

Voorbehouden voor de correctoren					
Vraag 1(1)	Vraag 1(2)	Vraag 2	Vraag 3	Vraag 4	Totaal

Examen Statistiek & Data-analyse

18 januari 2021

Naam :

Studierichting :

Lees de volgende aanwijzingen eerst voor het examen te beginnen :

- Wie de vragen aanneemt en bekijkt, moet minstens 1 uur blijven zitten.
- Het examen bestaat uit **4 bundels**: bundel 1 bestaat uit vraag 1, bundel 2 bestaat uit vraag 2, bundel 3 bestaat uit vragen 3 en 4 en bundel 4 bevat de **R** uitvoer die je nodig hebt voor vraag 2. Controleer of je alle bundels gekregen hebt en **noteer op elke bundel je naam!**
- Beantwoord de vragen op de voorziene witruimte. Schrijf net en leesbaar! Je mag ook de extra antwoordruimte achteraan de bundel gebruiken, zorg dat je dit duidelijk aanduidt.
- Je mag gebruik maken van je rekenmachine, een zelf afgeprint formulierum en je statistische tabellen om dit examen op te lossen. Op het formulierum en de tabellen mag niets geschreven zijn!
- Bij het indienen van je examen geef je, naast de 3 bundels met examenvragen, ook de bundel met **R** uitvoer en je kladpapier af (met de **R** uitvoer en het kladpapier wordt **geen** rekening gehouden tijdens verbetering).
- Alleen een niet-grafisch rekentoestel is toegelaten. Alle communicatieapparatuur is strikt verboden.
- **Bij de toepassing van formules geef je de ingrediënten van de berekeningen duidelijk weer met weergave van voldoende tussenresultaten. Als je tijdsgebrek hebt hoeft je de finale uitkomst niet weer te geven. Het is de methode die primeert.**
- Indien je een T-test of T-betrouwbaarheidsinterval voor het verschil van twee gemiddelden van twee onafhankelijke normaal verdeelde populaties moet uitvoeren *dan dien je de test voor gelijkheid van de varianties NIET uit te voeren en mag je opteren voor de test met gelijke varianties omwille van de eenvoud van de berekeningen.*
- Een hypothesetest dient steeds uitgevoerd te worden op het 5% significantieniveau.
- Je hebt drie uur tijd om het examen op te lossen.

VEEL SUCCES !

Vraag 1 (9pt)

Bij hersenonderzoek wordt vaak gebruik gemaakt van dierenproeven. Bij een experiment naar de rol van schade aan het hippocampus gedeelte in de hersenen en verlies aan recent geheugen werden 18 primaten geoefend om 100 objecten te onderscheiden, telkens 20 objecten op respectievelijk 16, 12, 8, 4 en 2 weken voor een behandeling. Bij 11 toevallig gekozen apen werd de toegang tot de hippocampus nadien met een operatieve ingreep afgeblokt, terwijl voor de overige 7 (de controle groep) de hersenen niet werden behandeld (terwijl er wel een operatief ingrijpen onder narcose werd uitgevoerd). De onderstaande gegevens geven voor elke primate weer bij welke groep deze werd ingedeeld, naast het percentage correct herkende objecten die werden aangeleerd op kortere afstand (weken 2 en 4) voor de operatie (short term) en langere afstand (weken 8, 12 en 16) voor de operatie (long term).

Groep	Short	Long
Controle	85.00	71.67
Controle	80.00	71.67
Controle	85.00	48.33
Controle	82.50	65.00
Controle	72.50	68.33
Controle	80.00	78.33
Controle	77.50	70.00
Behandeld	62.50	73.33
Behandeld	85.00	66.67
Behandeld	65.00	71.67
Behandeld	62.50	66.67
Behandeld	62.50	70.00
Behandeld	62.50	66.67
Behandeld	57.50	65.00
Behandeld	60.00	60.00
Behandeld	65.00	75.00
Behandeld	52.50	66.67
Behandeld	60.00	70.00

Analyseer bovenstaande gegevens in verband met de volgende twee onderzoeksvragen. Hierbij kan je gebruik maken van de volgende gemiddelden en standaarddeviaties (genoteerd als $\bar{x}(s)$):

