aantal Bernoulli-experimenten n en de succeskans per experiment p.

- $X \sim$ Poisson$(\lambda \cdot t)$: tellen van aantal "gebeurtenissen" in een "interval" van tijd / ruimte.

  Parameters: het gemiddelde aantal gebeurtenissen λ per meeteenheid (tijd / ruimte) en het aantal meeteenheden t.

  $\rightarrow$ *Voorbeeld: bij de meeteenheid van een dag bestaat een week uit $t = 7$ meeteenheden.*

- $T \sim$ Exponentieel(λ): meten van de tijd / ruimte tot de volgende gebeurtenis.

  Parameter: het gemiddelde aantal gebeurtenissen λ per meeteenheid (tijd / ruimte).

Verwachtingswaarde en variantie van veelgebruikte kansverdelingen:

Verdeling	Kans(dichtheids)functie	CDF	$E(X)$	Var(X)
Discreet				
Uniform(a,b)	$p(k)=\frac{1}{b-a+1}$ $(k=a,a+1,\dots,b)$	$F(k)=\begin{cases}0 & x<a\\ \frac{k-a+1}{b-a+1} & a\leq k<b\\ 1 & k\geq b\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Binomiaal(n,p)	$p(k)=\binom{n}{k}p^k(1-p)^{n-k}$	$F(k)=\sum\limits_{i=0}^{k}\binom{n}{i}p^i(1-p)^{n-i}$	np	$np(1-p)$
Poisson(λ)	$p(k)=e^{-\lambda}\cdot\frac{\lambda^k}{k!}$	$F(k)=\sum\limits_{i=0}^{k}e^{-\lambda}\cdot\frac{\lambda^i}{i!}$	λ	λ
Continuous				
Uniform(a,b)	$f(x)=\begin{cases}\frac{1}{b-a}, & a\leq x\leq b\\ 0, & \text{elders.}\end{cases}$	$F(x)=\begin{cases}0, & x<a\\ \frac{x-a}{b-a}, & a\leq x<b\\ 1, & x\geq b\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentieel(λ)	$f(x)=\lambda e^{-\lambda x}$, $x\geq 0$	$F(x)=\begin{cases}1-e^{-\lambda x}, & x\geq 0\\ 0, & x<0\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Veelgebruikte functies op de grafische rekenmachine

Type vraag	TI-84 Plus	Casio
Continue kansverdeling (willekeurig)		
$P(a\leq X\leq b)$	$\int_a^b f(x)\,dx$	$\int_a^b f(x)\,dx$
$X\sim$ **Binomiaal**(n,p)		
$P(X=k)$ $P(X\leq k)$	binompdf(n,p,k) binomcdf(n,p,k)	BinomialPD(k,n,p) BinomialCD(k,n,p)
$X\sim N(\mu,\sigma)$		
$P(a\leq X\leq b)$ Grenswaarde g zodat $P(X\leq g)=p$?	normalcdf(a,b,μ,σ) invNorm(p,μ,σ)	NormalCD(a,b,σ,μ) InvNormCD(tail=left, p,σ,μ)
$X\sim$ **Poisson**(λ)		
$P(X=k)$ $P(X\leq k)$	poissonpdf(λ,k) poissoncdf(λ,k)	PoissonPD(k,λ) PoissonCD(k,λ)

z-score:

$$z=\frac{x-\mu}{\sigma}$$

Centrale limietstelling: Gegeven n kansvariabelen $X_1,X_2,\dots,X_n$ die onderling onafhankelijk zijn en dezelfde kansverdeling hebben met een verwachtingswaarde μ en standaardafwijking σ, dan geldt (bij benadering) dat

- de som $\sum X = X_1+X_2+\ldots+X_n$ normaal verdeeld is met verwachtingswaarde $n\cdot\mu$ en standaardafwijking $\sqrt{n}\cdot\sigma$.
- het gemiddelde $\overline{X}=\frac{X_1+X_2+\ldots+X_n}{n}$ normaal verdeeld is met verwachtingswaarde μ en standaardafwijking $\frac{\sigma}{\sqrt{n}}$.

Statistiek deel 2:

Betrouwbaarheidsintervallen voor het gemiddelde μ (σ bekend)

- **$100 \cdot (1-\alpha)$%-betrouwbaarheidsinterval (BI) voor μ:**

$$z_{\alpha/2} = \text{InvNorm}(\text{opp} = 1-\alpha/2; \mu = 0; \sigma = 1)$$

$$[\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}; \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}]$$

- **Minimale steekproefomvang voor $100 \cdot (1-\alpha)$%-BI als μ maximaal $\pm a$ mag afwijken:**

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{a}\right)^2$$

Betrouwbaarheidsintervallen voor het gemiddelde μ (σ onbekend)

- **$100 \cdot (1-\alpha)$%-betrouwbaarheidsinterval (BI) voor μ:**

$$t = \text{InvT}(\text{opp} = 1-\alpha/2; \text{df} = n-1)$$

$$[\overline{x} - t \cdot \frac{s}{\sqrt{n}}; \overline{x} + t \cdot \frac{s}{\sqrt{n}}]$$

- **Minimale steekproefomvang voor $100 \cdot (1-\alpha)$%-BI als μ maximaal $\pm a$ mag afwijken:**

GR tabel (voor verschillende n): $\frac{s}{\sqrt{n}} \cdot \text{InvT}(\text{opp} = 1-\alpha/2; \text{df} = n-1) \leq a$

NB: zodra $n \geq 30$, vallen de normale en de t-verdeling nagenoeg samen. Je mag dan rekenen met de schatting s in plaats van de daadwerkelijke (onbekende) σ.

- **Onderscheidend vermogen (toets met $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, en gegeven $\mu = \mu_1$)**

$$1 - \beta = P(\overline{X} \text{ neemt waarde aan in het kritieke gebied} \mid \mu = \mu_1)$$

Betrouwbaarheidsintervallen voor de binomiale succeskans p

Betrouwbaarheidsinterval voor p (Clopper-Pearson): Gegeven een binomiale verdeling met n Bernoulli-experimenten en onbekende p, en uitkomst k.

