

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

Examen Statistiek & Data-analyse

27 januari 2021

Naam :

Studierichting :

Lees 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 1 en 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 formularium en je statistische tabellen om dit examen op te lossen. Op het formularium 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)

Duchenne Muscular Dystrophy (DMD) is een genetisch overdraagbare ziekte van moeder op de kinderen. Dochters die de ziekte dragen vertonen geen symptomen en kunnen de ziekte dus onbewust doorgeven aan de eigen kinderen. Zonder test zou een vrouw enkel een vermoeden van de ziekte kunnen hebben als een zoon de ziekte ontwikkelt. Er werd in het verleden gezocht naar een bloedtest om de ziekte te ontdekken. In de bijgaande data set worden metingen van creatine kinase (CK) en hemopexine (H) enzymen bij 19 controle proefpersonen (die geen drager zijn) en 19 DMD dragers weergegeven.

Groep	CK	H	gestandaardiseerde CK-scores	gestandaardiseerde H-scores
Controle	21	74.5	-0.98	-0.50
Controle	95	69.8	2.38	-0.82
Controle	40	72.7	-0.11	-0.62
Controle	48	76	0.25	-0.39
Controle	39	88.5	-0.16	0.47
Controle	30	82.7	-0.57	0.07
Controle	38	85	-0.20	0.23
Controle	27	87.2	-0.70	0.38
Controle	74	80.4	1.43	-0.09
Controle	34	80.5	-0.39	0.08
Controle	28	82.5	-0.66	0.06
Controle	37	93	-0.25	0.78
Controle	97.5	34	2.50	-3.29
Controle	37	98	-0.25	1.13
Controle	30	80	-0.57	-0.12
Controle	37	98	-0.25	0.13
Controle	24	100.5	-0.84	1.30
Controle	41	78.5	-0.07	-0.22
Controle	30	90.5	-0.57	0.61
DMD	167	89	0.00	-0.50
DMD	104	81	-0.39	-1.17
DMD	30	108	-0.85	-1.07
DMD	65	87	-0.63	-0.67
DMD	440	107	1.69	0.99
DMD	58	88.2	-0.67	-0.57
DMD	129	93.1	-0.24	-0.16
DMD	265	83.5	0.61	-0.96
DMD	285	79.5	0.73	-1.29
DMD	124	92	-0.27	-0.25
DMD	53	76	-0.70	-1.58
DMD	657	104	3.04	0.74
DMD	168	82.5	0.01	-1.04
DMD	286	109.5	0.74	1.20
DMD	73	105.5	-0.58	0.86
DMD	19	100.5	-0.92	0.45
DMD	113	97	-0.33	0.16
DMD	57	105	-0.68	0.82
DMD	78	118	-0.55	1.90

Bij de onderstaande vragen in verband met deze data kan je gebruik maken van de volgende gemiddelden en standaarddeviaties (genoteerd als $\bar{x}$ (s)):

	CK	H
Controle	42.5 (22.0)	84.9 (9.0)
DMD	166.9(161.1)	95.1(12.1)

- (1) In bijlage A worden de normale kwantielplots getoond voor CK en voor H en dit voor de twee groepen (controle en DMD dragers). Per variabele werden de normale kwantielplots voor de beide groepen op één assenkruis geplaatst om de vergelijking tussen beide groepen te faciliteren.
 - (a) Beoordeel de normaliteit van de CK en H gegevens per groep uit deze grafiek. Indien een variabele duidelijk niet normaal verdeeld is, welke transformatie zou je toepassen om de gegevens naar normaliteit proberen te transformeren? Motiveer je antwoord.

- (b) Wat kan je uit deze grafieken leren omtrent de verschillen in gemiddelde en variantie tussen beide groepen wat betreft CK, respectievelijk H?

- (2) Ga na of er een significant verschil is tussen beide groepen (controle en DMD) wat betreft de proporties H waarden die minstens 85 eenheden bedragen.
- (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.

(3) Ga na of er een significant lineair verband is tussen CK en H metingen *bij de DMD 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.

Kan je op basis van de bijgaande samenvattende kenmerken en grafieken nagaan of deze voorwaarden voldaan zijn? Geef eventuele afwijkingen weer.

- (c) Bereken de testwaarde en de P-waarde. Maak hiervoor ook een schets met testwaarde en P-waarde. Hier mag je dus eventuele afwijkingen gemeld in (b) negeren.

- (d) Formuleer je besluit.

Extra antwoordruimte vraag 1

Naam :

Vraag 2 (4pt)

Een van de belangrijkste contaminanten die over Europese landen werd verspreid na het incident met de kerncentrale van Chernobyl in 1987 was radioactief cesium. Om de invloed van vervuilde grond op de gewassen te meten werden in Noord Itali in 16 bosgebieden metingen uitgevoerd van concentraties van cesium (Bq/kg) in de grond en in paddenstoelen.

