

Naam:

1

Voornaam:

Studentennummer:

Opleiding:

G2N11c: Examen Statistiek en Data-analyse

Prof. dr. Mia Hubert

28 januari 2022

Instructies voor een goed verloop van het examen:

1. Wie de vragen aanneemt en bekijkt, moet minstens 1 uur blijven zitten.
2. Dit is een **gesloten boek** examen. Enkel schrijfgerief, een niet-grafisch rekenmachine, het formularium en de statistische tabellen van de cursus zijn toegestaan. **Op je formularium en statistische tabellen mag niets geschreven staan!**
3. Dit examen bevat 4 vragen, opgedeeld in 4 bundels in 4 verschillende kleuren. Controleer of je elke bundel gekregen hebt, en ook kladpapier.
4. Schrijf op elk van de antwoordbladen je naam en voornaam. Op de eerste pagina van elke bundel schrijf je ook je studentennummer en opleiding.
5. Gebruik in principe voor elk antwoord niet meer ruimte dan voorzien. Indien je toch meer antwoordruimte nodig hebt, dan mag je de achterkant of kladpapier gebruiken. Verwijs hier duidelijk naar en laat je kladpapier bij het indienen vastnieten aan de juiste bundel.
6. Schrijf zo leesbaar en duidelijk mogelijk. Wat niet leesbaar is, wordt fout gerekend.
7. Noteer je bewerkingen zorgvuldig. Een juist antwoord zonder tussenstappen levert zeer weinig punten op.
8. Einduitkomsten rond je af tot **3 cijfers** na de komma. Tussenresultaten worden niet afgerond.
9. Alle informatie uit gegeven R-output mag je rechtstreeks gebruiken en hoeft je dus niet na te rekenen.
10. Alle elektronische apparaten worden voor het examen uitgeschakeld.

Veel succes!

1. Duid per meerkeuzevraag aan welke van de beweringen juist zijn. Meerdere juiste antwoorden zijn mogelijk. Verklaar je antwoord in de voorziene ruimte. Een antwoord zonder verklaring levert geen punten op.

(a) De toevalsvariabelen (X, Y) zijn bivariaat normaal verdeeld met $\mu = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$ en $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

Hieruit kunnen we besluiten dat:

- ☐ X en Y zijn onafhankelijk
- ☐ $\rho(X, Y) = 0$
- ☐ $\text{Cov}(X, Y) = 0$
- ☒ $X + Y$ is normaal verdeeld
- ☐ $\text{Var}(X + Y) = 3$
- ☐ XY is normaal verdeeld
- ☒ $E(XY) = 16$
- ☐ geen enkele van de antwoorden is juist

Oplossing:

Er is gegeven dat $E(X) = \mu_x = 5$ en $E(Y) = \mu_y = 3$. Verder is $\text{Var}(X) = 2$ (waarde van Σ op de eerste rij, eerste kolom), $\text{Var}(Y) = 1$ (waarde van Σ op de tweede rij, tweede kolom), en $\text{Cov}(X, Y) = \sigma_{xy} = 1$ (waarde van Σ buiten de diagonaal, dus op de eerste rij, tweede kolom. Of op de tweede rij, eerste kolom).

- Hieruit volgt dat $\rho(X, Y) \neq 0$ want de correlatie is een veelvoud van de covariantie. Omdat (X, Y) bivariaat normaal verdeeld zijn, volgt hieruit ook dat X en Y niet onafhankelijk zijn.
- Als $(X, Y) \sim N_2(\mu, \Sigma)$ dan is elke lineaire combinatie normaal verdeeld. Bovendien is steeds $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y) = 2 + 1 + 2 \cdot 1 = 5$.
- Over de verdeling van het product XY kan geen uitspraak gedaan worden. Wel geldt steeds dat $E(XY) = \text{Cov}(X, Y) + E(X)E(Y) = 1 + 5 \cdot 3 = 16$.

- (b) We beschikken over een steekproef van 21 metingen van een normaal verdeelde variabele X . We berekenen een 99% betrouwbaarheidsinterval voor het gemiddelde van X . Dit levert het interval $[10.76; 13.24]$. Hieruit kunnen we besluiten dat:

- ☐ het populatiegemiddelde van X gelijk is aan 12.
- ☐ het populatiegemiddelde van X ligt tussen 10.76 en 13.24.
- ☐ wanneer we 1000 keer een nieuwe steekproef met 21 metingen beschouwen uit dezelfde populatie, zowat 990 keer het populatiegemiddelde van X ligt tussen 10.76 en 13.24.
- ☐ het populatiegemiddelde van X kleiner is dan 10.
- ☐ het populatiegemiddelde van X kleiner is dan 10. De kans dat we dit besluit nemen terwijl in werkelijkheid het populatiegemiddelde van X groter is dan 10 bedraagt 1%.
- ☐ het populatiegemiddelde van X niet gelijk is aan 10.
- ☒ **het populatiegemiddelde van X niet gelijk is aan 10. De kans dat we dit besluit nemen terwijl in werkelijkheid het populatiegemiddelde van X gelijk is aan 10 bedraagt 1%.**
- ☒ **het steekproefgemiddelde gelijk is aan 12.**
- ☒ **het steekproefgemiddelde ligt tussen 10.76 en 13.24.**
- ☐ wanneer we 1000 keer een nieuwe steekproef met 21 metingen beschouwen uit dezelfde populatie, zowat 990 keer het steekproefgemiddelde ligt tussen 10.76 en 13.24.
- ☐ het steekproefgemiddelde kleiner is dan 10.
- ☐ het steekproefgemiddelde kleiner is dan 10. De kans dat we dit besluit nemen terwijl in werkelijkheid het steekproefgemiddelde groter is dan 10 bedraagt 1%.
- ☐ het steekproefgemiddelde kleiner is dan 10. De kans dat we dit besluit nemen terwijl in werkelijkheid het populatiegemiddelde groter is dan 10 bedraagt 1%.
- ☒ **het steekproefgemiddelde niet gelijk is aan 10.**
- ☐ het steekproefgemiddelde niet gelijk is aan 10. De kans dat we dit besluit nemen terwijl in werkelijkheid het steekproefgemiddelde gelijk is aan 10 bedraagt 1%.
- ☐ het steekproefgemiddelde niet gelijk is aan 10. De kans dat we dit besluit nemen terwijl in werkelijkheid het populatiegemiddelde gelijk is aan 10 bedraagt 1%.
- ☐ geen enkele van de antwoorden is juist.

Oplossing:

Elke waarde μ_0 die niet tot het 99% BI behoort, komt overeen met het verwerpen van de nulhypothese $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ op het 1% significantieniveau.

Hier beschouwen we $\mu_0 = 10$. We aanvaarden dus H_1 , maar we zijn niet 100% zeker van ons besluit. We kunnen immers een type I fout maken. Dit wil zeggen dat de kans dat we H_0 verwerpen terwijl die in feite waar is, gelijk is aan 1%.

Alle voorgaande uitspraken in de lijst zijn onjuist. De derde uitspraak lijkt heel sterk op de definitie van een BI, maar bij elke nieuwe steekproef wordt een nieuw BI geconstrueerd dus we kunnen niets beweren over hoe vaak het populatiegemiddelde tot dat ene BI behoort.

Het BI is symmetrisch rond het steekproefgemiddelde, dus $\bar{x} = (10.76 + 13.24)/2$. We kunnen dus ook zeggen dat het steekproefgemiddelde tot het gegeven BI behoort en niet gelijk is aan 10.

Naam:

1

Voornaam:

Studentennummer:

Opleiding:

2. Het is algemeen geweten dat een deel van de Leuvense fietsers rondrijdt met een niet-werkend achterlicht. Uit een uitgebreid politie-onderzoek naar veiligheid bij fietsers blijkt dat 80% van de fietsers met een niet-werkend achterlicht geen fietshelm draagt. Van de fietsers met een werkend achterlicht draagt 15% wel een fietshelm. Bovendien rijdt 95% van de fietsers die een helm dragen rond met een werkend achterlicht.

- (a) Toon aan dat 96.2% van de Leuvense fietser rondrijdt met een werkend achterlicht.

Oplossing: Noem A de gebeurtenis dat een Leuvense fietser rondrijdt met een werkend achterlicht en B de gebeurtenis dat een fietser een fietshelm draagt. Dan is gegeven dat

$$P(B|A) = 15\%$$

$$P(B^c|A^c) = 80\%$$

$$P(A|B) = 95\%$$

We gebruiken de regel van Bayes om $P(A)$ te berekenen:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

$$\Downarrow$$

$$0.95 = \frac{0.15P(A)}{0.15P(A) + 0.2(1 - P(A))}$$

$$\Downarrow$$

$$0.95(0.2 - 0.05P(A)) = 0.15P(A)$$

$$\Downarrow$$

$$0.19 = (0.15 + 0.0475)P(A)$$

$$\Downarrow$$

$$P(A) = \frac{0.19}{0.1975} = 0.962025$$

-
- (b) Hoe groot is de kans dat een Leuvense fietser een fietshelm draagt?

Oplossing: We gebruiken de wet van de totale kans om $P(B)$ te berekenen:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = 0.15 \times 0.962 + 0.2 \times 0.038 = 0.1519$$

- (c) Is het dragen van een fietshelm onafhankelijk van het rondrijden met een werkend achterlicht? Verklaar je antwoord.