1. Bereken de succeskans p_1 zodat geldt $P(X \leq k) = \text{binomcdf}(n; p; k) = \alpha/2$
2. Bereken de succeskans p_2 zodat geldt $P(X \geq k) = 1 - \text{binomcdf}(n; p; k-1) = \alpha/2$
3. De berekende waarden voor p_1 en p_2 zijn de grenzen van het Clopper-Pearson interval.

Hypothesetoetsen

Stappenplan hypothesetoetsen

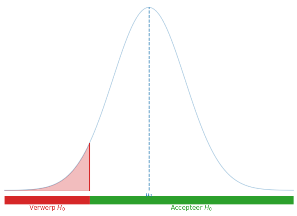
1. Definieer de nulhypothese H_0 en de alternatieve hypothese H_1.
2. Bepaal het significantieniveau α (kans op verwerpen van H_0 terwijl H_0 waar is $\to$ type-I fout)
3. Verzamel data voor de toetsingsgrootheid
4. Bereken de toetsingsgrootheid
   - Uitgaande van de nulhypothese H_0 maken we aannames over de kansverdeling van de toetsingsgrootheid!
5. Geef een conclusie (met behulp van het kritieke gebied / p-waarde) en vertaal deze terug naar de originele probleemcontext.

Drie typen hypothesetoetsen: linkszijdig, tweezijdig, rechtszijdig

Linkszijdige toets

Kritiek gebied:

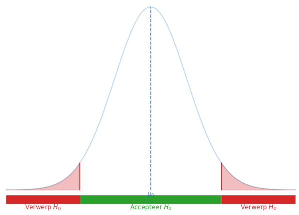
$(-\infty; g]$



Tweezijdige toets

Kritiek gebied:

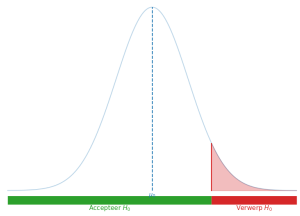
$(-\infty; g_1]$ en $[g_2; \infty)$



Rechtszijdige toets

Kritiek gebied:

$[g; \infty)$



Kansverdeling (onder H_0)	**Linkszijdig**	**Tweezijdig**	**Rechtszijdig**
$N(\mu; \sigma)$	$g = \text{InvNorm}(\alpha; \mu; \sigma)$	$g_1 = \text{InvNorm}(\text{opp} = \frac{\alpha}{2}; \mu; \sigma)$ $g_2 = \text{InvNorm}(\text{opp} = 1 - \frac{\alpha}{2}; \mu; \sigma)$	$g = \text{InvNorm}(1 - \alpha; \mu; \sigma)$
$t(\text{df})$	$g = \text{InvT}(\alpha; \text{df})$	$g_1 = \text{InvT}(\text{opp} = \frac{\alpha}{2}; \text{df})$ $g_2 = \text{InvT}(\text{opp} = 1 - \frac{\alpha}{2}; \text{df})$	$g = \text{InvT}(1 - \alpha; \text{df})$
Grenzen die met de solver functie moeten worden opgelost:			
$\chi^2(\text{df})$ (chikwadraat)	$\chi^2\text{cdf}(0; g; \text{df}) = \alpha$	$\chi^2\text{cdf}(0; g_1; \text{df}) = \frac{\alpha}{2}$ $\chi^2\text{cdf}(g_2; 10^{99}; \text{df}) = \frac{\alpha}{2}$	$\chi^2\text{cdf}(g; 10^{99}; \text{df}) = \alpha$
$F(\text{df}_A; \text{df}_B)$	$\text{Fcdf}(0; g; \text{df}_A; \text{df}_B) = \alpha$	$\text{Fcdf}(0; g_1; \text{df}_A; \text{df}_B) = \frac{\alpha}{2}$ $\text{Fcdf}(g_2; 10^{99}; \text{df}_A; \text{df}_B) = \frac{\alpha}{2}$	$\text{Fcdf}(g; 10^{99}; \text{df}_A; \text{df}_B) = \alpha$

p-waardes uitrekenen (gegeven een theoretische en geobserveerde toetsingsgrootheid T en t)

