
Tentamen Toegepaste Statistiek
voor studenten IMM en AI
Vrije Universiteit, Faculteit Exacte Wetenschappen
23 maart 2009, 12.00-14.00 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.

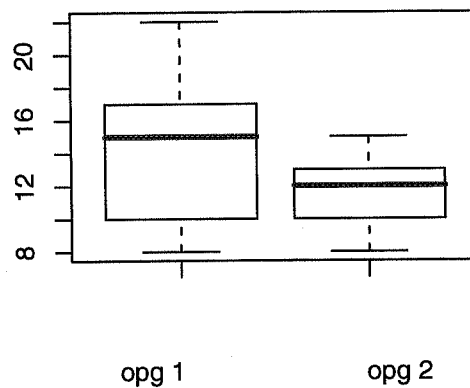
Bijlage: twee rekenregels en Tabel 5.1.

Onderstaande zeven opdrachten wegen even zwaar.

1. Zijn de volgende uitspraken zinvol (of correct)? Motiveer je antwoord.
 - a) Om een steekproef van lengten die geclassificeerd zijn naar haarkleur van een subject, te visualiseren wordt een histogram gebruikt.
 - b) De kans op een even aantal ogen bij het gooien van een dobbelsteen is 0.5. De kans op een zwarte kaart bij trekking van een kaart uit een kaartspel is 0.5. De experimenten zijn onafhankelijk, dus de gebeurtenissen zijn disjunct en de kans op beide gebeurtenissen tezamen is $0.5 + 0.5 = 1$.
 - c) Om een populatieproportie te schatten met 95% betrouwbaarheid en een foutenmarge E , moet de steekproefomvang tenminste $n = \frac{1}{E^2}$ zijn.
 - d) Wanneer het heel belangrijk is om bij toetsing van een claim de nulhypothese niet ten onrechte te verwerpen, kun je beter een significantieniveau van 0.01, dan van 0.05 gebruiken.
2. Geef, indien van toepassing, voor de volgende twee situaties de modus, de mediaan, het gemiddelde en de standaardafwijking van de data.
 - a) Een experiment werd uitgevoerd om te onderzoeken of gebrek aan koolstofdioxide in de grond het fenotype van erwtenplanten beïnvloedt. Voor elk van de planten werd het fenotype genoteerd als volgt: 1 = glad geel, 2 = glad groen, 3 = gerimpeld geel, 4 = gerimpeld groen. De resultaten van 20 planten waren:
2 1 1 1 1 1 1 1 4 1 2 2 1 3 3 2 1 3 1 2.
 - b) De lengte van de eerste 10 woorden in een roman zijn 2 3 2 6 3 2 2 4 4 3.

In Figuur 1 staan boxplots van de tijd (in uren) besteed door een groep studenten aan twee opdrachten van een computerpraktikum. (De whiskers 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.)

- c) Geef de maten voor lokatie en schaal met hun waarden die je kunt aflezen uit deze twee boxplots.
- d) Schets voor beide data sets een histogram dat bij de data zou kunnen passen en becommentarieer en vergelijk de twee data sets wat betreft lokatie, schaal en vorm van hun verdelingen.



Figuur 1: *Boxplot van uren besteed aan opgaven computerpraktikum.*

3. Bij een groot bedrijf behoort het afnemen van een IQ-test tot de standaard sollicitatieprocedure. De IQ-scores van 25 van de sollicitanten waren

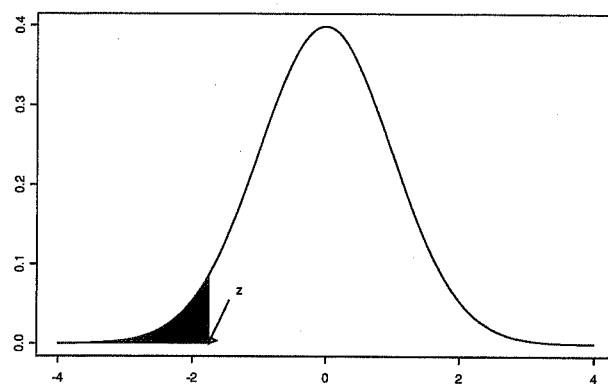
111	117	91	85	95
115	81	116	125	87
98	132	100	96	82
101	115	100	90	102
109	110	103	103	136.

- a) Wat zijn in dit geval de populatie, de steekproefgrootte en de populatieparameter?
- b) Bereken op basis van relatieve frequentie een schatting voor de kans dat de IQ-score van een sollicitant bij het bedrijf hoger is dan 115.

IQ-testen zijn zo geconstrueerd dat IQ-scores in de gehele bevolking normaal verdeeld zijn met verwachting (populatiegemiddelde) $\mu = 100$ en standaardafwijking $\sigma = 15$. Veronderstel dat dit ook geldt voor de IQ-scores van de sollicitanten bij het bedrijf.