Oplossing: Neen, want $P(A) \neq P(A|B)$

Naam:

Voornaam:

Studentennummer:

Opleiding:

3. Een luchtvaartmaatschappij beweert in haar jaarlijks rapport dat 83% van haar passagiers hun vrienden zou aanraden om ook met deze maatschappij te reizen. Een kritische journalist denkt dat de luchtvaartmaatschappij een te optimistisch cijfer publiceert en bevraagt daarom 200 willekeurig uitgekozen passagiers. Hij stelt vast dat 160 ervan hun vrienden aanbevelen om met die maatschappij te vliegen.

Kan de journalist besluiten dat zijn wantrouwen tegenover de luchtvaartmaatschappij terecht is?

- (a) Welke hypothesen worden hier getest? Formuleer zoveel mogelijk in symbolen en verklaar de gebruikte notatie indien van toepassing.

Oplossing:

$$H_0 : p \geq 0.83 \text{ versus } H_1 : p < 0.83$$

met p de proportie passagiers die hun vrienden aanbevelen om met die maatschappij te vliegen

- (b) Welke voorwaarde(n) moet(en) voldaan zijn om deze hypothesetest te mogen uitvoeren? Ga, indien mogelijk, na of hieraan voldaan is en leg uit waarom (niet).

Oplossing: Er is voldaan aan de voorwaarden:

- De steekproef is willekeurig: OK, dit is gegeven.
- Discrete variabele met 2 uitkomsten: OK, vrienden aanbevelen of niet
- De absolute frequenties zijn ≥ 5 : OK, want $0.83 \cdot 200 = 166$ en $0.17 \cdot 200 = 34$.

- (c) Geef de formule van de teststatistiek en de (eventueel benaderende) verdeling ervan onder de nulhypothese. Geef ook het aantal vrijheidsgraden indien van toepassing.

Oplossing:

$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1 - p_0)/n}} \approx N(0, 1)$$

- (d) Bereken de testwaarde.

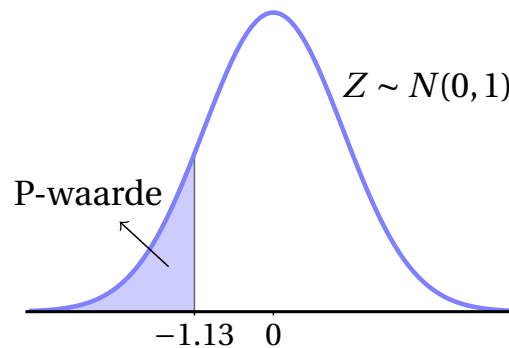
Oplossing: $z = \frac{0.8 - 0.83}{\sqrt{0.83 \cdot 0.17/200}} = -1.129$

- (e) Bereken de P-waarde. Geef zowel de uitdrukking met de kans als de (benaderende) waarde ervan.

Oplossing: P-waarde = $P(Z < -1.129) = 1 - P(Z < 1.129) = 1 - 0.871 = 0.129$.

- (f) Maak een schets waarop je de testwaarde en de P-waarde aanduidt.

Oplossing:



- (g) Wat is de betekenis van de P-waarde?

Oplossing: De P-waarde is dat kans dat we bij een hypothesetest gebaseerd op een andere willekeurige steekproef van grootte 200 uit dezelfde populatie, een testwaarde zouden vinden die minstens even klein is (en dus minstens even sterk in de richting van H_1 wijst) als de testwaarde die we nu gevonden hebben.

- (h) Formuleer een zo volledig mogelijk inhoudelijk besluit op significantieniveau $\alpha = 1\%$.

Oplossing: De P-waarde is groter dan het significantieniveau $\alpha = 0.01$, we verwerpen de nulhypothese dus niet op significantieniveau $\alpha = 0.01$. De journalist kan **niet** besluiten dat de cijfers in het jaarlijks rapport van de luchtvaartmaatschappij overdreven zijn. Hij kan niet besluiten dat de proportie zeer tevreden passagiers kleiner is dan 83%.

- (i) Welk type fout maak je hier mogelijk? Leg uit wat het maken van deze fout concreet betekent in deze context.

Oplossing: Een type II fout. We besluiten dat de proportie zeer tevreden passagiers minstens 83% is, terwijl die in feite kleiner is dan 83%.

Naam:

1

Voornaam:

Studentennummer:

Opleiding:

4. Bij een onderzoek naar de morfometrische kenmerken van een specifieke soort buidelrat in Australië, werd de hoofdlengte (“HL” in cm) en de totale lengte (“TotL”, in cm) gemeten van 101 buidelratten. Men is geïnteresseerd in de vraag of men de totale lengte kan voorspellen aan de hand van de hoofdlengte. Op bladzijde 4 kan je enkele resultaten van de analyse in R terugvinden.

- (a) Stel de vergelijking van de regressierechte op waarmee je de totale lengte kan voorspellen a.d.h.v. de hoofdlengte.

