

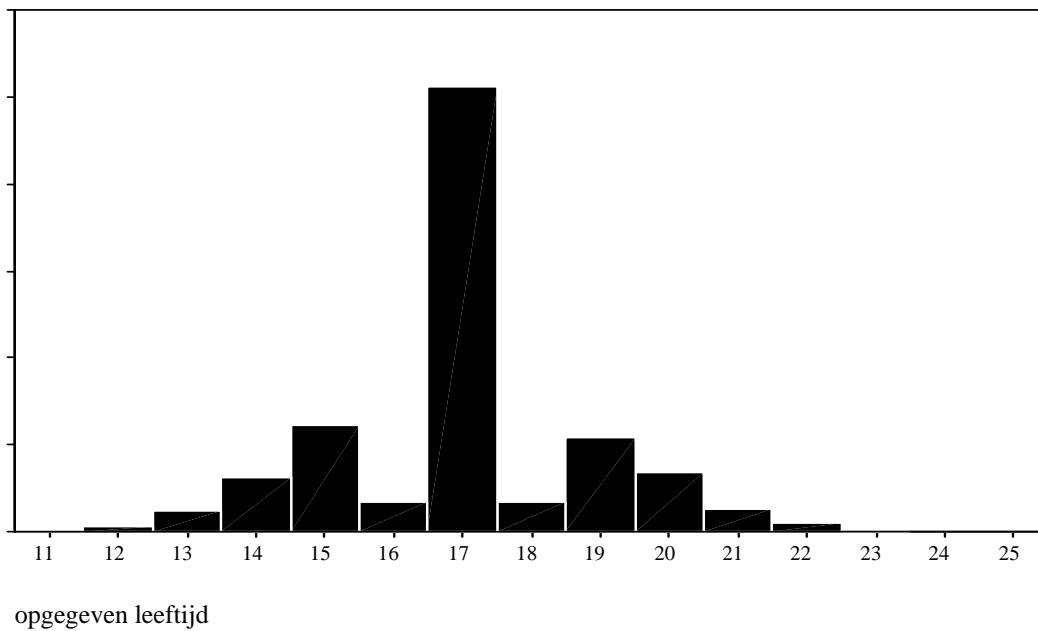
In dit tentamen gaan een aantal vragen over de (opgegeven) leeftijd van asielzoekers. Indien een alleenstaande asielzoeker minderjarig is (een zogenaamde AMA), gelden er andere regels met betrekking tot het verkrijgen van een voorlopige verblijfsvergunning dan indien hij of zij 18 jaar of ouder is. De regels voor asielzoekers van 18 jaar of ouder (meerderjarige asielzoekers) zijn veel strenger, tenminste dat waren ze tot het najaar van vorig jaar. Tegenwoordig wordt 80% van de AMA's na één maand het land weer uitgezet. Maar nog steeds schijnt het beleid met betrekking tot AMA's 'vriendelijker' te zijn dan met betrekking tot meerderjarige asielzoekers. Aangezien dit een tentamen statistiek is, zal ik het hier verder bij laten en overgaan aan de statistische kwesties.

Het grootste probleem bij het vaststellen of een asielzoeker meerder- dan wel minderjarig is, is het feit dat de meeste asielzoekers geen paspoort (meer) hebben. Je zou natuurlijk de leeftijd kunnen vragen aan de asielzoeker, maar men is van mening dat dat niet helemaal betrouwbaar is (en je kunt dat de asielzoeker in kwestie ook niet kwalijk nemen, me dunkt). Vandaar dat er enige tijd geleden nogal wat te doen was over een methode om aan de hand van röntgenfoto's de leeftijd van een persoon te bepalen. De personen die deze techniek presenteerden claimden dat er nu een wetenschappelijke en dus (?) betrouwbare manier was om iemands leeftijd te bepalen. Het tentamen is gebaseerd op dit verhaal. De data die gepresenteerd worden zijn niet echt, maar de (statistische) strekking van het verhaal is dat wel.

1.

De opgegeven leeftijden van 8 willekeurig gekozen alleenstaande asielzoekers waren: 14, 15, 16, 17, 17, 17, 19, 21.

