
Tentamen Toegepaste Statistiek
voor studenten IMM en AI
Vrije Universiteit, Faculteit Exacte Wetenschappen
9 juni 2009, 18.30-20.30 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 zes opdrachten wegen even zwaar.

1. Zijn de volgende uitspraken zinvol (of correct)? Motiveer je antwoord.
 - a) Bij een verdeling die scheef naar rechts is (right-skewed), ligt de mediaan links van het steekproefgemiddelde.
 - b) Tussen IQ en schoenmaat is geen correlatie. Op grond van metingen van schoenmaat en IQ aan 30 personen is de vergelijking voor de best passende rechte lijn berekend tussen schoenmaat en IQ. Deze vergelijking werd gebruikt om van iemand met schoenmaat 42 het IQ te voorspellen.
 - c) Om een populatiegemiddelde te schatten met 95% betrouwbaarheid en een foutenmarge E , moet de steekproefomvang tenminste $n = \frac{1}{E^2}$ zijn.
 - d) Het gemiddelde van de verdeling van gemiddelden is bij benadering een normale verdeling als de steekproefomvang groot is. Dit geldt alleen als de steekproeven afkomstig zijn uit een normale verdeling.
2. Overgewicht is een belangrijke oorzaak voor diabetes. Volgens het voedingscentrum is iemand met een BMI-waarde hoger dan 30 veel te zwaar. Van 25 vrouwen met diabetes werd de BMI bepaald. De waarden waren

35.8	38.3	41.3	39.8	41.0
39.9	36.0	35.0	34.2	34.8
35.5	39.3	32.7	37.9	31.8
28.8	40.7	26.4	39.5	28.3
37.3	34.9	37.4	30.9	40.2

Het steekproefgemiddelde is $\bar{x} = 35.9$.

- a) Bereken op basis van relatieve frequentie een schatting voor de kans dat de BMI van een vrouw met diabetes hoger is dan 30.0.

Neem aan dat de BMI indices van vrouwen met diabetes normaal verdeeld zijn met verwachting (populatiegemiddelde) $\mu = 34.7$ en standaardafwijking $\sigma = 4.8$.

- b) Geef op basis van deze veronderstelling een schatting voor de kans dat de BMI van een vrouw met diabetes hoger is dan 30.0. Je kunt hiervoor bijgeleverde Tabel 5.1 gebruiken.
- c) Wat kan een reden zijn voor verschillende resultaten bij de onderdelen a) en b)?
- d) Geef, ook op basis van bovenstaande normaliteitsveronderstelling, de kans dat de gemiddelde BMI van 25 vrouwen met diabetes hoger is dan 35.9. Ook hiervoor kun je Tabel 5.1 gebruiken.

3. Een kaartspel heeft 10 rode, 10 gele, 10 blauwe en 10 groene kaarten. De kaarten van een kleur zijn genummerd 1 t/m 10. Iemand trekt willekeurig een kaart uit het kaartspel.

- a) Wat is de kans op de gebeurtenis dat de kaart rood of even is?
- b) Zij A de gebeurtenissen dat de kaart rood is en B de gebeurtenis dat de kaart even is. Zijn A en B onafhankelijk? Waarom (niet)?

Vervolgens wordt de getrokken kaart terug gelegd, worden de kaarten geschud en trekt de persoon opnieuw een kaart.

- c) Construeer een kansruimte voor het experiment van de twee trekkingen.

Bij trekking van een even kaart krijgt de persoon die trekt 2 euro, bij een oneven kaart krijgt de persoon niets.

- d) Beschouw de stochastische variabele X die het totaal verkregen bedrag bij het experiment van de twee trekkingen 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 een onderzoek naar verschuiving in beroepsgroepen tussen de generatie van vaders en die van hun zonen, werden beroepen ingedeeld in de niveaugroepen laag (L), midden (M) en hoog (H). Een index 'V' geeft de generatie van vaders aan; een index 'Z' die van de zonen. In de generatie van vaders had 50% een beroep in L_V , 40% in M_V en 10% in H_V . De volgende tabel geeft de voorwaardelijke kansen dat een zoon een beroep in een van de groepen heeft, gegeven de verschillende mogelijkheden voor het niveau van het beroep van de vader. Bijvoorbeeld, als een vader een hoog niveau beroep had, dan is de kans dat zijn zoon dit ook heeft $P(H_Z|H_V) = 0.45$; als een vader een midden niveau beroep had, dan is de kans dat zijn zoon een laag niveau beroep heeft $P(L_Z|M_V) = 0.25$, enz.

