

Tentamen Toegepaste Statistiek

voor studenten Informatiekunde en Kunstmatige Intelligentie
vrije Universiteit, Faculteit Exacte Wetenschappen
8 juni 2006, 18.30-20.30u.

1. a) Een farmaceutisch bedrijf denkt een verbeterde formule voor een anti-hooikoorts pil te hebben gevonden. Om te onderzoeken of de pil inderdaad beter werkt, worden twee (zo gelijk mogelijke) groepen van hooikoortspatiënten geselecteerd. De ene groep werkt met de oude pil, de andere met de nieuwe. Na de zomer worden de deelnemers geïnterviewd om te bepalen of de nieuwe pil inderdaad beter werkt dan de oorspronkelijke. Welk type statistische studie is dit? Beargumenteer uw antwoord.
b) Geef met behulp van een voorbeeld aan wat cluster steekproeftrekken is.
c) Bij een tv-programma worden kijkers opgeroepen telefonisch hun mening te geven over een politiek onderwerp, om de mening van 'het Nederlandse volk' te peilen. Geef, gebruikmakend van de geleerde terminologie, aan welke gevaren er aan dit type onderzoek kleven.
d) Let aan de hand van een voorbeeld uit wat 'confounding' ('verstrengeling') is.
e) De tijdens het college behandelde kentallen voor locatie hebben de eigenschap dat deze locatie-equivariant zijn. Leg uit wat hiermee wordt bedoeld.
2. a) Een autofabrikant is geïnteresseerd in de kleur-voorkeuren van autobezitters in een bepaald land. Studenten wordt gevraagd gedurende vijf dagen langs een snelweg en in een binnenstad te administreren welke kleur (25 mogelijkheden) auto's voorbij komen. Welk type data levert dit op en wat is de meetschaal?
b) Vanuit de ruimte wordt de gemiddelde dikte van de ozonlaag op de zuidpool continu gemeten. Geef aan met welk type data we hier te maken hebben. Geef ook aan wat de meetschaal is.
3. In het kader van een onderzoek naar gewichten van konijnen in een natuurreservaat is aselect een steekproef van omvang 36 uit de populatie getrokken. Hieronder zijn de gewichten, uitgedrukt in kilogram, weergegeven.

0.43	0.47	0.53	0.68	0.68	0.70	0.72	0.76	0.77
0.81	0.83	0.84	0.84	0.85	0.91	0.92	0.92	0.93
0.95	0.96	0.98	0.98	0.99	0.99	1.00	1.00	1.00
1.03	1.04	1.07	1.07	1.10	1.11	1.12	1.12	1.14

- a) Geef een tabel met 4 kolommen. In de eerste de *klassen*, oplopend van $[0.40, 0.50)$ tot en met $[1.10, 1.20)$; 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.
- Voor statistische analyses zijn verscheidene pakketten bruikbaar. Een software-ontwikkelaar heeft onderzoek verricht naar het gebruik van de pakketten R, Splus, SPSS, Minitab, Excel en 'overig' ten behoeve van statistische studies, en wil de resultaten inzichtelijk maken.
 - Een oliemaatschappij wil weten hoe het totale jaarlijkse benzinegebruik in België zich vanaf 1960 heeft ontwikkeld.
 - Drie kandidaten streden om het lijsttrekkerschap bij de VVD. De uitkomst van de verkiezing wil men presenteren op een manier zodat ook kan worden gezien hoe de stemverhoudingen lagen onder de mannelijke, resp. vrouwelijke stemmers.
5. Een fabrikant van cola doet onderzoek naar de preciese hoeveelheid cola die in de blikjes zit die uit zijn productie komen. Van 25 aselekt gekozen blikjes uit de dagproductie wordt de inhoud (in *cl.*) gemeten. Hieronder staan de resultaten van de metingen.

33.25	33.16	33.42	33.16	33.34
33.31	33.30	33.21	33.52	33.17
32.98	32.91	33.50	33.23	33.53
33.35	33.20	33.09	33.18	33.17
33.27	33.28	33.47	33.32	32.94

- Bereken op basis van relatieve frequenties een schatting voor de kans dat een willekeurig blikje uit de dagproductie minder dan 33.00 *cl.* cola bevat.