	Short	Long
Controle	80.4 (4.43)	67.6 (9.42)
Behandeld	63.2 (8.07)	68.3 (4.22)

- (1) De onderzoekers wensen na te gaan of er een significant lager gemiddelde van *Short Term Memory Scores* is bij de 'Behandelde groep' ten opzichte van de 'Controle groep'.
- (a) Maak een grafiek waarmee je de beslissing grafisch kunt ondersteunen. Beschrijf het type van grafiek die je voorstelt en duidt 3 eenvoudige kenmerken van elke verdeling (naar keuze) op de grafiek aan.
- (b) Formuleer een geschikte H_0 en H_1 waarmee je deze onderzoeksvraag kan beantwoorden, zoveel mogelijk in symbolen. Verklaar je symbolen.

(c) Geef de teststatistiek met de verdeling ervan onder H_0 . Geef de voorwaarden weer die voldaan moeten zijn om deze verdeling te kunnen gebruiken

(d) Bereken de testwaarde en de P-waarde. Maak hiervoor ook een schets met testwaarde en P-waarde

(e) Formuleer je besluit.

- (2) Men wenst eveneens na te gaan of er een significant lineair verband is tussen de short en long term metingen *bij de behandelde groep*.
- (a) Formuleer een geschikte H_0 en H_1 waarmee je deze onderzoeksvraag kan beantwoorden, zoveel mogelijk in symbolen. Verklaar je symbolen.
- (b) Geef de teststatistiek met de verdeling ervan onder H_0 . Geef de voorwaarden weer die voldaan moeten zijn om deze verdeling te kunnen gebruiken

(c) Bereken de testwaarde en de P-waarde. Maak hiervoor ook een schets met testwaarde en P-waarde

(d) Formuleer je besluit

Extra antwoordruimte vraag 1

Naam :

Vraag 2 (4pt)

Reeds in 1993 was er opschudding bij een senaatsverkiezing in Pennsylvania omtrent het gebruik van Absentee Ballots en het gebruik van stemmachines. De Democratische kandidaat behaalde 19127 stemmen via de stemmachines en de Republikein 19691. Anderzijds was het net andersom bij Absentee ballots: 1391 voor de Democraat en 366 voor de Republikein. Hieronder staan deze resultaten (verschil Democratisch - Republikeins resultaat) voor 21 voorafgaande verkiezingen in hetzelfde district in de 10 jaren daarvoor:

Verkiezing	Absentee ballot	Machine	Verkiezing	Absentee ballot	Machine
1	-841	-1375	12	364	2620
2	-436	-599	13	384	3596
3	-224	-129	14	320	3619
4	160	47	15	240	4195
5	597	7132	16	1365	4395
6	-356	1116	17	392	4512
7	300	940	18	404	5088
8	288	987	19	408	5570
9	609	1927	20	693	6286
10	412	2150	21	400	6533
11	167	2573	22	1025	-564

Aan de hand van de 21 eerdere observaties werd een lineair regressiemodel opgesteld (y =absentee, x =machine) Beantwoord nu volgende vragen aan de hand van de R uitvoer en de figuren in de extra bundel. Vermeld telkens welke uitvoer/figuur je gebruikt.

1. Welk regressiemodel wordt er gebruikt en wat zijn de schattingen van de parameters uit het model?

2. Welke modelveronderstellingen worden er gemaakt? Ga na of deze voldaan zijn en vermeld daar telkens bij op basis van welke grafiek(en)/output je deze conclusie stelt.

3. Vul de ontbrekende waarden (1) t.e.m. (3) aan in de bijgevoegde regressieuitvoer. Noteer de waarden en berekeningen hieronder, niet in de uitvoer zelf!

(1)

(2)

(3)

4. Bereken de standaardisering e_{22}/s_ϵ van de residu e_{22} op basis van deze regressieresultaten (gebaseerd op de overige gegevens) die hoort bij de bewuste 22ste verkiezing in 1993.

5. Doe op basis van dit nieuwe residu een uitspraak omtrent de uitzonderlijkheid van deze laatste verkiezing mede op basis van de residuplot en de normale kwantielplot van de residuen. Met andere woorden: vind je dat er voldoende evidentie is dat het resultaat van verkiezing 22 weinig verklaard wordt door de eerdere gegevens?

