

Tentamen Toegepaste Statistiek voor studenten Informatiekunde
 vrije Universiteit, Faculteit Exacte Wetenschappen
 29 maart 2005, 9.30-12.30u.

1.
 - a) In een studie waarin honderden Zweedse tweelingen zijn betrokken, is vastgesteld dat de verstandelijke vermogens onder eenenijge tweelingen meer met elkaar overeenstemmen dan die van tweeeijge tweelingen. Welk type statistische studie is dit? Beargumenteer uw antwoord.
 - b) Geef met behulp van een voorbeeld aan wat gestratificeerd steekproeftrekken (stratified sampling) is.
 - c) Honderd mensen zijn aselekt geselecteerd uit een bepaalde doelgroep om het effect van 'wandелgewoonte' op de hartslag te bestuderen. Alle geselecteerden mogen kiezen tot welke groep ze zullen behoren: wel dagelijks een wandeling of niet. Aan het begin van het experiment wordt de rusthartslag van alle deelnemers gemeten, alsmede na afloop van de zes weken waarin het experiment wordt uitgevoerd. De verschillen in hartslagmeting worden vervolgens per groep bestudeerd om het effect van de wandelgewoonte op de rusthartslag na te gaan. Geef aan wat er bij dit onderzoek niet goed is en geef een alternatieve opzet waar u meer vertrouwen in heeft.
 - d) Leg aan de hand van een voorbeeld uit wat het begrip *verstremgeling* of *confounding* inhoudt.
 - e) Naar aanleiding van talloze studies die zijn gedaan naar het effect van een vroeggeboorte op de verdere ontwikkeling van het kind, wordt in een studie geconcludeerd dat vroeggeboorte een negatief effect heeft op die ontwikkeling. Geef aan welk type studie hier is gedaan en licht uw antwoord toe.
2.
 - a) Op een bepaalde plaats wordt het CO_2 -gehalte in de lucht dagelijks gemeten. Geef aan met welk type data we hier te maken hebben. Geef ook aan wat de meetschaal is.
 - b) In een call center worden dagelijks telefoontjes voor verschillende bedrijven afgehandeld. Over de periode van een jaar worden de telefoontjes ingedeeld naar het bedrijf waarvoor ze bedoeld zijn. Welk type data levert dit op en wat is de meetschaal?
3. In het kader van een onderzoek naar lengten van zestienjarigen is een steekproef van omvang 36 uit de populatie van zestienjarigen getrokken. Hieronder zijn de gegevens, uitgedrukt in centimeters, weergegeven.

163	180	170	164	170	164	185	162	167
167	159	171	161	168	178	164	168	173
175	195	158	162	171	159	164	180	182
160	158	158	181	160	165	174	169	158

- a) Geef een tabel met 4 kolommen. In de eerste de *klassen*, lopend van 155-160 tot en met 190-195; in de tweede de bijbehorende frequentie, de derde de relatieve frequentie en in de vierde kolom de cumulatieve relatieve frequentie.
 - b) Schets op basis hiervan een histogram voor de data en beschrijf de vorm van de dataverdeling kwalitatief.
 - c) Geef de modale klasse aan en bereken de mediaan van de data.
4. Geef aan welke visualisatie het meest geschikt zou zijn voor de volgende data. Motiveer uw antwoord.
- a) Een ijsproducent brengt al gedurende tien jaar vier smaken ijs op de markt. Hij is met name geïnteresseerd in het verloop (in de tijd) van het onderlinge marktaandeel van zijn smaakjes.
 - b) Van 1.5 V batterijen zijn er zo'n 40 verschillende merken. Deze worden in verschillende landen in verschillende hoeveelheden verkocht. De brancheorganisatie heeft informatie over deze aantallen voor de landen Nederland en België, en is met name ook geïnteresseerd in de verschillen in verkopen.
 - c) Bij een natuursteenhandel komt op een bepaalde dag een nieuwe partij natuursteen binnen. Alle stenen worden ingedeeld in klassen, genummerd 1 t/m 5, afhankelijk van 'schoonheid'. Stenen uit klasse 1 zijn het mooist en die uit klasse 5 het minst mooi. Men wil in één oogopslag een indruk krijgen van de schoonheid van de partij stenen.
5. In de VS wordt op scholen gebruikgemaakt van een zogenaamde SAT score voor scholieren. Dit is een score die vergelijkbaar is met onze CITO score. Hieronder staan SAT scores van 25 aselect gekozen Amerikaanse scholieren.

510	686	601	388	504
366	473	483	510	424
553	503	455	533	314
528	677	673	568	251
411	451	611	585	527

- a) Bereken op basis van relatieve frequenties een schatting voor de kans dat een willekeurige scholier in de VS een score heeft die lager is dan 450.