Staal	Bodem	Paddenstoel	Staal	Bodem	Paddenstoel
1	33	1	9	279	34
2	55	9	10	329	41
3	138	14	11	82	46
4	319	17	12	86	49
5	415	20	13	55	53
6	425	17	14	60	60
7	442	14	15	144	79
8	475	15	16	292	99

Aan de hand van deze 16 stalen werd een lineair regressiemodel opgesteld (y =’Paddenstoel’, x =’Bodem’). Beantwoord nu de volgende vragen aan de hand van de R uitvoer en de figuren in bijlage B. 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. 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)

3. 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. Specificeer ook de eventuele afwijkingen die je detecteert

Extra antwoordruimte vraag 2

Naam :

Vraag 3 (4pt)

Een systeem bevat twee componenten A en B. De kans dat B blijft functioneren gedurende de garantieperiode is 0.9. De kans dat ze beide blijven functioneren onder garantie is 0.75, terwijl de kans dat er tenminste één van beide blijft functioneren gelijk is aan 0.96 .

- a. Bepaal de kans dat B blijft functioneren onder garantie als gegeven is dat de A component blijft functioneren

- b. Functioneren de beide componenten onafhankelijk van elkaar? Verklaar.

Vraag 4 (4pt)

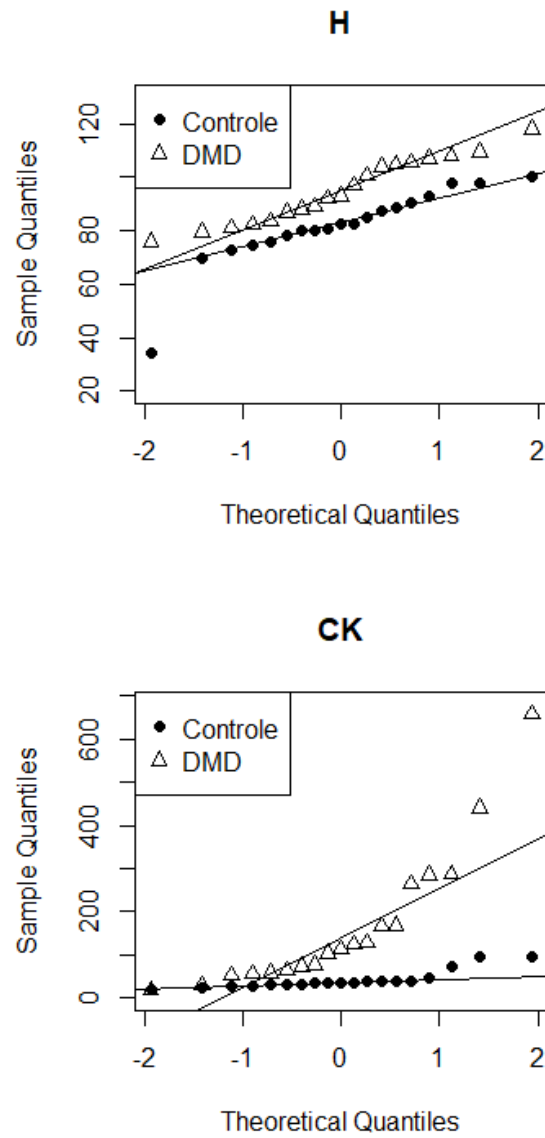
a. Bereken benaderend de kans dat in een groep van 50 starters tenminste 25 onder hen een foutloze oplossing aanbiedt.

Page 13

- c. Stel nu dat men de echte proportie p van foutloze oplossingen niet kent. Bereken een 95% betrouwbaarheidsinterval voor p op basis van een steekproef van 25 foutloze oplossingen in een groep van 50 studenten.

Extra antwoordruimte vraag 3 en vraag 4

BIJLAGE A



Figuur 1: Normale kwantielplots H (boven) en CK (onder), voor controle groep én voor DMD groep op één assenkruis geplaatst

