

Tentamen Toegepaste Statistiek voor studenten Informatiekunde
vrije Universiteit, Faculteit Exacte Wetenschappen
9 juni 2005, 18.30-21.30u.

1.
 - a) In de Splts van 6 juni 2005 wordt melding gemaakt van een oproep van de Inspectie voor de Gezondheidszorg. Ziekenhuizen worden opgeroepen zeer terughoudend te zijn met het narcosemiddel lachgas in verband met eventuele gevaren voor ongeboren babies tijdens de zwangerschap. Reden voor deze oproep is dat in het Haagse ziekenhuis Leyenburg over een periode van tien jaar van negentien zwangere verpleegkundigen die mogelijk lachgas hebben ingeademd er zes een gehandicapt kind hebben gekregen. Geef aan welk type statistische studie dit is.
 - b) Geef met behulp van een voorbeeld aan wat systematisch steekproeftrekken is.
 - c) De exit poll voor het referendum over de Europese grondwet op 1 juni, 21.00 uur, gaf een percentage nee-stemmers dat heel dicht bij het uiteindelijke percentage lag. Geef, gebruikmakend van de aangeleerde terminologie, een methode aan die in deze situatie kan zijn gebruikt om tot een betrouwbaar resultaat te komen.
 - d) Leg aan de hand van een voorbeeld uit wat het begrip *verstrengeling* of *confounding* inhoudt.
 - e) Een autofabriek wil voor een bepaald type auto (Fiat Punto) nagaan welke van twee soorten katalysatoren het best werkt. Twintig exemplaren van dit type auto worden gebruikt in een onderzoek. Tien hiervan worden uitgerust met katalysator A en tien met katalysator B. Met de twintig wagens worden testritten gemaakt. Naar aanleiding van de gemeten prestaties van de katalysatoren moet worden besloten welke in de toekomst standaard zal worden ingebouwd. Geef aan met welk type statistische studie we hier te maken hebben.
2.
 - a) Bij de Rijn in Lobith is apparatuur geïnstalleerd waarmee continu gemeten wordt wat de concentratie van verschillende vervuilende stoffen is, in *parts per million*. Welk type data levert dit op en wat is de meetschaal?
 - b) Een krant wil overgaan van het standaardformaat naar een tabloid formaat. Hierbij zal ook een nieuwe layout worden gebruikt. Om na te gaan welke layout voor de lezers van deze specifieke krant het meest wenselijk is, wordt een groep aselect gekozen abonnees gevraagd de verschillende layouts te scoren op een schaal van *A* (heel goed) tot *F* (heel slecht) Welk type data levert dit op en wat is de meetschaal?

3. In een demografisch onderzoek in een bepaalde regio is voor een aselekt gekozen steekproef van gezinnen de omvang bepaald. Hieronder zijn de resultaten weergegeven.

1	2	2	1	2
5	4	3	2	3
5	2	3	4	1
2	2	3	1	3
6	4	6	1	2

- Geef een tabel met daarin de frequenties, de relatieve- en cumulatieve relatieve frequenties voor corresponderend met deze data.
 - Schets op basis hiervan een staafdiagram voor de data en beschrijf de vorm van de dataverdeling kwalitatief.
 - Geef de modus, de mediaan en het gemiddelde van de data.
4. Een fabrikant van chocolade doet onderzoek naar het preciese gewicht van een bepaald type chocoladereep die uit zijn productieproces komen. Van 25 aselekt gekozen repen uit de dagproductie wordt het gewicht (in gram) gemeten. Hieronder staan de resultaten van de metingen.

101.71	101.21	101.70	99.90	100.14
100.42	101.26	100.75	101.16	101.43
101.32	99.99	100.61	100.11	101.34
101.87	101.87	100.97	101.56	99.76
100.87	100.72	100.60	100.61	100.84

- Bereken op basis van relatieve frequenties een schatting voor de kans dat een willekeurige reep uit de dagproductie minder dan 100.00 gram weegt.

Uit eerder onderzoek is gebleken dat de populatieverdeling van het gewicht van een chocoladereep bij benadering normaal is met populatiegemiddelde $\mu = 101.00$ en standaardafwijking 0.50.

- Geef op basis van deze bewering de kans dat een willekeurige reep minder dan 100 gram weegt. Tabel 1 aan het eind van dit tentamen kan bij het berekenen van deze kans van nut zijn.
 - Geef op basis van deze bewering de kans dat het gemiddelde gewicht van 25 repen uit de dagproductie minder dan 100.90 gram is. Tabel 1 aan het eind van dit tentamen kan ook bij het berekenen van deze kans van nut zijn.
5. Geef aan welke visualisatie het meest geschikt zou zijn voor de volgende data. Motiveer uw antwoord.
- Een autofabrikant brengt al gedurende tien jaar vier types auto's op de markt. Hij is met name geïnteresseerd in het verloop (in de tijd) van het onderlinge marktaandeel van zijn producten.

