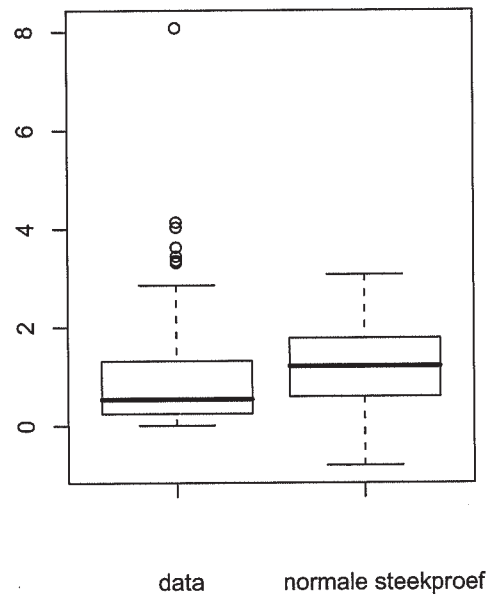

Tentamen Empirische Methoden
voor studenten AI, IMM en Informatica
Vrije Universiteit, Faculteit Exacte Wetenschappen
2 juni 2010, 18.30-21.15 uur

NB. Alleen het gebruik van een simpele rekenmachine is toegestaan; gebruik van grafische en/of programmeerbare rekenmachines, mobiele telefoons, e.d. is niet toegestaan.

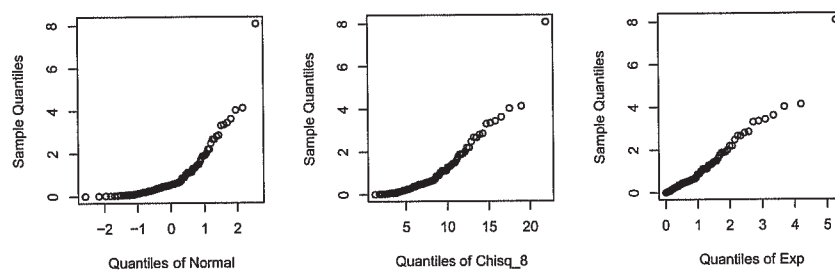
Bijlage: formuleblad (2 blz) en tabellenblad

Onderstaande negen opdrachten wegen even zwaar.

1. a) De linker boxplot in Figuur 1 is een boxplot van data afkomstig van een onbekende verdeling, de rechter is een boxplot van een even grote steekproef uit een normale verdeling met dezelfde verwachting en standaardafwijking als de onbekende verdeling. Wat kun je aflezen uit deze boxplots met betrekking tot de locatie, schaal en vorm van de onbekende verdeling van de data van de linker boxplot?
(De whiskers van de boxplots verbinden de box in dit geval met de meest extreme datapunten die niet meer dan 1.5 maal de interkwartiel afstand van de box afliggen.)
- b) Welk van de twee locatiematen zal groter zijn voor de dataset uit de onbekende verdeling, het steekproefgemiddelde of de steekproefmediaan en waarom? Geef indien mogelijk, een schatting van de populatieparameters mediaan, modus en verwachting van de onbekende verdeling waaruit de data afkomstig zijn.
- c) In Figuur 2 staan voor de dataset uit de onbekende verdeling een normale QQ -plot, een QQ -plot tegen de quantielen van de chikwadraatverdeling met 8 vrijheidsgraden en een QQ -plot tegen de quantielen van de exponentiële verdeling. Wat kun je uit deze grafieken afleiden over de *vorm* van de verdeling van de data ten opzichte van die van de standaard normale verdeling? Zeg in het bijzonder ook iets over de staarten van de verdeling ten opzichte van die van de standaard normale. Beargumenteer je antwoord.
- d) Welke locatieschaal-familie zou op grond van je antwoorden bij de voorgaande onderdelen mogelijk geschikt kunnen zijn voor deze data?