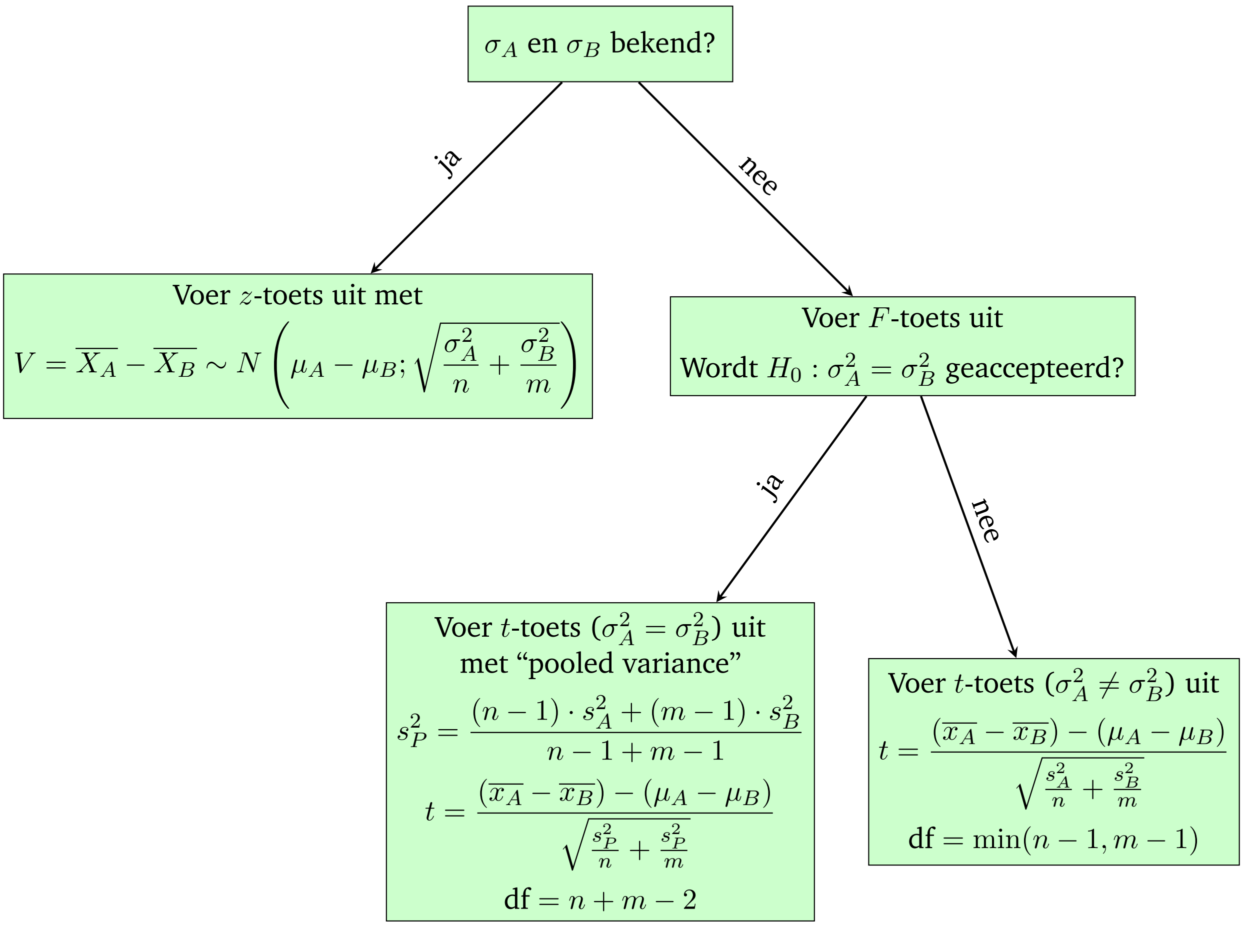
Kansverdeling (onder H_0)	**Linkszijdig ($P(T \leq t)$)**	**Rechtszijdig ($P(T \geq t)$)**
$N(\mu; \sigma)$	$p = \text{normalcdf}(-10^{99}; t; \mu; \sigma)$	$p = \text{normalcdf}(t; 10^{99}; \mu; \sigma)$
$t(\text{df})$	$p = \text{tcdf}(-10^{99}; t; \text{df})$	$p = \text{tcdf}(t; 10^{99}; \text{df})$
$\chi^2(\text{df})$	$p = \chi^2\text{cdf}(0; t; \text{df})$	$p = \chi^2\text{cdf}(t; 10^{99}; \text{df})$
$F(\text{df}_A; \text{df}_B)$	$p = \text{Fcdf}(0; t; \text{df}_A; \text{df}_B)$	$p = \text{Fcdf}(t; 10^{99}; \text{df}_A; \text{df}_B)$

NB: Om met de p-waarde een conclusie te trekken uit een hypothesetoets vergelijken we de p-waarde met het significantieniveau α. Let op: bij tweezijdige toetsen neem je het minimum van de linkszijdige en rechtszijdige p-waarde en vergelijk je deze met $\alpha/2$!

Soorten toetsen

Soort toets	**Toetsingsgrootheid**	**Kansverdeling (onder H_0)**
Toetsen voor het gemiddelde $\mu \leq \mu_0$ of $\mu = \mu_0$ of $\mu \geq \mu_0$		
z-toets (σ bekend)	$\overline{X}$	$N(\mu_0; \frac{\sigma}{\sqrt{n}})$
t-toets (σ onbekend)	$T = \frac{\overline{X}-\mu_0}{\frac{s}{\sqrt{n}}}$	$t(\text{df} = n-1)$
Chikwadraattoetsen (χ^2)		
Onafhankelijkheid	$X^2 = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$	$\chi^2(\text{df} = (\text{\#rijen-1}) \cdot (\text{\#kolommen-1}))$
Aanpassing (goodness-of-fit)	$X^2 = \sum_i \frac{(O_i-E_i)^2}{E_i}$	$\chi^2(\text{df} = (\text{\#categorieën-1}))$
Verschiltoetsen (op basis van twee populaties A en B)		
F-toets: $\sigma_A^2 = \sigma_B^2$	$F = \frac{S_A^2}{S_B^2}$	$F(\text{df}_A, \text{df}_B)$
z-toets	$V = \overline{X_A} - \overline{X_B}$	$N\left(\mu_A - \mu_B; \sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}}\right)$
t-toets ($\sigma_A^2 = \sigma_B^2$)	$T = \frac{(\overline{X_A}-\overline{X_B})-(\mu_A-\mu_B)}{\sqrt{\frac{S_P^2}{n}+\frac{S_P^2}{m}}}$	$t(\text{df} = n+m-2)$
t-toets ($\sigma_A^2 \neq \sigma_B^2$)	$T = \frac{(\overline{X_A}-\overline{X_B})-(\mu_A-\mu_B)}{\sqrt{\frac{S_A^2}{n}+\frac{S_B^2}{m}}}$	$t(\text{df} = \min(n-1; m-1))$

Beslisboom verschiltoetsen



Correlatie en regressie

Correlatiecoëfficiënt van Pearson:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2) \cdot (\overline{y^2} - \overline{y}^2)}}$$

Correlatiecoëfficiënt van Spearman:

$$r_s = 1 - \frac{6 \cdot \sum_i d_i^2}{n^3 - n}$$

Coëfficiënten van de lineaire regressielijn $Y = a + b \cdot X$:

$$b = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}$$

$$a = \overline{y} - b \cdot \overline{x}$$

Schatting van de variantie van de storingsterm ε:

$$s_\varepsilon^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum \left(y_i - (a + b \cdot x_i)\right)^2}{n-2} = \frac{n}{n-2} \cdot \left(\overline{y^2} - a \cdot \overline{y} - b \cdot \overline{xy}\right)$$