- b) Bij een personeelswerfingsbureau komen op een bepaalde vacature 120 sollicitaties binnen. Alle sollicitanten worden ingedeeld in klassen 1 t/m 5, afhankelijk van geschiktheid. Kandidaten uit de eerste klasse zijn het meest geschikt en die uit klasse 5 het minst. Men wil in één oogopslag een indruk krijgen van de geschiktheid van de sollicitanten.
6. Beschouw een kleinschalig roulettespel waarbij het balletje in zes verschillende vakjes kan vallen. De (volledig vergelijkbare) vakjes zijn genummerd 1 t/m 5 en 'blanco' respectievelijk. Het balletje wordt losgelaten en de schijf draait. Uiteindelijk komt het balletje in een vakje terecht.
- a) Construeer een kansruimte voor dit experiment.
- Om het spel te spelen moeten we 1 euro betalen. We spelen en zetten in op vakje 2. Dit betekent dat we 5 euro ontvangen wanneer het balletje in vakje 2 terecht komt. Anders krijgen we niets.
- b) Definieer de stochastische variabele X die aan iedere uitkomst van het experiment de *winst* toekent die we opstrijken. Merk op dat 'winst' in dit geval ook negatief kan zijn.
- c) Construeer de kansfunctie van X netjes op basis van de formele definitie.
- d) Bereken de verwachting EX van X .
7. Men wil om 21.00 uur op de avond van 1 juni 2005 een schatting geven van het percentage ja-stemmers in het referendum over de Europese grondwet. Dit gebeurt op basis van een representatieve exit poll onder 100 mensen die hun stem hebben uitgebracht. Van deze 100 mensen gaven er 36 aan vóór te hebben gestemd.
- a) Geef de gebruikelijke *puntschatting* $\hat{p}$ voor het onbekende percentage p van ja-stemmers.

De Centrale Limietstelling is een stelling die in de huidige situatie het volgende zegt. Bij een grote steekproefomvang n , is de grootte

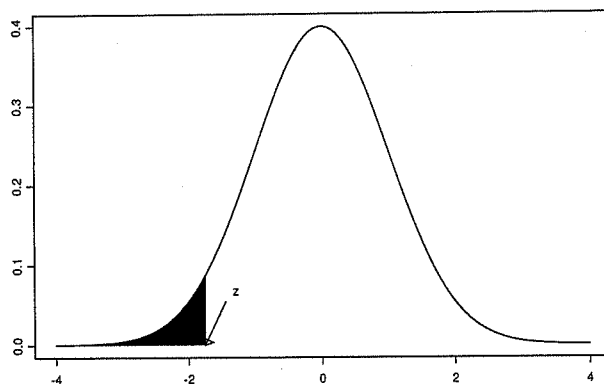
$$\sqrt{n} \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})}}$$

bij benadering standaard normaal verdeeld.

- b) Leid op basis van de bijbehorende normale benadering de *algemene uitdrukking* af voor een 95% betrouwbaarheidsinterval voor p .

In de onderdelen c) en d) van deze opgave mogen breuken en wortels in het antwoord blijven staan.

- c) Geef het 95% betrouwbaarheidsinterval voor p in de huidige situatie en licht de interpretatie van dat interval toe.
 - d) Hoeveel mensen zou men bij benadering moeten vragen naar hun stemgedrag om een 95% betrouwbaarheidsinterval ter breedte 0.02 te krijgen?
8. In opgave 4 is voor 25 aselekt gekozen chocoladerepen uit de dagproductie van een fabrikant het preciese gewicht gemeten. Het gemiddelde van die metingen is $\bar{x} = 100.87$ en de standaardafwijking 0.65. In deze opgave willen we op basis van die metingen toetsen of het gemiddelde gewicht van alle repen in de dagproductie 101.00 gram kan zijn, zoals uit eerder onderzoek zou zijn gebleken.
- a) Stel de nulhypothese H_0 en de alternatieve hypothese H_a op met betrekking tot het gemiddelde gewicht van de chocoladerepen uit de dagproductie.
 - b) Bereken de standaardscore z (toetsingsgrootheid) gebaseerd op de data.
 - c) Bereken de P -waarde behorend bij deze score (tabel 1 kan hierbij van nut zijn).
 - d) Wordt H_0 ten gunste van H_a verworpen bij een significantieniveau $\alpha = 0.05$? Licht uw antwoord toe.



Figuur 1: standaard normale curve corresponderend met tabel 1

z	Oppervlakte onder de curve <i>links van</i> z
-2.000	0.023
-1.960	0.025
-0.500	0.309
0.040	0.516
0.250	0.599
1.000	0.841
1.645	0.950
1.960	0.975

Tabel 1: oppervlakten onder de standaard normale curve corresponderend met figuur 1

1a: 3	1d: 3	2b: 3	3c: 4	4c: 3	6a: 3	6d: 3	7c: 4	8b: 4
1b: 3	1e: 3	3a: 6	4a: 3	5a: 3	6b: 3	7a: 3	7d: 3	8c: 4
1c: 3	2a: 3	3b: 4	4b: 3	5b: 3	6c: 3	7b: 3	8a: 4	8d: 3

Tabel 2: te verdienen punten per onderdeel

Cijfer=(aantal punten + 10)/10

Succes!