Figuur 1: Boxplot van data uit onbekende verdeling en boxplot van steekproef uit normale verdeling met zelfde verwachting en standaardafwijking als onbekende verdeling.



Figuur 2: QQ -plots van de data.

2. Toen ook vrouwen gevechtspiloot konden worden, diende het ontwerp van schietstoelen te worden aangepast. Hieronder staan de gewichten in kg, naar grootte geordend, van 25 willekeurig gekozen vrouwen.

41	42	43	45	48
50	56	59	62	65
67	67	68	68	68
72	72	73	73	74
79	85	87	91	96.

Het steekproefgemiddelde is $\bar{x} = 66.0$ en de steekproefstandaarddeviatie is $s = 15.4$.

- Welke parameter wordt geschat met het steekproefgemiddelde?
- De oorspronkelijke schietstoelen waren ontworpen voor mannen met een gewicht tussen 63 en 95 kg. Bepaal op basis van relatieve frequentie een schatting voor de kans dat het gewicht van een vrouw tussen de 63 en 95 kg ligt.

Veronderstel nu dat het gewicht van vrouwen normaal verdeeld is met verwachting $\mu = 65$ kg en standaardafwijking $\sigma = 13$ kg.

- Geef op basis van bovenstaande veronderstelling van de normaliteit van het gewicht van vrouwen een schatting voor de kans dat het gewicht van een willekeurige vrouw tussen de 63 en 95 kg ligt. Je kunt hiervoor bijgeleverde Tabel 5.1 gebruiken.
 - De schattingen van de kansen onder b) en c) verschillen. Noem twee mogelijke oorzaken.
 - Geef, op basis van bovenstaande veronderstelling van de normaliteit van het gewicht van vrouwen, de kans dat het gemiddelde gewicht van 25 vrouwen tussen de 63 en 95 kg ligt. Ook hiervoor kun je Tabel 5.1 gebruiken.
3. Beschouw het experiment waarbij iemand een zuivere munt opgooit en een zuivere dobbelsteen werpt.
- Construeer een kansruimte voor dit experiment.
 - Wat is binnen deze kansruimte de kans op de gebeurtenis 'kop en 6' en wat is de kans op de gebeurtenis 'kop of 6'?
 - Zij A de gebeurtenis 'kop en 6' en B de gebeurtenis 'kop of 6'. Zijn A en B onafhankelijk? Waarom (niet)?

Als bij het opgooien van de munt kop bovenkomt, krijgt de persoon 2 euro; als munt bovenkomt niets. Bovendien krijgt de persoon 1 euro als er 2 of 4 wordt gegooid met de dobbelsteen en 2 euro als er 6 wordt gegooid met de dobbelsteen.