$100 \cdot (1 - \alpha)\%$-betrouwbaarheidsinterval voor de gemiddelde Y bij een gegeven $X = x_0$:

$$t = \mathrm{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 2)$$

$$s_\mu = s_\varepsilon \cdot \sqrt{\frac{1}{n} \cdot \left(1 + \frac{(x_0 - \overline{x})^2}{\overline{x^2} - \overline{x}^2}\right)}$$

$$[a + b \cdot x_0 - t \cdot s_\mu; a + b \cdot x_0 + t \cdot s_\mu]$$

$100 \cdot (1 - \alpha)\%$-betrouwbaarheidsinterval voor Y bij een gegeven $X = x_0$:

$$t = \mathrm{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 2)$$

$$s_f = s_\varepsilon \cdot \sqrt{1 + \frac{1}{n} \cdot \left(1 + \frac{(x_0 - \overline{x})^2}{\overline{x^2} - \overline{x}^2}\right)}$$

$$[a + b \cdot x_0 - t \cdot s_f; a + b \cdot x_0 + t \cdot s_f]$$

Opgave 1 (20 punten) Een onderzoeksteam van het Maritime Warfare Center (MWC) onderzoekt de signaalsterkte (in decibel) van sonarpulsen die worden gereflecteerd door een nieuw type onderzeeboot. Het huidige (oude) ontwerp heeft een gemiddelde gereflecteerde signaalsterkte van $\mu = \mu_0 = 65$ decibel.

Het MWC wil toetsen of bij het nieuwe ontwerp de gemiddelde signaalsterkte significant verschilt van het oude ontwerp. In een steekproef van tien onafhankelijke metingen met het nieuwe ontwerp is de gereflecteerde signaalsterkte gemiddeld $63,75$ decibel met een standaardafwijking van $0,78$. Je mag aannemen dat de signaalsterktes normaal verdeeld zijn met onbekende standaardafwijking σ.

Gebruik voor de hypothesetoets een significantieniveau van $\alpha = 0,05$.

1a [5pt] Definieer de nulhypothese en de alternatieve hypothese van de hypothesetoets. Verklaar het gekozen type (tweezijdig, linkszijdig of rechtszijdig) van de toets.

Uitwerking

Aangezien we willen toetsen of de gemiddelde gereflecteerde signaalsterkte significant afwijkt van $\mu = \mu_0 = 65$, kunnen we de hypothesetoets als volgt definiëren:

$$H_0 : \mu = 65 \quad \text{(geen significant verschil)} \qquad \text{(2pt)}$$

$$H_1 : \mu \neq 65 \quad \text{(wel een significant verschil)} \qquad \text{(1pt)}$$

Dit is een tweezijdige toets, omdat de nulhypothese uitgaat van een status-quo (geen verschil) en de alternatieve hypothese juist wel een verschil aanduidt. (2pt)

1b [9pt] Voer de bijbehorende hypothesetoets uit met behulp van het kritieke gebied.

Uitwerking

We mogen uitgaan van een significantieniveau van $\alpha = 0,05$ en verder is gegeven uit de steekproefdata dat $\overline{x} = 63,75$ en $s = 0,78$. Uit de centrale limietstelling volgt onder H_0 dat de toetsingsgrootheid $\overline{X}$ normaal verdeeld is met gemiddelde $\mu = \mu_0 = 65$ en standaardafwijking $\frac{\sigma}{\sqrt{n}}$. (1pt)

Aangezien de steekproefgrootte $n = 10 < 30$ en de (populatie)standaardafwijking σ onbekend is, moeten we de t-verdeling gebruiken. (1pt)

Deze t-verdeling heeft df $= n - 1 = 9$ vrijheidsgraden. (1pt)

De bijbehorende t-waarde (tweezijdig toetsen) is gelijk aan

$$t = \text{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 1) = \text{InvT}(\text{opp} = 0.975; \text{df} = 9) \approx 2,2622. \quad \text{(2pt)}$$

Hieruit volgt een voorspellingsinterval voor $\overline{X}$ op basis van deze steekproef gelijk aan

$$\begin{aligned} &[\mu_0 - t \cdot \frac{s}{\sqrt{n}}; \mu_0 + t \cdot \frac{s}{\sqrt{n}}] \\ =&[65 - 2,2622 \cdot \frac{0,78}{\sqrt{10}}; 65 + 2,2622 \cdot \frac{0,78}{\sqrt{10}}] \\ \approx&[64,4420; 65,5580] \end{aligned} \quad \text{(2pt)}$$

Het gemeten steekproefgemiddelde $\overline{x} = 63,75$ ligt niet in dit voorspellingsinterval – dus wel in het kritieke gebied – dus verwerpen we de nulhypothese. (1pt) Er is voldoende reden om aan te nemen dat de gemiddelde gereflecteerde signaalsterkte significant afwijkt van het nieuwe ontwerp. (1pt)

1c [6pt] Bereken de kans op een type-II fout β als geldt dat het daadwerkelijke populatiegemiddelde van het nieuwe ontwerp gelijk is aan $\mu = 64,5$ decibel, en dat de standaardafwijking σ gelijk is aan $0,8$.

Uitwerking

Bij een type-II fout wordt H_0 aangenomen, terwijl H_1 waar is. (1pt) In dit geval geldt dat H_1 waar is, omdat $\mu = 64,5 \neq 65$. (1pt) We willen dus de kans bepalen dat de toetsingsgrootheid een waarde in het acceptatiegebied van H_0 aanneemt (oftewel $[64,4420; 65,5580]$), gegeven $\mu = 64,5$. (1pt)

In andere woorden:

$$\begin{aligned} \beta &= P(64,4420 \leq \overline{X} \leq 65,5580 \mid \mu = 64.5) \\ &= \text{normalcdf}(\text{lower} = 64,4420; \text{upper} = 65,5580; \mu = 64.5; \sigma = \frac{0.8}{\sqrt{10}}) \\ &\approx 0,5907. \end{aligned} \quad \text{(2pt)}$$

De kans op een type-II fout is dus gelijk aan $\beta \approx 0,5907$, oftewel 59% kans dat de nulhypothese wordt aangenomen terwijl deze in werkelijkheid fout is. (1pt)

Opgave 2 (30 punten) De afdeling Dienstencentrum Human Resources (DCHR) van Defensie wil nagaan of het wervingsproces effectief en vlot verloopt. De doelstelling is dat sollicitanten binnen **45 dagen** (exclusief veiligheidsonderzoek) na hun schriftelijke sollicitatie een formeel aanstellingsvoorstel ontvangen.