Oplossing: We zoeken de parameters van het regressiemodel

$$\text{TotL} = \alpha + \beta \text{HL}$$

De parameters dienen nog berekend te worden. Voor $\hat{\beta}$ geeft dit:

$$\hat{\beta} = r \frac{s_{\text{TotL}}}{s_{\text{HL}}} = 0.6704 \frac{4.1968}{0.3519} = 7.9959$$

Hieruit kunnen we $\hat{\alpha}$ bepalen als zijnde

$$\hat{\alpha} = \overline{\text{TotL}} - \hat{\beta} \overline{\text{HL}} = 87.2693 - 7.9959 \cdot 9.2731 = 13.1224$$

De vergelijking van de regressierechte is dan

$$\text{TotL} = 13.1224 + 7.9959 \text{HL}$$

- (b) Vul de ontbrekende gegevens (a) t.e.m. (c) aan in de R-uitvoer.

Oplossing:

(a) Dit is de res. df., en wordt gegeven door $n - 2 = 99$.

(b) Dit is de SSM en is hier gelijk aan de MSM, dus 791.60

(c) Dit is de F-waarde en wordt gegeven door $\frac{\text{MSM}}{\text{MSE}} = \frac{791.6}{9.8} = 80.78$

(c) Test of dit regressiemodel zinvol is. Je mag aannemen dat alle modelveronderstellingen voldaan zijn.

i. Wat zijn de hypothesen die je test?

Oplossing:

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

ii. Geef de teststatistiek en de (eventueel benaderende) verdeling ervan onder H_0 . Geef ook het aantal vrijheidsgraden indien van toepassing.

Oplossing: $T = \frac{\hat{B}}{s.e.(\hat{B})} \sim_{H_0} t_{99}$

iii. Bepaal de testwaarde.

Oplossing: $t = \frac{7.9959}{0.8894} = 8.99$

iv. Bereken de P-waarde. Geef zowel de uitdrukking met de kans als de (benaderende) waarde ervan.

Oplossing: P-waarde is $2P(T > 8.99) = 2k$ met $T \sim t_{99}$ en $k < 0.005$.

v. Formuleer een zo volledig mogelijk inhoudelijk besluit op significantieniveau $\alpha = 5\%$.

Oplossing: We verwerpen H_0 en besluiten op 5% s.n. dat het regressiemodel zinvol is en we dus de totale lengte kunnen voorspellen a.d.h.v. de hoofdlengte.

- (d) Er blijkt een buidelrat van dezelfde populatie te zijn waarvan men vergeten is de totale lengte te meten. Men heeft echter wel zijn hoofdlength bepaald. Deze is 8.7 cm.
- i. Wat is de verwachte totale lengte van deze buidelrat?

Oplossing: Volgens ons regressiemodel is de verwachte totale lengte gegeven door

$$E[\widehat{\text{TotL}} \mid x = 8.7] = 13.1224 + 7.9959 \cdot 8.7 = 82.6867$$

Deze notatie is ook juist: $y_0 = 13.1224 + 7.9959 \cdot 8.7 = 82.6867$.

- ii. Construeer een 95% betrouwbaarheidsinterval voor de totale lengte van deze specifieke buidelrat.

Oplossing: Het 95% BI voor de individuele respons van een buidelrat met hoofdlength 8.7cm wordt gegeven door

$$\begin{aligned} & \left[\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2, 0.025} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{\text{HL}})^2}{(n-1)s_{\text{HL}}^2}} \right] \\ = & \left[13.1224 + 7.9959 \cdot 8.7 \pm 1.984 \cdot 3.13 \sqrt{1 + \frac{1}{101} + \frac{(8.7 - 9.2731)^2}{100 \cdot 0.1238}} \right] \\ = & [76.3647; 89.0088] \end{aligned}$$

R-uitvoer

```
> fit <- lm(TotL~HL)
> summary(fit)
Call: lm(formula = TotL ~ HL)
Residuals:
    Min       1Q   Median       3Q      Max
-6.9624 -2.3227  0.2755  2.0761  7.3139

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -          8.2538     -        -
HL           -          0.8894     -        -

Residual standard error: 3.13 on (a) degrees of freedom
Multiple R-squared:  0.4494, Adjusted R-squared:  0.4439
F-statistic: (c) on 1 and (a) DF,  p-value: -

> anova(fit)
Analysis of Variance Table
Response: TotL
      Df Sum Sq Mean Sq F value Pr(>F)
HL      1    (b)    791.6    (c)     -
Residuals (a) 969.71    9.8

> mean(HL)
[1] 9.273069
> mean(TotL)
[1] 87.26931
> var(HL)
[1] 0.1238135
> var(TotL)
[1] 17.61315
> cor(HL, TotL)
[1] 0.6704015
```