	L_Z	M_Z	H_Z
L_V	0.49	0.50	0.01
M_V	0.25	0.70	0.05
H_V	0.07	0.48	0.45

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.)

- Wat is de kans dat twee willekeurige vaders beiden een beroep op het midden niveau hadden?
- Wat is de kans dat tenminste één van drie willekeurige vaders een beroep op het midden niveau had?
- Wat is de kans dat een willekeurige zoon een beroep op hoog niveau heeft?
- Een willekeurige zoon heeft een beroep op hoog niveau. Wat is de kans dat zijn vader ook een beroep op hoog niveau had?

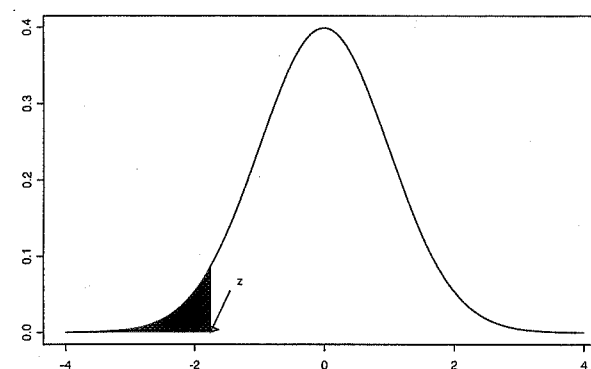
5. In een klinische studie werd het effect van twee medicijnen A en B vergeleken. Hiertoe werd aan 100 patiënten eerst een week het ene medicijn en daarna een week het andere medicijn gegeven. De volgorde van toediening was aselekt, door loting bepaald. Vervolgens werd aan elke patiënt gevraagd welk medicijn beter beviel. Van de 100 patiënten gaven 65 patiënten de voorkeur aan medicijn A en 35 aan medicijn B. Om te onderzoeken of deze uitslag een sterke indicatie is dat de meeste patiënten A beter vinden dan B, werd op basis van de studie de bewering getoetst dat de fractie van de patiënten die medicijn A beter vindt dan medicijn B groter is dan 0.5. Zij p de fractie van patiënten in de gehele populatie van patiënten die A beter vinden dan B.
- Stel de nulhypothese H_0 en de alternatieve hypothese H_a met betrekking tot de fractie p .
 - Bereken de gebruikelijke schatting $\hat{p}$ voor p .
 - Bereken de standaardscore z (toetsingsgrootheid) voor $\hat{p}$.
 - Bepaal de P -waarde behorend bij deze score. Je kunt hiervoor weer Tabel 5.1 gebruiken.
 - Wat is de conclusie van de toets bij een significantieniveau van 5%?
6. In onderdeel b) van de vorige opgave is een puntschatting $\hat{p}$ voor de populatieparameter p bepaald. In deze opgave bepalen we, gebaseerd op dezelfde data, een intervalschatting voor deze parameter. De Centrale Limietstelling is een stelling die in de huidige situatie het volgende zegt: bij een grote steekproefomvang n , is de grootheid

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}}$$

bij benadering standaard normaal verdeeld.

(In onderdeel b) van deze opgave mogen breuken en wortels in het antwoord blijven staan.)

- Leid op basis van de bijbehorende normale benadering de *algemene uitdrukking* af voor een 95% betrouwbaarheidsinterval voor p .
- Bepaal op basis hiervan het 95% betrouwbaarheidsinterval voor p uitgaande van de data van opgave 6, en licht de interpretatie van dat interval toe.
- Is het op grond van het in b) gevonden interval aannemelijk dat de meeste patiënten medicijn A beter vinden dan medicijn B? Licht je antwoord kort toe.



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 (Percentages are from the table)					
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	48.07	0.95	82.89
-1.1	13.57	0.0	50.00	1.0	84.13
				1.1	86.43
				1.2	88.49
				1.3	90.32
				1.4	91.92
				1.5	93.32
				1.6	94.52
				1.7	95.54
				1.8	96.41
				1.9	97.13
				2.0	97.72
				2.1	98.21
				2.2	98.61
				2.3	98.93
				2.4	99.18
				2.5	99.38
				2.6	99.53
				2.7	99.65
				2.8	99.74
				2.9	99.81
				3.0	99.87
				3.5	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.