BIJLAGE B

```

> Chernobyl<- read.csv(file=file.choose(), sep=" ")
> bodem <- Chernobyl$Bodem
> paddenstoel <- Chernobyl$Paddenstoel
>
>
> resultaat=lm(paddenstoel~bodem)
> summary(resultaat)

Call:
lm(formula = paddenstoel ~ bodem)

Residuals:
    Min       1Q   Median       3Q      Max
-41.658 -13.938  -4.061   9.744  65.908

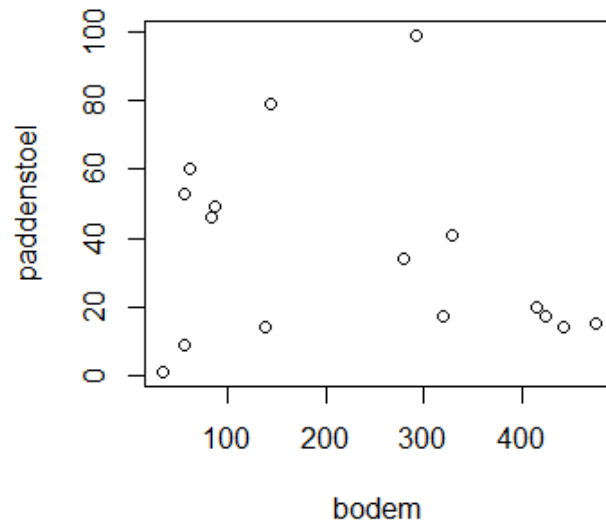
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  43.87726   12.25571   3.580  0.00301 **
bodem        -0.03693    0.04454   (1)    (2)
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: (3) on 14 degrees of freedom
Multiple R-squared:  0.04682, Adjusted R-squared: -0.02126
F-statistic: 0.6877 on 1 and 14 DF, p-value: 0.4208

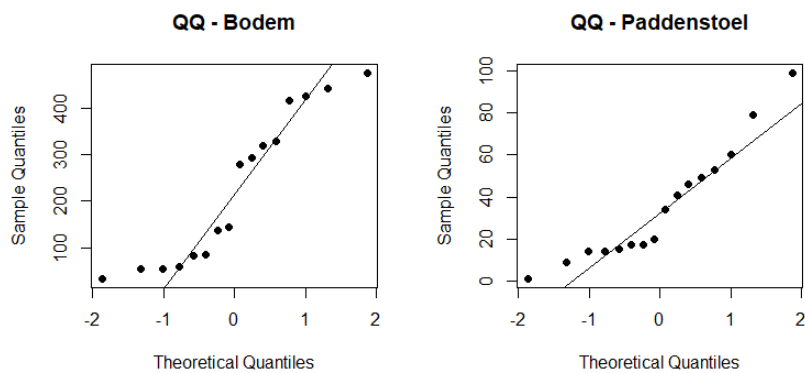
>
> anova(resultaat)
Analysis of Variance Table

Response: paddenstoel
      Df Sum Sq Mean Sq F value Pr(>F)
bodem   1    530   529.96  0.6877 0.4208
Residuals 14  10788   770.57
>
> mean(paddenstoel)
[1] 35.5
> var(paddenstoel)
[1] 754.5333
> mean(bodem)
[1] 226.8125
> var(bodem)

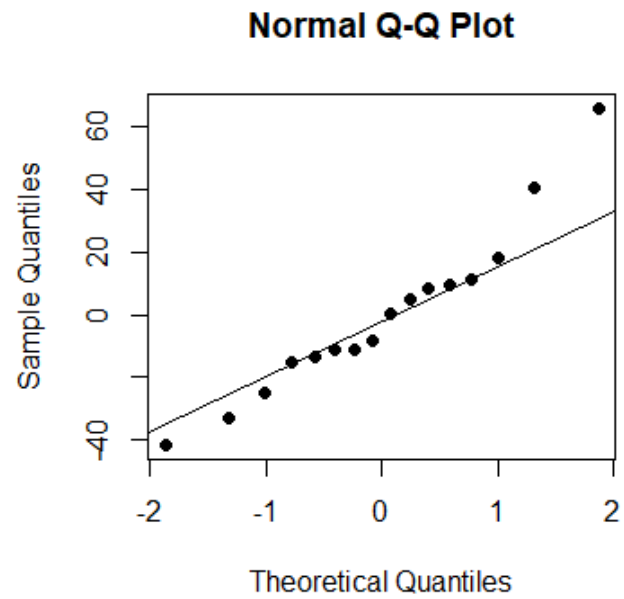
```



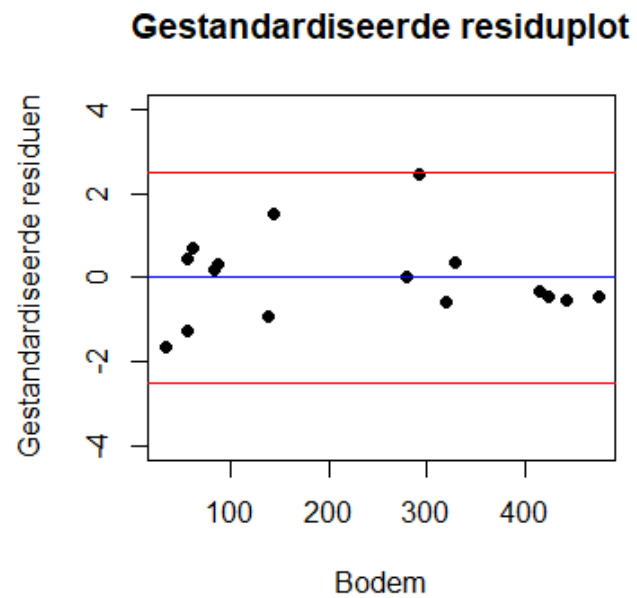
Figuur 2: Scatterplot 'Paddenstoel' versus 'Bodem' scores



Figuur 3: Normale kwantielplots van bodem metingen (links) en paddenstoel metingen (rechts)



Figuur 4: Normale kwantielplot residuen



Figuur 5: Residuplot