- d) Beschouw binnen de kansruimte uit onderdeel a de stochastische variabele X die het totaal verkregen bedrag bij het experiment van de worpen met munt en dobbelsteen beschrijft. Construeer de kansfunctie $p(x) = P(X = x)$ van X netjes op basis van de formele definitie.
- e) Bereken de verwachting EX van X ; geef niet alleen de uitkomst, maar laat zien hoe je hier aan komt.
4. Bij toepassing van een leugendetector geldt: als de proefpersoon liegt, geeft het apparaat dit aan met kans 0.88; als de proefpersoon de waarheid vertelt, geeft het apparaat dit aan met kans 0.86. Bij het screenen van personen om veiligheidsredenen geldt dat bij het beantwoorden van een bepaalde vraag 99% van de personen de waarheid spreekt. Personen liegen of spreken de waarheid onafhankelijk van elkaar. Geef bij onderstaande onderdelen niet alleen het antwoord, maar laat zien hoe je hieraan komt en geef kort aan welke rekenregel of eigenschap van kansen je bij de berekening gebruikt.
(*Je mag in je antwoord een breuk of product laten staan.*)
- a) Wat is de kans dat bij het beantwoorden van de vraag drie willekeurige personen liegen?
- b) Wat is de kans dat bij het beantwoorden van de vraag tenminste één van drie willekeurige personen liegt?
- c) Wat is de kans dat bij het beantwoorden van de vraag de detector aangeeft dat een persoon liegt, als deze persoon in werkelijkheid de waarheid spreekt?
- d) Wat is de kans dat bij het beantwoorden van de vraag de detector aangeeft dat een persoon liegt?
- e) Wat is de kans dat bij het beantwoorden van de vraag een persoon de waarheid spreekt, gegeven dat de detector aangeeft dat hij liegt?
5. In opgave 2 zijn de gewichten gegeven van $n = 25$ vrouwen. Het steekproefgemiddelde van de gewichten is $\bar{x} = 66.0$ en de steekproefstandaarddeviatie $s = 15.4$. In deze opgave bepalen we op basis van deze gewichten een betrouwbaarheidsinterval voor het populatiegemiddelde μ van het gewicht van vrouwen. Op basis van de Centrale Limietstelling weten we dat, als n voldoende groot is, het gemiddelde van n gewichten met een kans van ongeveer 95% tussen de waarden $\mu - 2s_n/\sqrt{n}$ en $\mu + 2s_n/\sqrt{n}$ ligt.
- a) Leid uit deze kansuitspraak de *algemene uitdrukking* af voor een 95% betrouwbaarheidsinterval voor de onbekende waarde μ .
- b) Geef op basis van deze algemene uitdrukking het 95% betrouwbaarheidsinterval voor μ uitgaande van de gegeven dataset.
- c) Wat is de interpretatie van het in b) bepaalde interval?

- d) Is het op grond van het gevonden betrouwbaarheidsinterval aannemelijk dat het populatiegemiddelde μ van het gewicht van vrouwen gelijk is aan $\mu = 65$ kg? Waarom (niet)?
- e) Vind je het gebruik van de Centrale limietstelling voor deze data gerechtvaardigd? En als verondersteld mag worden dat het gewicht van vrouwen normaal verdeeld is? Licht beide antwoorden kort toe.
6. Voorstanders van de kilometerheffing beweren dat 90% van de autobezitters voor kilometerheffing is. Iemand die dit wat overdreven acht, neemt een steekproef van 100 autobezitters. Van deze 100 personen gaven er 84 aan voor kilometerheffing te zijn. We zijn geïnteresseerd in de vraag of deze uitslag een voldoende indicatie is dat minder dan 90% van personen in de populatie voor kilometerheffing is. Zij p de fractie van personen in de populatie die voor kilometerheffing is.
- a) Stel de nulhypothese H_0 en de alternatieve hypothese H_a met betrekking tot de fractie p .
- b) Bepaal de gebruikelijke (punt)schatting $\hat{p}$ voor p .
- c) Bereken de standaardscore z (toetsingsgrootte) voor $\hat{p}$.
- d) Bepaal met behulp van Tabel 5.1 de P -waarde behorend bij deze score.
- e) Wat is de conclusie van de toets bij een significantieniveau van 0.05?
7. Hieronder staan twee sets lichaamstemperaturen (in $^{\circ}\text{C}$).

steekproef 1: 36.1 35.7 36.4 35.8 36.6 37.3

steekproef 2: 36.7 37.0 37.1 36.7 37.0 36.4

De steekproefgemiddelden en steekproefstandaarddeviaties voor deze data zijn:

steekproef 1: $\bar{x}_1 = 36.3$, $s_1 = 0.59$

steekproef 2: $\bar{x}_2 = 36.8$, $s_2 = 0.26$

paarsgewijze verschillen: $\bar{x}_d = -0.5$, $s_d = 0.75$.

Neem aan dat alle metingen zijn gedaan om 8 uur in de ochtend: steekproef 1 aan een groep van 6 mannen, steekproef 2 aan een groep van 6 vrouwen. Veronderstel dat de populatievarianties van de lichaamstemperatuur om 8 uur in de ochtend van mannen en vrouwen gelijk zijn.