- a* (6 punten) Bereken het gemiddelde, de mediaan en het 25% getrimd gemiddelde van deze data.
- b* (4 punten) Stel dat onderstaand histogram de verdeling van de opgegeven leeftijd van zeer veel asielzoekers geeft. Is het gemiddelde dan een goede maat voor locatie in een willekeurige steekproef van 8 personen die uit deze populatie getrokken wordt? Motiveer je antwoord.



Figuur 1. Histogram bij opgave 1.b.

De methode om de leeftijd te bepalen is onder andere gebaseerd op de botdichtheid: hoe groter de botdichtheid, hoe ouder. Tijdens dit tentamen nemen we aan dat botdichtheid gemeten wordt op een arbitraire schaal waarvoor geldt: hoe hoger de score, hoe groter de botdichtheid.

2.

Stel dat de botdichtheid bij 17-jarige jongens/mannen normaal verdeeld is met als gemiddelde 10 en als standaarddeviatie 2.

- a (3 punten) Wat is de kans dat van een willekeurig gekozen 17-jarige jongen/man de botdichtheid groter is dan 12?
- b (3 punten) Wat is de kans dat van een willekeurig gekozen steekproef van 4 personen uit de populatie 17-jarige jongens/mannen de gemiddelde botdichtheid groter is dan 12?
- c (4 punten) Stel dat wel een populatie mannen/jongens hebben die voor 60% uit 17-jarigen bestaat en voor 40% uit 18-jarigen. Voor de 17-jarigen geldt dat de kans op een botdichtheid van meer dan 10 gelijk is aan 0.5. Stel dat voor de 18-jarigen geldt dat de kans op een botdichtheid van meer dan 10 gelijk is aan 0.75. We kiezen uit deze populatie een willekeurig persoon en bepalen zijn botdichtheid. Deze blijkt meer dan 10 te zijn. Wat is de kans dat deze persoon 18 jaar is?

3. Indien je een meetinstrument voor leeftijd wilt baseren op iets als botdichtheid dan is het nodig om te weten of de botdichtheid van jongens/mannen en meisjes/vrouwen gelijk aan elkaar is (op eenzelfde leeftijd natuurlijk). Stel dat de botdichtheid van een meisje van 17 doorgaans lager is dan die van een jongen van 17 en je zou in je formule voor de voorspelde leeftijd geen rekening met geslacht houden dan zul je jongens altijd als ouder inschatten en/of meisjes als jonger (immers hoe hoger de leeftijd, hoe groter de botdichtheid) en als jongens juist een lagere botdichtheid hebben dan gebeurt het omgekeerde. Om een beeld te krijgen over mogelijke verschillen tussen jongens en meisjes met betrekking tot de botdichtheid heeft een onderzoeker van 7 willekeurig gekozen jongens en van 7 willekeurig gekozen meisjes, allen van 17 jaar, de botdichtheid bepaald. De gegevens staan in onderstaande tabel en tevens is er al wat uitgerekend.

	Botdichtheid jongens	Botdichtheid meisjes
	7.68	5.43
	8.26	5.95
	9.32	8.11
	9.37	9.03
	9.63	10.17
	12.85	10.47
	13.27	11.03
Gemiddelde	10.05	8.6
sd	2.17	2.21

gepoolde standaarddeviatie  $s_p = 2.19$ . Bovendien heeft ze verschillen berekend ('jongen' – 'meisje'), aangegeven met 'd'. Hiervoor heeft ze ook wat gegevens berekend:  $\bar{d} = 1.46$ ;  $s_d = 1.16$ .

- a* (15 punten) Kun je op basis van deze gegevens een uitspraak doen over verschillen in botdichtheid tussen jongens en meisjes? Gebruik  $\alpha = 0.05$ .
- b* (5 punten) Stel dat de onderzoeker zich niet beperkt had tot 17-jarigen maar een steekproef uit de populatie jongens/mannen onder de 25 en een steekproef uit de populatie meisjes/vrouwen onder de 25 genomen zou hebben. Wat zou het effect zijn op de *p*-waarde of de breedte van het betrouwbaarheidsinterval dat je bij *a*. berekend hebt?