Extra antwoordruimte vraag 2

Naam :

Vraag 3 (4pt)

Een methode om een onderscheid te maken tussen graniet (G) en basalt (B) rotsformaties bestaat erin om de spectrum intensiteiten te meten van het infrarood spectrum van de zonnenergie dat reflecteert op het rotsoppervlak. Stel door R_1 , R_2 en R_3 de gemeten spectrum intensiteiten weer bij drie verschillende golflengtes. Voor graniet heeft men typisch $R_1 < R_2 < R_3$ terwijl bij basalt $R_3 < R_1 < R_2$ eerder voorkomt. Deze metingen worden vanuit een vliegtuig gemaakt en er komt daardoor variatie op de resultaten. Na veel proefvluchten komen de volgende twee voorwaardelijke kansverdelingen tot stand:

	Graniet	Basalt
$R_1 < R_2 < R_3$	60%	10%
$R_1 < R_3 < R_2$	25%	20%
$R_3 < R_1 < R_2$	15%	70%

Stel bovendien dat in een bepaalde regio de proportie graniet rots gelijk is aan 25% en basalt aan 75

- a. Bepaal voor die regio de kans op graniet rots als de meting $R_1 < R_3 < R_2$ oplevert. Evenzo voor basalt rots bij een $R_1 < R_3 < R_2$ resultaat.

- b. Als een meting de volgorde $R_1 < R_3 < R_2$ oplevert voor welk type rots zou je dan kiezen? Motiveer je keuze.

Naam :

Vraag 4 (4pt)

Beschouw een systeem van componenten $C_1, C_2, \dots, C_n$ waarbij de corresponderende levensduren $T_1, T_2, \dots, T_n$ van deze componenten onafhankelijk zijn en exponentieel verdeeld zijn met een gemiddelde van $1/\lambda$ jaar. Voor de exponentiële verdeling geldt dat de verwachtingswaarde en standaarddeviatie gelijk zijn aan $1/\lambda$.

- a. Met $S_n = T_1 + T_2 + \dots + T_n$ de totale levensduur van alle componenten samen, bepaal een benadering voor $P(S_n > a)$ met $n = 100$, $\lambda = 1/5$ en $a = 600$

- b. Benader de mediaan van de totale levensduur S_n op basis van de benadering in b. Maak terug gebruik van dezelfde waarden voor n , λ en a .