- a) Zijn dit gepaarde of onafhankelijke steekproeven?
- b) Geef een (punt)schatting van het verschil tussen de gemiddelde lichaamstemperatuur om 8 uur van mannen en die van vrouwen.
- c) Toets met een geschikte t -toets de claim dat de gemiddelde lichaamstemperatuur om 8 uur van mannen verschillend is van die van vrouwen. Neem significantieniveau $\alpha = 0.1$. Je kunt ook hierbij bijgeleverde Tabel 10.1 gebruiken.
- d) Is het gebruik van de t -toets voor deze situatie gerechtvaardigd? Waarom wel/niet?

	menselijk ingrijpen	natuurlijke factoren
jonger dan 30	101 (130)	40 (49)
ouder dan 50	120 (129)	70 (61)

Tabel 1: Opinie-onderzoek opwarming aarde.

8. Tabel 1 geeft enkele resultaten van een opinieonderzoek naar opwarming van de aarde. Aan respondenten die antwoordden dat naar hun mening de aarde inderdaad langzaam opwarmt, is gevraagd of zij menselijk ingrijpen of natuurlijke factoren als voornaamste oorzaak zien. De vraag is of jongere personen (jonger dan 30) anders over de oorzaken van opwarming van de aarde denken dan oudere personen (ouder dan 50). Dit wordt onderzocht d.m.v. een toets. De waarden tussen haakjes in de tabel zijn de verwachte frequenties als leeftijd geen rol speelt in de mening over de oorzaken van opwarming van de aarde.

- Formuleer de nul- en alternatieve hypothese.
- Toets de hypothesen met een chikwadraattoets. Neem significantieniveau $\alpha = 0.05$. Je kunt hierbij bijgeleverde Tabel 10.7 gebruiken.
- Beschrijf kort hoe men de claim zou kunnen onderzoeken of oudere personen relatief vaker menselijk ingrijpen als voornaamste oorzaak zien. (Er zijn meerdere antwoorden mogelijk.) N.B. Je hoeft dit onderzoek niet zelf te doen.

9. In Figuur 3 staat een scatter plot van metingen aan de hoeveelheid geproduceerd materiaal en de temperatuur tijdens de vervaardiging van het materiaal.

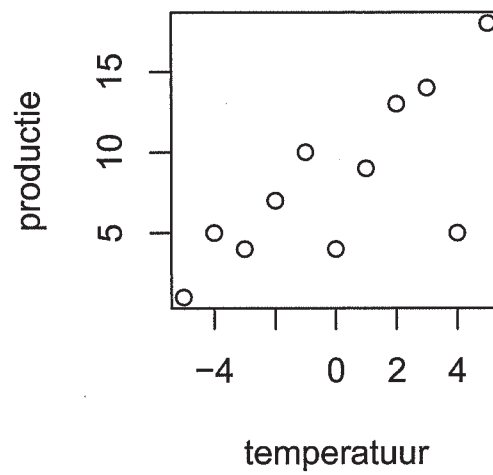
- Trek in gedachten de ‘beste’ lijn door de punten en schat op het oog de steekproefcorrelatiecoëfficiënt.
- Hoeveel variatie in de y -variabele wordt ongeveer verklaard door de ‘beste’ lijn?
- Zijn er uitbijters? Zo ja, bespreek dan hun effect op de sterkte van de correlatie en de ‘beste’ lijn.

Laat b_1 de te schatten richtingscoëfficiënt (helling) van de ‘beste’ lijn zijn, $\hat{b}_1 = 1.1$ de bijbehorende schatting, $\hat{s}_{\hat{b}_1} = 0.35$ de geschatte standaardafwijking van $\hat{b}_1$ en n het aantal data paren. Dan heeft onder aanname van normaliteit van de meetfouten de grootte

$$T_{b_1} = \frac{\hat{b}_1 - b_1}{\hat{s}_{\hat{b}_1}}$$

een t -verdeling met $n - 2$ vrijheidsgraden.