Uit eerder onderzoek is gebleken dat de populatieverdeling van de inhoud van een colablikje bij benadering normaal is met populatiegemiddelde $\mu = 33.20$ en standaardafwijking 0.20.

- Geef op basis van deze bewering de kans dat een willekeurig blikje minder dan 33.00 *cl.* cola bevat. 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 de gemiddelde inhoud van 25 blikjes uit de dagproductie lager dan 33.10 *cl.* ligt. Tabel 1 aan het eind van dit tentamen kan ook bij het berekenen van deze kans van nut zijn.
6. In een fabriek produceren machines 1 t/m 3 respectievelijk 30%, 45% en 25% van het product. Van de machines is bekend dat deze respectievelijk 2%, 3% en 2% defecten produceren. Een exemplaar wordt aselekt uit de gehele dagproductie gekozen. Dit exemplaar blijkt defect te zijn.
- Formuleer de regel van Bayes in het algemeen.

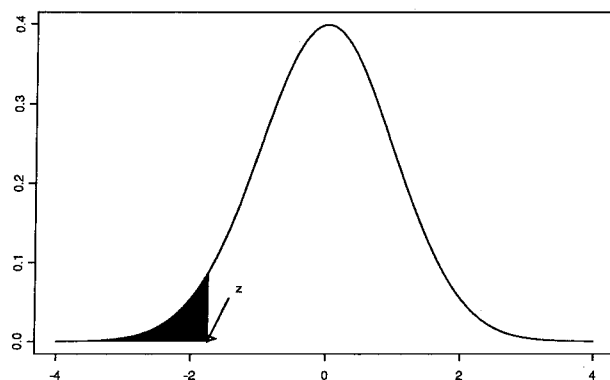
- b) Bereken (bij voorkeur gebruikmakend van onder a) genoemde regel) de kans dat het gekozen exemplaar door machine 1 is geproduceerd. Het antwoord mag niet-uitgewerkte breuken bevatten.
7. Een radioprogramma doet onderzoek naar het percentage huisvrouwen dat van plan is de voetbalwedstrijd Nederland-Ivoorkust op TV te gaan bekijken. Om hierover informatie te krijgen, wordt onderzoek onder $n = 100$ aselekt gekozen huisvrouwen gehouden. Van deze 100 huisvrouwen blijken er 36 van plan te zijn genoemde wedstrijd te gaan bekijken.
- a) Geef de gebruikelijke *puntschatting* $\hat{p}$ voor het onbekende percentage p van huisvrouwen dat de wedstrijd op de TV wil gaan kijken.

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.

- b) Geef op basis hiervan het 95% betrouwbaarheidsinterval voor p in de huidige situatie en licht de interpretatie van dat interval toe. (In het antwoord mogen wortels en producten onuitgewerkt blijven staan.)
8. In opgave 4 is voor 25 aselekt gekozen colablikjes uit de dagproductie van een fabrikant de preciese inhoud gemeten. Het gemiddelde van die metingen is $\bar{x} = 33.26$ en de standaardafwijking 0.16. In deze opgave willen we op basis van die metingen toetsen of de gemiddelde inhoud van alle blikjes in de dagproductie 33.20 cl kan zijn, zoals uit eerder onderzoek zou blijken.
- a) Stel de nulhypothese H_0 en de alternatieve hypothese H_a op met betrekking tot de gemiddelde inhoud van de colablikjes uit de dagproductie.
- b) Bereken de standardscore 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
-2.500	0.006
-1.960	0.025
-1.000	0.159
-0.500	0.309
0.250	0.599
0.375	0.646
1.188	0.883
1.875	0.970
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	5c: 3	7a: 3	8b: 3
1b: 3	1e: 3	3a: 4	4a: 3	5a: 3	6a: 2	7b: 3	8c: 3
1c: 3	2a: 3	3b: 3	4b: 3	5b: 3	6b: 3	8a: 3	8d: 3

Tabel 2: *te verdienen punten per onderdeel*

$$\text{Cijfer} = (\text{aantal punten} + 8) / 8$$

Dit is een gesloten boek tentamen. Een rekenmachine is niet toegestaan.

Succes!