Om de huidige omstandigheden van werving en selectie te onderzoeken, zijn doorlooptijden van dertien recent aangestelde burgermedewerkers verzameld (in dagen):

$$39, 42, 44, 48, 47, 43, 31, 46, 40, 42, 43, 49, 44$$

Neem aan dat de doorlooptijden normaal verdeeld zijn, elke procedure met dezelfde verwachtingswaarde en standaardafwijking.

2a [7pt] Bereken het steekproefgemiddelde en de steekproefstandaardafwijking van de doorlooptijden van sollicitatieprocedures.

Uitwerking

We berekenen het steekproefgemiddelde $\overline{x}$ als volgt:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{39 + 42 + \ldots + 44}{13} \approx 42,9231. \qquad \text{(3pt)}$$

We berekenen de steekproefvariantie s^2 als volgt:

$$\begin{aligned} s^2 &= \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} \\ &= \frac{(39 - 42,9231)^2 + (42 - 42,9231)^2 + \ldots + (44 - 42,9231)^2}{13 - 1} \\ &\approx 21,5769. \end{aligned} \qquad \text{(3pt)}$$

We berekenen de steekproefstandaardafwijking s door te wortel van de steekproefvariantie s^2 te nemen:

$$s = \sqrt{s^2} = \sqrt{21,5769} \approx 4,6451. \qquad \text{(1pt)}$$

2b [6pt] Bereken een $94\,\%$-betrouwbaarheidsinterval voor de gemiddelde doorlooptijd μ van sollicitatieprocessen, op grond van bovengenoemde steekproef, zonder daarbij gebruik te maken van de optie TESTS/Interval van de grafische rekenmachine. Rond de grenzen van het interval af op gehele dagen, dusdanig dat het betrouwbaarheidsniveau gewaarborgd blijft.

Uitwerking

Laat $X \sim N(\mu =?; \sigma =?)$ de doorlooptijd zijn van een willekeurige sollicitatieprocedure. Volgens de centrale limietstelling is de gemiddelde doorlooptijd $\overline{X} = \frac{X_1+X_2+\ldots+X_{13}}{13}$ van 13 procedures ook normaal verdeeld, met onbekende verwachtingswaarde μ en standaardafwijking $\frac{\sigma}{\sqrt{13}}$. Omdat naast de verwachtingswaarde μ ook de standaardafwijking σ onbekend is (en bovendien de steekproefgrootte $n = 13 < 30$), moet de t-verdeling worden gebruikt. Omdat σ onbekend is en de steekproefgrootte $n = 13 < 30$ is, moeten we gebruik maken van de t-verdeling. De t-waarde die hoort bij $94\,\%$ betrouwbaarheid, oftewel $\alpha = 0.06$, is (in het geval van tweezijdige intervallen) gelijk aan (1pt) (1pt)

$$t = \mathrm{InvT}(\text{opp} = 1 - \frac{\alpha}{2}; \text{df} = n - 1) = \mathrm{InvT}(\text{opp} = 0,97; \text{df} = 12) \approx 2,0764. \qquad \text{(1pt)}$$

Het $94\,\%$-betrouwbaarheidsinterval wordt dan gegeven door

$$\begin{aligned}
&[\overline{x} - t \cdot \frac{s}{\sqrt{n}}; \overline{x} + t \cdot \frac{s}{\sqrt{n}}] \\
&= [42,9231 - 2,0764 \cdot \frac{4,6451}{\sqrt{13}}; 42,9231 + 2,0764 \cdot \frac{4,6451}{\sqrt{13}}] \\
&= [40,2480; 45,5982]. \qquad \text{(2pt)}
\end{aligned}$$

Om de betrouwbaarheid te waarborgen, mag het interval niet kleiner worden, dus naar buiten afronden: $[40; 46]$. Met $94\,\%$ betrouwbaarheid ligt de gemiddelde doorlooptijd van een sollicitatieprocedure tussen de 40 en 46 dagen. (1pt)

2c [9pt] Toets op basis van de gegeven steekproef of de gemiddelde doorlooptijd van sollicitatieprocedures voldoet aan de bovengenoemde doelstelling van maximaal 45 dagen. Geef je conclusie op basis van de p-waarde. Kies in dit geval als onbetrouwbaarheid $\alpha = 0,05$. Leg in simpele bewoordingen uit wat de uitslag van deze toets betekent voor de doorlooptijd van sollicitatieprocedures binnen Defensie.

Uitwerking

In deze hypothesetoets hebben we we te maken met een t-toets met nul- en alternatieve hypothese als volgt: (1pt)

$$H_0 : \mu \leq 45 \qquad \text{(gemiddelde voldoet aan de doelstelling)}$$

$$H_1 : \mu > 45 \qquad \text{(gemiddelde voldoet NIET aan de doelstelling)} \qquad (1pt)$$

In dit geval werken we met een rechtszijdige toets (alternatieve hypothese heeft vorm $>$), dus het kritieke gebied is van de vorm $[g, \infty)$. We willen de (kleinste) grens g zoeken zodanig dat de kans op een type-I fout (H_0 verwerpen terwijl die waar is) kleiner is dan $\alpha = 0,05$. (1pt)
Onder de nulhypothese is de gemiddelde doorlooptijd $\overline{X}$ van 13 willekeurige sollicitatieprocedures normaal verdeeld met verwachtingswaarde $\mu = 45$ en standaardafwijking $\frac{\sigma}{\sqrt{n}}$. Omdat σ onbekend is en $n = 18 < 30$, gebruiken we opnieuw (1pt)
de t-verdeling en $s = 4,6451$. Er geldt dus, omdat we eenzijdig toetsen, dat (1pt)

$$t = \text{InvT}(opp = 1 - \alpha; df = n - 1) = \text{InvT}(opp = 0,95; df = 12) \approx 2,1788. \qquad (1pt)$$