- d) Toets de claim dat $b_1 = 0$, ofwel dat er geen lineair verband is tussen de variabelen productie en temperatuur, met behulp van bovenstaande grootte. Neem significantieniveau $\alpha = 0.05$.



Figuur 3: Scatter plot.

Formuleblad bij Tentamen Empirische Methoden

Kansrekening

$(\Omega, \mathcal{A}, P)$ kansruimte

$A, B_1, B_2, \dots, B_m \in \mathcal{A}$ gebeurtenissen

$B_1, B_2, \dots, B_m$ een partitie van Ω met $P(B_i) > 0$ voor alle $i \in \{1, 2, \dots, m\}$; $r \in \{1, 2, \dots, m\}$.

Regel voor Totale Waarschijnlijkheid:

$$P(A) = \sum_{i=1}^m P(A \cap B_i) = \sum_{i=1}^m P(A|B_i)P(B_i)$$

Regel van Bayes:

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^m P(A|B_i)P(B_i)} = \frac{P(A|B_r)P(B_r)}{\sum_{i=1}^m P(A|B_i)P(B_i)}$$

Twee steekproeven

(Onderstaande formules gelden onder geschikte voorwaarden.)

(i) Voor twee *onafhankelijke* steekproeven en als $\sigma_1^2 = \sigma_2^2 = \sigma^2$, heeft de grootte

$$T^{(2)} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\bar{s}_{n_1, n_2} \sqrt{1/n_1 + 1/n_2}}$$

een t -verdeling met $n_1 + n_2 - 2$ vrijheidsgraden. Hier is $\bar{s}_{n_1, n_2}$ de wortel van de ‘gepoolde’ steekproefvariantie $\bar{s}_{n_1, n_2}^2$ gegeven door

$$\bar{s}_{n_1, n_2}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

(ii) Voor twee *onafhankelijke* steekproeven met $\sigma_1^2 \neq \sigma_2^2$, wordt de noemer van $T^{(2)}$ vervangen door

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

en is de verdeling van $T^{(2)}$ ongeveer een t -verdeling met *aangepast* aantal vrijheidsgraden het volgende getal afgerond naar het dichtsbijzijnde gehele getal:

$$df_{adjust} = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

(iii) Voor twee *onafhankelijke* steekproeven is een betrouwbaarheidsinterval voor $(p_1 - p_2)$ met betrouwbaarheid van ongeveer 95%:

$$[(\hat{p}_1 - \hat{p}_2) - E, (\hat{p}_1 - \hat{p}_2) + E],$$

met

$$E = 2\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

(iv) Als bovendien $p_1 = p_2$, dan heeft

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1 - \bar{p})/n_1 + \bar{p}(1 - \bar{p})/n_2}}$$

bij benadering een standaard normale verdeling. Hier is $\bar{p} = (x_1 + x_2)/(n_1 + n_2)$ de ‘gepoolde’ steekproef fractie.

Correlatie

Onder voorwaarden heeft de grootte

$$T_{cor} = \frac{r - \rho}{\sqrt{(1 - r^2)/(n - 2)}}$$

een t -verdeling met $n - 2$ vrijheidsgraden. Hier is r de steekproefcorrelatiecoëfficiënt gegeven door

$$r = \frac{1}{n - 1} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}.$$

Lineaire regressie

De richtingscoëfficiënt (helling) $\hat{b}_1$ en de intercept $\hat{b}_0$ van de ‘beste’ rechte lijn worden gegeven door

$$\hat{b}_1 = r \frac{s_y}{s_x}$$

en

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}.$$

Tabellen standaard normale, t- en chikwadraatverdeling bij tentamen Empirische Methoden