Extra antwoordruimte vraag 3 en vraag 4

```

PA <- read.csv(file=file.choose(), sep=";")
> machine <- PA$Machine
> absentee <- PA$Absentee
>
> resultaat=lm(absentee~machine)
> summary(resultaat)

Call:
lm(formula = absentee ~ machine)

Residuals:
    Min       1Q   Median       3Q      Max
-566.47 -197.46  -58.72   239.88   908.42

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -100.30776   114.95191   -0.873   0.39377
machine      0.12671     0.03039     (1)   0.00052 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

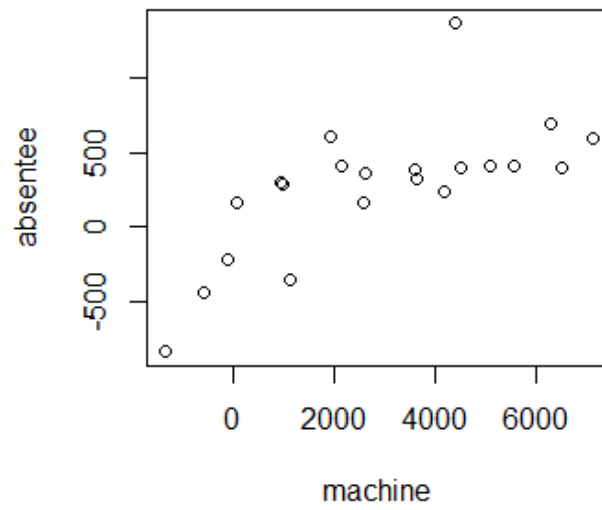
Residual standard error: 336 on 19 degrees of freedom
Multiple R-squared:  0.4778,    Adjusted R-squared:  0.4503
F-statistic: 17.38 on 1 and 19 DF,  p-value: 0.0005203

>
> anova(resultaat)
Analysis of Variance Table

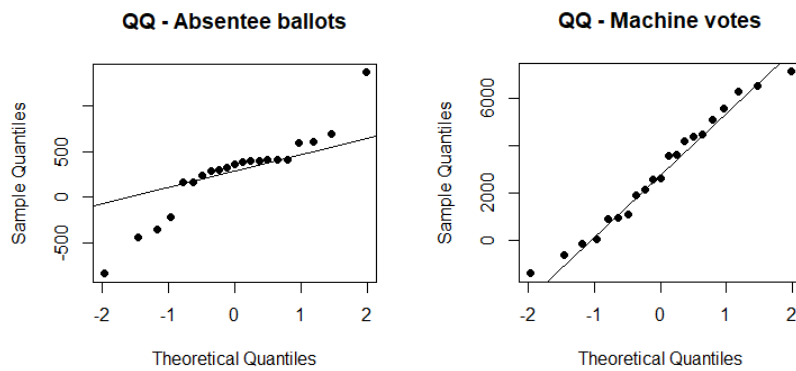
Response: absentee
      Df Sum Sq Mean Sq F value    Pr(>F)
machine  1 1962106 1962106   (3) 0.0005203 ***
Residuals 19 2144453   (2)
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

>
> mean(absentee)
[1] 268.8571
> var(absentee)
[1] 205327.9
> mean(machine)
[1] 2913.476
> var(machine)
[1] 6110468

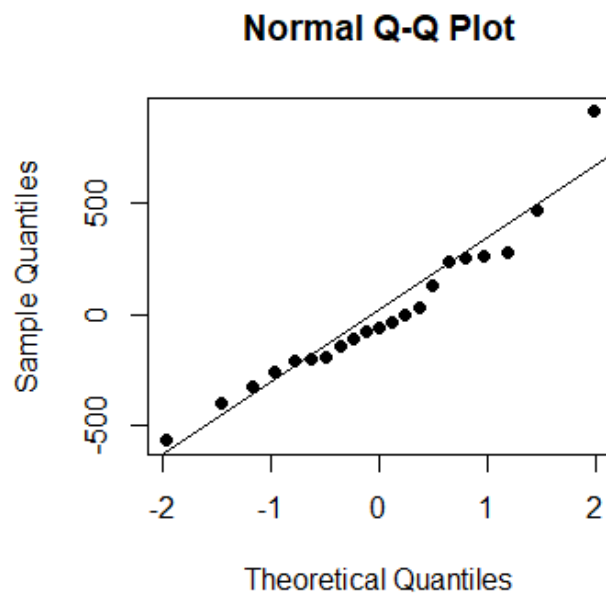
```



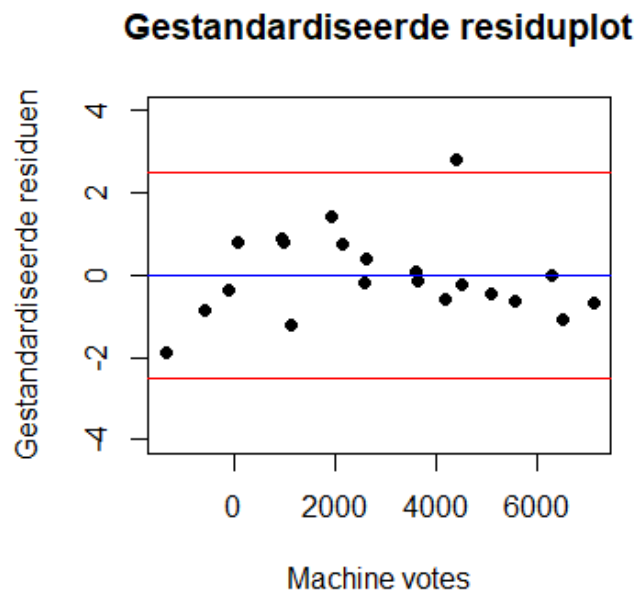
Figuur 1: Scatterplot Absentee versus Machine verschillen



Figuur 2: Normale kwantielplots van Absentee verschillen (links) en machine stemverschillen (rechts)



Figuur 3: Normale kwantielplot residuen r_i



Figuur 4: Residuplot