(merk op dat de t-waarde anders is, omdat we met een ander betrouwbaarheidsniveau werken!).
De bijbehorende grenswaarde vinden we nu met

$$g = \mu + t \cdot \frac{s}{\sqrt{18}} = 45 + 2,1788 \cdot \frac{4,6451}{\sqrt{13}} \approx 47,8070. \qquad (1pt)$$

Het steekproefgemiddelde $\overline{x} \approx 45,9231$ is kleiner dan deze ondergrens $g \approx 47,8070$ van het kritieke gebied, dus H_0 wordt niet verworpen. Dit betekent dat met deze toets niet kan worden aangetoond dat de gemiddelde doorlooptijd buiten de gestelde doelstelling valt. (2pt)

2d [8pt] Het Dienstencentrum Human Resources breidt haar onderzoek uit. In een grootschalige onderzoek met 386 burgermedewerkers, blijkt dat 102 burgermedewerkers binnen 45 dagen een aanstellingsvoorstel heeft ontvangen. Bereken met de Clopper-Pearson methode een $95\,\%$-betrouwbaarheidsinterval voor het percentage burgermedewerkers dat binnen 45 dagen een aanstellingsvoorstel heeft ontvangen.

Uitwerking

Laat X het aantal burgermedewerkers zijn dat binnen 45 dagen een aanstellingsvoorstel heeft ontvangen. Aangezien de steekproefomvang gelijk is aan 386 kiesgerechtigde Nederlanders, is X binomiaal verdeeld met $n = 386$ en nog onbekende succeskans p. (1pt)

We bepalen een $95\,\%$-betrouwbaarheidsinterval voor p met behulp van de Clopper-Pearson methode. Omdat de gewenste betrouwbaarheid $95\,\%$ is, geldt dat $\alpha = 0,05$.

De Clopper-Pearson methode werkt als volgt:

1. Bepaal de succeskans p_1 waarvoor geldt dat de linkeroverschrijdingskans van de uitkomst $k = 102$ gelijk is aan $\alpha/2$, oftewel $P(X \leq 102) = \alpha/2 = 0,025$. Voer hiervoor in het functiescherm van de grafische rekenmachine:

$$y_1 = \text{binomcdf}(n = 386; p_1 = X; k = 102)$$
$$y_2 = 0,025$$ (2pt)

   De solver optie geeft een waarde van $p_1 \approx 0,3112$. (1pt)

2. Bepaal de succeskans p_2 waarvoor geldt dat de rechteroverschrijdingskans van de uitkomst $k = 102$ gelijk is aan $\alpha/2$, oftewel $P(X \geq 102) = 1 - P(X \leq 101) = \alpha/2 = 0,025$. Voer hiervoor in het functiescherm van de grafische rekenmachine:

$$y_1 = 1 - \text{binomcdf}(n = 386; p_1 = X; k = 101)$$
$$y_2 = 0,025$$ (2pt)

   De solver optie geeft een waarde van $p_2 \approx 0,2209$. (1pt)

We vinden het Clopper-Pearson interval door de twee gevonden waarden als grenzen te nemen, oftewel met $95\,\%$-betrouwbaarheid ligt het percentage burgermedewerkers dat binnen 45 dagen een aanstellingsvoorstel heeft ontvangen tussen ongeveer $22\,\%$ en $31\,\%$. (1pt)

Opgave 3 (20 punten) Bij de sollicitatiegesprekken voor cadetten en adelborsten wordt door de aanstellingscommissie aan de aspirant-officier gevraagd naar zijn of haar primaire motivatie om officier te willen worden. De meest genoemde antwoorden van de aspiranten zijn ruwweg in te delen in drie categorieën: leiderschap, avontuur en dienen aan het vaderland.

Naar aanleiding van de sollicitatiegesprekken kunnen de antwoorden als volgt worden samengevat:

	Cadetten	Adelborsten	Totaal
Leiderschap	52	22	74
Avontuur	23	9	32
Vaderland dienen	38	28	66
Totaal	113	59	172

C-NLDA wil inzicht krijgen in of de primaire motivatie van cadetten significant afwijkt van die van adelborsten.

3a [4pt] Welk type hypothesetoets dienen we hiervoor uit te voeren? Formuleer de nulhypothese H_0 en de alternatieve hypothese H_1 van de bijbehorende hypothesetoets.

Uitwerking

Omdat we willen weten of de primaire motivatie van cadetten significant afwijkt van die van de adelborsten, dienen we een χ^2-toets (chikwadraat) voor onafhankelijkheid uit te voeren. Dit kan ook omdat we te maken hebben met twee categorische variabelen! (2pt)

De bijbehorende nul- en alternatieve hypothese zijn in dit geval gelijk aan

H_0 : de verdeling van de primaire motivaties van cadetten
is onafhankelijk van het type aspirant-officieren. (1pt)

H_1 : de verdeling van de primaire motivaties van cadetten
is WEL afhankelijk van het type aspirant-officieren. (1pt)

3b [11pt] Bepaal de p-waarde van de desbetreffende hypothesetoets.

Uitwerking

Om een chi-kwadraattoets voor onafhankelijkheid uit te voeren, moeten we eerst de verwachte frequenties uitrekenen. Onder de aanname dat H_0 waar is, is deze voor elk van de negen cellen te berekenen als $\frac{\text{rijtotaal}\cdot\text{kolomtotaal}}{\text{totaal}}$. De verwachte frequenties zijn te berekenen met behulp van de formule:

$$E_{ij} = \frac{\text{rijtotaal}_i \cdot \text{kolomtotaal}_j}{\text{totaal}}$$

Tabel 1: "Observed" frequenties

	C	A	Totaal
Leiderschap	52	22	74
Avontuur	23	9	32
Vaderland dienen	38	28	66
Totaal	113	59	172

Tabel 2: "Expected" frequenties

C	A	Totaal
$48,6163$	$25,3837$	74
$21,0233$	$10,9767$	32
$43,3605$	$22,6395$	66
113	59	172

(4pt)