Table 5.1 Standard Scores and Percentiles for a Normal Distribution
(cumulative values from the left)

Standard score	%	Standard score	%	Standard score	%	Standard score	%
-3.5	0.02	-1.0	15.87	0.0	50.00	1.1	86.43
-3.0	0.13	-0.95	17.11	0.05	51.99	1.2	88.49
-2.9	0.19	-0.90	18.41	0.10	53.98	1.3	90.32
-2.8	0.26	-0.85	19.77	0.15	55.96	1.4	91.92
-2.7	0.35	-0.80	21.19	0.20	57.93	1.5	93.32
-2.6	0.47	-0.75	22.66	0.25	59.87	1.6	94.52
-2.5	0.62	-0.70	24.20	0.30	61.79	1.7	95.54
-2.4	0.82	-0.65	25.78	0.35	63.68	1.8	96.41
-2.3	1.07	-0.60	27.43	0.40	65.54	1.9	97.13
-2.2	1.39	-0.55	29.12	0.45	67.36	2.0	97.72
-2.1	1.79	-0.50	30.85	0.50	69.15	2.1	98.21
-2.0	2.28	-0.45	32.64	0.55	70.88	2.2	98.61
-1.9	2.87	-0.40	34.46	0.60	72.57	2.3	98.93
-1.8	3.59	-0.35	36.32	0.65	74.22	2.4	99.18
-1.7	4.46	-0.30	38.21	0.70	75.80	2.5	99.38
-1.6	5.48	-0.25	40.13	0.75	77.34	2.6	99.53
-1.5	6.68	-0.20	42.07	0.80	78.81	2.7	99.65
-1.4	8.08	-0.15	44.04	0.85	80.23	2.8	99.74
-1.3	9.68	-0.10	46.02	0.90	81.59	2.9	99.81
-1.2	11.51	-0.05	48.01	0.95	82.89	3.0	99.87
-1.1	13.57	0.0	50.00	1.0	84.13	3.5	99.98

let op: dit zijn percentages!

Table 10.7 Critical Values of χ^2 : Reject H_0 Only If $\chi^2 >$ Critical Value

Table size (rows $\times$ columns)	Significance level	
	0.05	0.01
2 $\times$ 2	3.841	6.635
2 $\times$ 3 or 3 $\times$ 2	5.991	9.210
3 $\times$ 3	9.488	13.277
2 $\times$ 4 or 4 $\times$ 2	7.815	11.345
2 $\times$ 5 or 5 $\times$ 2	9.488	13.277

0.95- en 0.99-quantielen
chikwadraatverdeling
(de kritieke waarden)

Table 10.1 Critical Values of t

Degrees of freedom ($n - 1$)	Area in one tail	
	0.025	0.05
	Area in two tails	
	0.05	0.10
1	12.706	6.314
2	4.303	2.920
3	3.182	2.353
4	2.776	2.132
5	2.571	2.015
6	2.447	1.943
7	2.365	1.895
8	2.306	1.860
9	2.262	1.833
10	2.228	1.812
11	2.201	1.796
12	2.179	1.782
13	2.160	1.771
14	2.145	1.761
15	2.131	1.753
16	2.120	1.746
17	2.110	1.740
18	2.101	1.734
19	2.093	1.729
20	2.086	1.725
21	2.080	1.721
22	2.074	1.717
23	2.069	1.714
24	2.064	1.711
25	2.060	1.708
26	2.056	1.706
27	2.052	1.703
28	2.048	1.701
29	2.045	1.699
30	2.042	1.697
31	2.040	1.696
32	2.037	1.694
34	2.032	1.691
36	2.028	1.688
38	2.024	1.686
40	2.021	1.684
50	2.009	1.676
100	1.984	1.660
Large	1.960	1.645

aantal
vrijheids-
graden df

$t_{df, 0.975}$
quantiel

$t_{df, 0.95}$
quantiel