- c) Geef op basis van deze veronderstelling een schatting voor de kans dat de IQ-score van een willekeurige sollicitant bij het bedrijf hoger is dan 115. Je kunt hiervoor bijgeleverde Tabel 5.1 gebruiken.
 - d) Geef, ook op basis van deze veronderstelling, de kans dat de gemiddelde IQ-score van 25 willekeurige sollicitanten bij het bedrijf lager is dan 95. Ook hiervoor kun je Tabel 5.1 gebruiken.
 - e) De schattingen van de kansen onder b) en c) verschillen nogal. Wat zou hiervan de oorzaak kunnen zijn?
4. Iemand gooit twee keer met een onzuiver muntstuk. De kans op kop is 0.3 en de kans op munt is 0.7.
(Bij onderstaande onderdelen mag je in je antwoord een breuk of produkt laten staan.)
- a) Construeer een kansruimte voor dit experiment.
- Als kop bovenkomt krijgt de persoon die gooit 1 euro, bij munt verliest ze 1 euro.
- b) Bepaal de kans dat het aantal verkregen euro's in de tweede worp *ongelijk* is aan dat van de eerste worp. Doe dit door netjes de gebeurtenis te formuleren binnen de in a) geconstrueerde kansruimte.
 - c) Beschouw de stochastische variabele X die het bij het experiment totaal verkregen bedrag beschrijft. Construeer de kansfunctie $p(x)$ van X netjes op basis van de formele definitie.
 - d) Bereken de verwachting EX van X ; geef niet alleen de uitkomst, maar laat zien hoe je hier aan komt.
5. 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 voor 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 produkt laten staan.)
- a) Wat is de kans dat drie willekeurige personen liegen?
 - b) Beschrijf het complement van bovenstaande gebeurtenis en geef de kans op dit complement.
 - c) Wat is de kans dat tenminste één van twee willekeurige personen liegt?
 - d) Wat is de kans dat de detector aangeeft dat een persoon liegt?

- e) Wat is de kans dat een persoon de waarheid spreekt, gegeven dat de detector aangeeft dat hij liegt?
6. In opgave 4 zijn de IQ-scores gegeven voor $n = 25$ sollicitanten van een bedrijf. Het gemiddelde van die scores is $\bar{x} = 104$ en de standaardafwijking $s_n = 14.68$. In deze opgave bepalen we op basis van de scores een betrouwbaarheidsinterval voor het gemiddelde μ van de IQ-scores van de sollicitanten van het bedrijf. Op basis van de Centrale Limietstelling weten we dat het gemiddelde van n scores 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 de gemiddelde IQ-score van de sollicitanten van het bedrijf gelijk is aan 100, d.w.z. hetzelfde als dat van de gehele bevolking? Licht je antwoord kort toe.
7. Een extreme vorm van angst kan leiden tot een fobie. In een aselechte steekproef van 100 Nederlanders gaven 10 mensen te kennen dat ze lijden aan een angstfobie. We willen op basis van deze steekproef de bewering toetsen dat 12% van de Nederlanders last heeft van een angstfobie.
- a) Stel de nulhypothese H_0 en de alternatieve hypothese H_a met betrekking tot de fractie p van Nederlanders met een angstfobie.
- b) Bereken de gebruikelijke schatting $\hat{p}$ voor p .
- c) Bereken de standaardscore z (toetsingsgrootte) voor $\hat{p}$.
- d) Bepaal de P -waarde behorend bij deze score. Je kunt hiervoor weer Tabel 5.1 gebruiken.
- e) Wat is de conclusie van de toets bij een significantieniveau van 1%?



Figuur 2: *Standaard normale curve corresponderend met Tabel 5.1.*

Rekenregels voor kansen bij het tentamen Toegepaste Statistiek

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

$A, B \in \mathcal{A}$ gebeurtenissen

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

Regel van 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:

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

Tabel 5.1 Standaard scores bij het tentamen Toegepaste Statistiek

Table 5.1 Standard Scores and Percentiles for a Normal Distribution (cumulative values from the left)					
Standard score	%	Standard score	%	Standard score	%
-3.5	0.02	-1.0	15.87	0.0	50.00
-3.0	0.13	-0.95	17.11	0.05	51.99
-2.9	0.19	-0.90	18.41	0.10	53.98
-2.8	0.26	-0.85	19.77	0.15	55.96
-2.7	0.35	-0.80	21.19	0.20	57.93
-2.6	0.47	-0.75	22.66	0.25	59.87
-2.5	0.62	-0.70	24.20	0.30	61.79
-2.4	0.82	-0.65	25.78	0.35	63.68
-2.3	1.07	-0.60	27.43	0.40	65.54
-2.2	1.39	-0.55	29.12	0.45	67.36
-2.1	1.79	-0.50	30.85	0.50	69.15
-2.0	2.28	-0.45	32.64	0.55	70.88
-1.9	2.87	-0.40	34.46	0.60	72.57
-1.8	3.59	-0.35	36.32	0.65	74.22
-1.7	4.46	-0.30	38.21	0.70	75.80
-1.6	5.48	-0.25	40.13	0.75	77.34
-1.5	6.68	-0.20	42.07	0.80	78.81
-1.4	8.08	-0.15	44.04	0.85	80.23
-1.3	9.68	-0.10	46.02	0.90	81.59
-1.2	11.51	-0.05	48.01	0.95	82.89
-1.1	13.57	0.0	50.00	1.0	84.13
				1.1	85.43
				1.2	86.49
				1.3	87.32
				1.4	88.49
				1.5	89.44
				1.6	90.32
				1.7	91.15
				1.8	91.92
				1.9	92.65
				2.0	93.32
				2.1	93.98
				2.2	94.52
				2.3	95.04
				2.4	95.54
				2.5	96.01
				2.6	96.47
				2.7	96.91
				2.8	97.33
				2.9	97.72
				3.0	98.08
				3.1	98.41
				3.2	98.71
				3.3	98.98
				3.4	99.18
				3.5	99.38
				3.6	99.53
				3.7	99.65
				3.8	99.74
				3.9	99.81
				4.0	99.87
				4.1	99.90
				4.2	99.93
				4.3	99.95
				4.4	99.97
				4.5	99.98

Note: The table shows percentiles for standard scores between -3.5 and +3.5, though much lower and higher standard scores are possible. (Appendix A has a more detailed standard score table.) The % column gives the percentage of values in the distribution less than the corresponding standard score.