De SAT score is een gestandaardiseerde score waarvoor geldt die een normale verdeling heeft met populatiegemiddelde $\mu = 500$ en standaardafwijking 100.

- b) Geef op basis van deze bewering de kans dat de score van een willekeurige scholier lager is dan 450. Tabel 1 aan het eind van dit tentamen kan bij het berekenen van deze kans van nut zijn.

6. We gooien twee keer met een zuivere munt. De afzonderlijke uitkomsten worden gecodeerd met -1 en +1. Denk aan een spelletje waarbij iemand die kop gooit 1 punt krijgt en iemand die munt gooit 1 punt inlevert.
- Construeer een kansruimte voor dit experiment.
 - Bereken de kans dat de score in de tweede worp minstens zo gunstig is als in de eerste worp. Doe dit door netjes de gebeurtenis te formuleren binnen de in a) geconstrueerde kansruimte.
 - Beschouw de stochastische variabele X die de puntenopbrengst beschrijft bij het experiment. Construeer de kansfunctie van X netjes op basis van de formele definitie.
 - Bereken de verwachting EX van X .
 - Maak de gebeurtenissen die worden omschreven als 'eerste worp is +1' en 'tweede worp is -1' expliciet binnen de kansruimte en laat (volgens de formele definitie van onafhankelijkheid) zien dat deze gebeurtenissen *onafhankelijk* zijn.
7. De STER is geïnteresseerd in het percentage Nederlanders dat geregeld (gemiddeld vaker dan drie keer per week) het reclameblok voor het acht uur journaal bekijkt. Om hierover informatie te krijgen, wordt onderzoek onder $n = 500$ aselekt gekozen Nederlanders gehouden. Van deze 500 personen bleken er 55 geregeld dit reclameblok te bekijken.
- Geef de gebruikelijke *puntschatting* $\hat{p}$ voor het onbekende percentage p Nederlanders dat geregeld dit blok bekijkt.

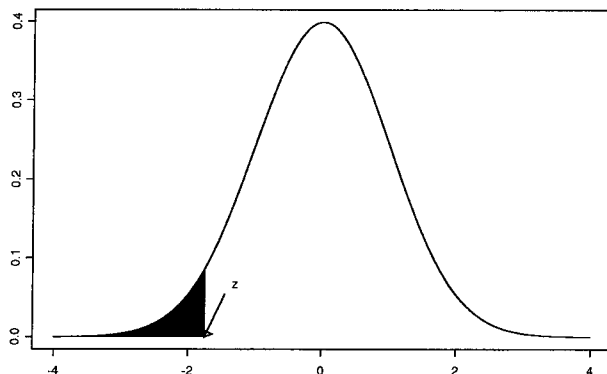
De Centrale Limietstelling is een stelling die in de huidige situatie het volgende zegt. Bij een grote steekproefomvang n , is de grootheid

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

bij benadering standaard normaal verdeeld.

- Leid op basis van deze uitspraak de *algemene uitdrukking* af voor een 95% betrouwbaarheidsinterval voor p .
 - Geef het 95% betrouwbaarheidsinterval voor p in de huidige situatie en licht de interpretatie van dat interval toe.
8. In opgave 5 zijn de SAT scores van 25 scholieren gegeven. Op basis hiervan willen we de claim onderzoeken dat de SAT score een gestandaardiseerde score is met populatiegemiddelde 500. Gegeven is dat het gemiddelde van de steekproef 504 is en de standaardafwijking 100.
- Stel de nulhypothese H_0 en de alternatieve hypothese H_a op met betrekking tot de gemiddelde SAT score van alle scholieren in de VS.

- b) Bereken de standaardscore z (toetsingsgrootte) gebaseerd op de data.
- c) Bereken de P -waarde behorend bij deze score (tabel 1 kan hierbij van nut zijn).
- d) Geef aan wat er gebeurt als er in de huidige context een *fout van type 2* wordt gemaakt.



Figuur 1: standaard normale curve corresponderend met tabel 1

z	Oppervlakte onder de curve <i>links van</i> z
-2.932	0.002
-1.960	0.025
-0.500	0.309
0.040	0.516
0.250	0.599
1.188	0.883
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: 3	4c: 3	6a: 3	6d: 3	7b: 3	8b: 3
1b: 3	1e: 3	3a: 6	4a: 3	5a: 3	6b: 3	6e: 3	7c: 3	8c: 6
1c: 3	2a: 3	3b: 3	4b: 3	5b: 3	6c: 6	7a: 3	8a: 3	8d: 3

Tabel 2: te verdienen punten per onderdeel

$$\text{Cijfer} = (\text{aantal punten} + 10) / 10$$

Succes!