We berekenen de toetsingsgrootheid χ^2 als volgt, waarbij O_{ij}/E_{ij} de geobserveerde / verwachte frequentie is in rij i, kolom j:

$$\begin{aligned}\chi^2 &= \frac{(O_{1,1}-E_{1,1})^2}{E_{1,1}} + \frac{(O_{1,2}-E_{1,2})^2}{E_{1,2}} + \ldots + \frac{(O_{3,2}-E_{3,2})^2}{E_{3,2}} \\ &= \frac{(52-48,6163)^2}{48,6163} + \frac{(22-25,3837)^2}{25,3837} + \ldots + \frac{(28-22,6395)^2}{22,6395} \\ &\approx 3,1603.\end{aligned}$$ (2pt)

De bijbehorende p-waarde is de rechteroverschrijdingskans $P(X > \chi^2)$, waarbij X^2 de toetsingsgrootheid van de chikwadraattoets voor onafhankelijkheid. De (1pt) toetsingsgrootheid X^2 volgt onder de nulhypothese een χ^2-verdeling met df $=$ $(\#\text{rijen} - 1)\cdot(\#\text{kolommen} - 1) = (3-1)\cdot(2-1) = 2$ vrijheidsgraden. (1pt)

$$\begin{aligned}p &= P(\chi^2 > 3.1603) \\ &= \chi^2\text{cdf}(\text{lower} = 3.1603; \text{upper} = 10^{99}; \text{df} = 2) \\ &\approx 0.2059.\end{aligned}$$ (2pt)

Omdat $p > \alpha$, wordt de nulhypothese H_0 aangenomen. Er is onvoldoende reden om de aanname te verwerpen dat de twee variabelen onafhankelijk zijn van elkaar. (1pt)

3c [5pt] Geef de conclusie van de hypothesetoets (op basis van een significantieniveau $\alpha = 0,05$) met behulp van de berekende p-waarde.

Uitwerking

Uit de hypothesetoets volgt een redelijk grote rechteroverschrijdingskans $p \approx 0,2059$. Omdat $p > \alpha = 0,05$, geldt dat H_0 wordt aangenomen. (2pt) We kunnen dus de conclusie trekken dat de verdeling van primaire motivaties om officier te worden niet significant verschilt tussen cadetten en adelborsten. (1pt)

Als we kijken naar de tabel van geobserveerde en verwachte frequenties, dan zien we dat deze redelijk dichtbij elkaar in de buurt liggen. Er zijn geen grote uitschieters te bekennen in de verwachte frequenties ten opzichte van de daadwerkelijk geobserveerde frequenties. (2pt)

Opgave 4 (30 punten) In een gezamenlijk project van sportinstructeurs van de KMA en de geneeskundige dienst wordt het verband onderzocht tussen de gemiddelde slaaptijd van een cadet (in uren) en de hersteltijd na een zware inspanning (in uren). Voor de hersteltijd wordt gekeken naar een spierpijn-index, en een cadet is volledig hersteld wanneer de spierpijn-index onder een gegeven drempelwaarde komt.

Van negen cadetten wordt gemeten hoe lang ze gemiddeld hebben geslapen en hoe lang hun hersteltijd is na een zware inspanning.

Slaaptijd (in uren)	$4,0$	$4,5$	$5,0$	$5,5$	$6,0$	$6,5$	$7,0$	$7,5$	$8,0$
Hersteltijd (in uren)	66	61	63	62	65	64	57	59	60

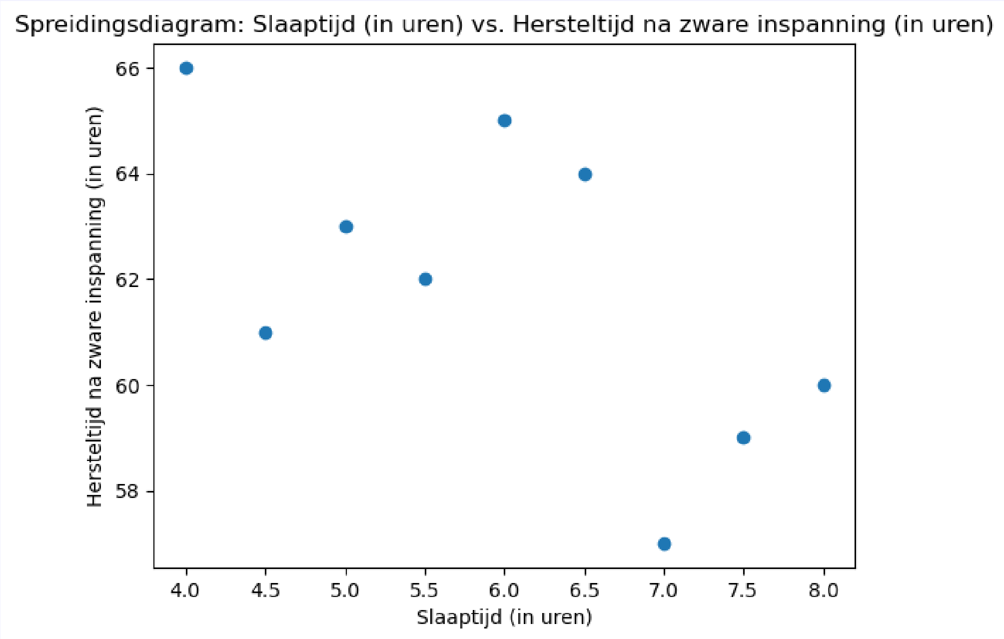
4a [2pt] Als we een regressie-analyse willen uitvoeren, welke variabele zou dan de afhankelijke variabele Y zijn en welke variabele zou de onafhankelijke variabele X zijn?

Uitwerking

Bij een regressie-analyse is de meest logische keuze om slaaptijd als onafhankelijke variabele X te kiezen, en de hersteltijd als afhankelijke variabele Y. Een langere hersteltijd verklaart niet waarom iemand langer slaapt, terwijl dat andersom wellicht wel het geval kan zijn. (2pt)

4b [5pt] Teken het bijbehorende spreidingsdiagram op basis van je antwoord op vraag (a).