4. (10 punten) Momenteel vindt een beoordeling over het al dan niet minderjarig zijn plaats op basis van een visuele inspectie door getrainde beoordelaars. Er kunnen bij zo'n inspectie allerlei beoordelingsfouten gemaakt worden. Zo kan het zijn dat er systematische verschillen tussen de beoordeling van b.v. Kongolese en Chinese personen bestaan. Om dit te onderzoeken werden aan een beoordelaar 12 paren van foto's van telkens één Kongolese

jongen en één Chinese jongen voorgelegd. De beoordelaar moest aangeven welke volgens hem de oudste was. De foto's waren echter van personen waarvan de leeftijd bekend was en deze personen waren steeds even oud! In totaal waren er dus foto's van 12 Kongolese jongens en van 12 Chinese jongens.

De beoordelaar gaf in 9 van de 12 keer aan de volgens hem de Kongolese jongen het oudste was en slechts in 3 van de 12 keer aan de volgens hem de Chinese jongen het oudste was (nergens gaf hij als antwoord dat ze volgens hem even oud waren). Toets de hypothese dat er sprake is van een oneerlijke beoordeling. Gebruik een  $\alpha$  van 0.10!

5.

Je zou bij bovenstaand onderzoek kunnen opmerken dat het niet de taak is van deze beoordelaars om personen te vergelijken, maar om op een correcte wijze aan te geven of een persoon minder- dan wel meerderjarig is. Daarom werd aan een beoordelaar gevraagd om 100 personen te beoordelen. Van deze personen (waarvan de leeftijd bekend waren) was 50 minderjarig en 50 meerderjarig. De gegevens staan in onderstaande tabel

Oordeel	Echte leeftijd		Totaal
	Minderjarig	Meerderjarig	
“Minderjarig”	30	20	50
“Meerderjarig”	20	30	50
Totaal	50	50	100

- a. (10 punten) Toets of er sprake is van samenhang de echte leeftijd en het oordeel van de beoordelaar. Gebruik  $\alpha = 0.05$ .
- b. (5 punten) Stel dat je bij a. tot de conclusie komt dat er sprake is van samenhang tussen de echte leeftijd en het oordeel; zou je dan op basis van deze gegevens kunnen zeggen dat de beoordelaar zijn werk goed doet?

6.

We gaan weer even terug naar het voorgestelde meetinstrument voor het bepalen van de leeftijd. Aan de hand van een röntgenfoto van het sleutelbeen worden een aantal karakteristieken gemeten en deze resultaten worden in een formule gevoerd waar dan weer een voorspelde leeftijd uit komt. In de case wordt nader ingegaan op hoe deze formule bepaald is. Om de kwaliteit van dit instrument te onderzoeken werd bij een groep van 40 personen (variërend in leeftijd van 12 tot 23 jaar) waarvan de echte leeftijd bekend was de

leeftijd met dit instrument bepaald. Met  $Y$  geven we de echte leeftijd weer (aantal geleefde dagen / 365.25) en met  $X$  geven we de op basis van het instrument bepaalde leeftijd weer. Met  $x$  (resp.  $y$ ) worden de deviatiescores van  $X$  (resp.  $Y$ ) aangegeven:  $\Sigma x^2 = 296.31$ ;  $\Sigma y^2 = 276.5$ ;  $\Sigma xy = 268.96$ ;  $\bar{x} = 17.2$ ;  $\bar{y} = 17.4$ ; residuele standaarddeviatie ( $s$ ) = 0.956.

- (2 punten) Bereken de waarde van de (Pearson product moment) correlatiecoëfficiënt.
- (3 punten) Stel de regressievergelijking voor de voorspelling van echte leeftijd met op basis van het instrument bepaalde leeftijd op.
- (2 punten) Wat is de voorspelde echte leeftijd voor een persoon waarvoor met het meetinstrument een leeftijd van 19.5 bepaald wordt?
- (3 punten) Indien met het instrument een leeftijd van 19.5 bepaald wordt, kunnen we dan met 95% zekerheid stellen dat deze persoon niet minderjarig is?

### CASE

In de onderstaande output worden de met SPSS berekende resultaten van de analyse van de gegevens van een onderzoek gepresenteerd. In dit onderzoek werd gepoogd om een formule op te stellen om op basis van een aantal op een röntgenfoto van het sleutelbeen te bepalen variabelen een voorspelling te geven van de echte leeftijd van een persoon. In dit onderzoek waren 100 personen betrokken. De variabelen die gebruikt werden om de echte leeftijd (aangeduid met LEEFTYD) te voorspellen waren: botdichtheid bepaald met methode Xeno (BOTX), botdichtheid bepaald met methode Yolande (BOTY), percentage kraakbeen op een bepaalde locatie (KRAAK), botdikte op een bepaalde plaats (DIK) en het aantal ringen in het bot (RING).