Uitwerking



4c [8pt] Bereken Pearson's correlatiecoëfficiënt $r(x, y)$. Wat kun je concluderen over de samenhang van de twee variabelen?

Uitwerking

We beginnen met het uitrekenen van Pearson's correlatiecoëfficiënt

$$r(x,y) = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2) \cdot (\overline{y^2} - \overline{y}^2)}}$$

Hiervoor hebben we dus de waardes van $\overline{x}$, $\overline{y}$, $\overline{xy}$, $\overline{x^2}$ en $\overline{y^2}$ nodig. Deze bepalen we aan de hand van de volgende rekentabel:

x	y	xy	x^2	y^2
4	66	264	16	4356
$4,5$	61	$274,5$	$20,25$	3721
5	63	315	25	3969
$5,5$	62	341	$30,25$	3844
6	65	390	36	4225
$6,5$	64	416	$42,25$	4096
7	57	399	49	3249
$7,5$	59	$442,5$	$56,25$	3481
8	60	480	64	3600
$\overline{x} = 6$	$\overline{y} = 61,8889$	$\overline{xy} = 369,1111$	$\overline{x^2} = 37,6667$	$\overline{y^2} = 3837,8889$

(4pt)

De correlatiecoëfficiënt van Pearson is dus gelijk aan

$$
\begin{aligned}
r(x,y) &= \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x}^2 - \overline{x^2}) \cdot (\overline{y}^2 - \overline{y^2})}} \\
&= \frac{369,1111 - 6 \cdot 61,8889}{\sqrt{(6^2 - 37,6667) \cdot (61,8889^2 - 3837,8889)}} \\
&= \frac{-2,2222}{3,5717} \\
&\approx -0,6222.
\end{aligned}
$$

(3pt)

De correlatiecoëfficiënt ligt redelijk ver van 0 en is negatief. Dit houdt in dat er een vrij duidelijke dalende trend zichtbaar is, maar dat er toch onzekerheid zit in hoe het exacte lineaire verband eruit zal zien. (1pt)

4d [7pt] Bereken de regressielijn $Y = a + b \cdot X$ door berekening van de coëfficiënten a en b. Bepaal aan de hand van de regressielijn een statistisch verantwoorde voorspelling voor de hersteltijd van een cadet die 6 uur en 45 minuten heeft geslapen.

Uitwerking

Om de coëfficiënten a en b van de regressielijn te berekenen, gaan we de rekentabel van vraag (a) hergebruiken. Er volgt namelijk dat

$$\begin{aligned} b &= \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \\ &= \frac{369,1111 - 6 \cdot 61,8889}{37,6667 - (6)^2} \\ &= \frac{-2,2222}{1,6667} \approx -1,3333 \\ a &= \overline{y} - b \cdot \overline{x} \\ &= 61,8889 - (-1,3333 \cdot 6) \\ &\approx 69,8889. \end{aligned}$$

(2pt)

(2pt)

De formule van de regressielijn is dus gelijk aan $Y = 69,8889 - 1,3333 \cdot X$. Een statistisch verantwoorde voorspelling voor de hersteltijd van een cadet die gemiddelde 6 uur en 45 minuten heeft geslapen vinden we door $X = 6,75$ in te vullen: Dit geeft een waarde van $Y = 69,8889 - 1,3333 \cdot 6,75 = 60,88887$ uur, oftewel iets minder dan 61 uur. (1pt) (2pt)

4e [8pt] Bereken een 90%-voorspellingsinterval voor de hersteltijd van een willekeurige cadet die 6 uur en 45 minuten heeft geslapen.

Uitwerking

DIn opdracht (d) hebben we een puntschatting van $y_0 = 69,8889 - 1,3333 \cdot 6,75 \approx 60,8889$ bepaald. Daarnaast kunnen we de standaardafwijking σ van de storingsterm ε schatten:

$$\begin{aligned} s_\varepsilon &= \sqrt{\frac{n}{n-2} \cdot \left(\overline{y^2} - a \cdot \overline{y} - b \cdot \overline{xy}\right)} \\ &= \sqrt{\frac{9}{7} \cdot (3837,8889 - 69,8889 \cdot 61,8889 - (-1,3333 \cdot 369,1111))} \\ &\approx 2,456. \end{aligned}$$

(3pt)

Vervolgens kunnen we een puntschatting berekenen van de standaardafwijking van Y voor gegeven $X = x_0$:

$$\begin{aligned} s_f &= s_\varepsilon \cdot \sqrt{1 + \frac{1}{n} \cdot \left(1 + \frac{(x_0 - \overline{x})^2}{\overline{x^2} - \overline{x}^2}\right)} \\ &= 2,456 \cdot \sqrt{1 + \frac{1}{9} \cdot \left(1 + \frac{(6,75 - 6)^2}{37,6667 - 6^2}\right)} \\ &\approx 2,6321. \end{aligned} \qquad \text{(2pt)}$$

Omdat we de standaardafwijkingen geschat hebben en de storingstermen normaal verdeeld zijn, moeten we werken met de t-verdeling met $df = n - 2 = 7$ vrijheidsgraden. De t-waarde die hoort bij een betrouwbaarheidsniveau $\alpha = 0,1$ is gelijk aan

$$t = \mathrm{InvT}(\text{opp} = 1 - \alpha/2; \text{df} = n - 2) = \mathrm{InvT}(\text{opp} = 0,95; \text{df} = 7) \approx 1,8946. \qquad \text{(1pt)}$$

Het $90\,\%$-betrouwbaarheidsinterval voor de gemiddelde Y voor gegeven $X = x_0$ kan dus worden beschreven door

$$\begin{aligned} &[y_0 - t \cdot s_f; y_0 - t \cdot s_f] \\ &= [60,8889 - 1,8946 \cdot 2,6321; 60,8889 + 1,8946 \cdot 2,6321] \\ &\approx [55,9021; 65,8757]. \end{aligned} \qquad \text{(2pt)}$$

Met $90\,\%$ zekerheid zal de hersteltijd van een willekeurige cadet die 6 uur en 45 minuten heeft geslapen tussen ongeveer 56 en 66 uur liggen.