Voor de eerste analyse probeerde de onderzoeker de leeftijd te voorspellen op basis van beide botdichtheden. De output volgt hieronder:

### Regression

#### Descriptive Statistics

	Mean	Std. Deviation	N
LEEFTYD	16.6165	3.34305	100
BOTX	9.7856	2.08920	100
BOTY	9.7939	2.08427	100

**Correlations**

		LEEFTYD	BOTX	BOTY
Pearson Correlation	LEEFTYD	1.000	.845	.847
	BOTX	.845	1.000	.999
	BOTY	.847	.999	1.000
Sig. (1-tailed)	LEEFTYD	.	.000	.000
	BOTX	.000	.	.000
	BOTY	.000	.000	.
N	LEEFTYD	100	100	100
	BOTX	100	100	100
	BOTY	100	100	100

**Variables Entered/Removed<sup>d</sup>**

Model	Variables Entered	Variables Removed	Method
1	BOTY <sup>a</sup> BOTX	.	Enter

a. All requested variables entered.

b. Dependent Variable: LEEFTYD

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.848 <sup>a</sup>	.719	.713	1.79009

a. Predictors: (Constant), BOTY, BOTX

**ANOVA<sup>b</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	795.594	2	397.797	124.139	.000 <sup>a</sup>
	Residual	310.830	97	3.204		
	Total	1106.424	99			

a. Predictors: (Constant), BOTY, BOTX

b. Dependent Variable: LEEFTYD

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.276	.865		3.786	.000
	BOTX	-1.394	2.044	-.871	-.682	.497
	BOTY	2.755	2.049	1.717	1.345	.182

a. Dependent Variable: LEEFTYD

- a. (5 punten) Hoe hoog is de samenhang tussen BOTX en LEEFTYD en hoe hoog is de samenhang tussen BOTY en LEEFTYD? Zijn deze significant afwijkend van 0 ( $\alpha = 0.05$ )?

b. (3 punten) Zijn de regressiegewichten voor BOTX en BOTY significant afwijkend van 0 ( $\alpha = 0.05$ )?

c. (2 punten) Verklaar het verschil of de overeenkomst tussen de uitkomsten van a. en b.

Vervolgens is in een stapsgewijze procedure (een zogenaamde BACKWARD) procedure een regressiemodel bepaald dat met weinig prediktoren toch een goede voorspelling van de echte leeftijd kan geven. Als prediktoren werden BOTY, KRAAK, DIK en RING gebruikt. Relevante stukken van de output volgen:

#### Descriptive Statistics

	Mean	Std. Deviation	N
LEEFTYD	16.6165	3.34305	100
BOTY	9.7939	2.08427	100
KRAAK	49.1471	15.42481	100
DIK	20.1370	5.25206	100
RING	4.0000	1.02494	100

#### Model Summary<sup>c</sup>

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.933 <sup>a</sup>	.870	.864	1.23157
2	.933 <sup>b</sup>	.870	.866	1.22551

a. Predictors: (Constant), RING, DIK, BOTY, KRAAK

b. Predictors: (Constant), RING, BOTY, KRAAK

c. Dependent Variable: LEEFTYD

#### Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.699	.811		5.795	.000
	BOTY	1.413	.064	.881	22.031	.000
	KRAAK	6.898E-02	.009	.318	7.291	.000
	DIK	-6.02E-03	.025	-.009	-.242	.809
	RING	-1.297	.137	-.398	-9.499	.000
2	(Constant)	4.593	.681		6.748	.000
	BOTY	1.414	.064	.882	22.238	.000
	KRAAK	6.826E-02	.009	.315	7.638	.000
	RING	-1.295	.136	-.397	-9.546	.000

a. Dependent Variable: LEEFTYD

d. (5 punten) Stel de regressievergelijking voor de optimale formule voor de voorspelling van LEEFTYD op.

e. (3 punten) Hoe hoog is de multiple correlatie voor de voorspelling van LEEFTYD?

f. (2 punten) Wat is de beste voorspelling voor het percentage onverklaarde variantie?