

# Natuurwetenschappen en Innovatiemanagement

## Uitwerkingen tentamen Statistiek 1

1.

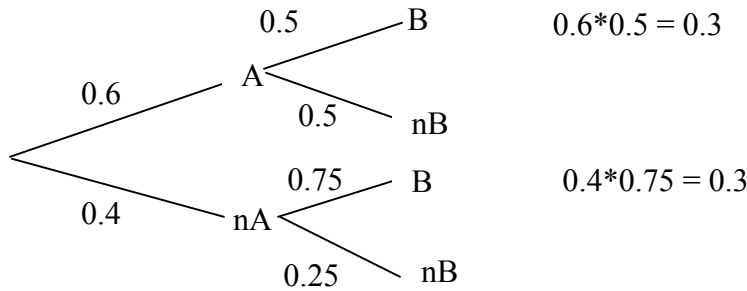
De opgegeven leeftijden van 8 willekeurig gekozen alleenstaande asielzoekers waren: 14, 15, 16, 17, 17, 17, 19, 21.

- a. Het gemiddelde: alles optellen en delen door 8 ( $= n$ ) levert  $136/8 = 17$ . De mediaan is in dit geval het gemiddelde van het 4<sup>de</sup> en het 5<sup>de</sup> getal (17 en 17): 17. Voor het 25% getrimd gemiddelde moeten we eerst aan beide zijden 25% ( $= 25\%$  van  $8 = 2$ ) van de observaties verwijderen. Resteert het gemiddelde van 16, 17, 17, 17 en dit is 16.75
- b. Op zich is het gemiddelde wel een aardige maat voor een steekproef van 8 personen uit deze populatie. De verdeling is in ieder geval symmetrisch en er lijken geen grote uitschieters te zijn. Toch heeft het 25% getrimd gemiddelde de voorkeur. De ‘hobbels’ bij 19 (en hoger) en 15 (en lager) zijn een indicatie dat deze, toch wat ‘extreme’ waardes gezien de grote piek bij 17, regelmatig zullen voorkomen en bij een kleine steekproefomvang is het niet gezegd dat de (te) grote en de (te) kleine waarden tegen elkaar weg zullen vallen. Dit zal blijken uit het feit dat de standaarderror voor het gemiddelde groter is dan die voor het getrimde gemiddelde. NB robuuste statistieken als het getrimd gemiddelde worden eigenlijk te weinig gebruikt (voornamelijk omdat de statistiek nodig voor het opstellen van b.v. betrouwbaarheidsintervallen wat ingewikkelder is). Overigens zal ik er geen punten aftrekken wanneer je voor het gemiddelde gaat (met een goede motivatie!!!), want dat zou 99% van de onderzoekers zeggen, maar ik kan het niet nalaten om jullie nogmaals op het nuttige van robuuste statistieken te wijzen (geef je het antwoord ‘liever het 25% getrimd gemiddelde’ met een goede motivatie dan krijg je bonuspunten)

2

- a. Botdichtheid geven we aan met  $X$ :  $X \sim \text{Normaal}(\mu = 10, \sigma = 2)$ , bij 17-jarige jongens. Gevraagd wordt:  $P(X > 12)$ . Standaardiseren levert:  $P(Z > (10 - 12)/ 2) = P(Z > 1)$  en tabel IV geeft dan:  $P(X > 12) = 0.159$

- b. De standaarddeviatie van het gemiddelde ( $\bar{X}$ ) van de botdichtheid van 4 17-jarige jongens ( $SE(\bar{X}) = \sigma/\sqrt{n} = 2/\sqrt{4} = 1$ ). Dus  $\bar{X} \sim \text{Normaal}(\mu = 10, \sigma = 1)$ . Gevraagd wordt  $P(\bar{X} > 12) = P(Z > (12 - 10)/1) = P(Z > 2) = 0.023$ .
- c. Met A wordt aangegeven ‘een 17-jarige jongen’ (met nA wordt aangegeven ‘een 18-jarige jongen’) met B wordt aangegeven ‘Botdichtheid > 10’ (met nB wordt aangegeven ‘Botdichtheid  $\leq 10$ ’);



$P(nA | B) = 0.3 / (0.3 + 0.3) = 0.5$ , met andere woorden de kans dat deze persoon met een botdichtheid > 10 ook een 18-jarige is 0.5! De apriori kans (zonder enige kennis van de botdichtheid) was 0.4 en we zien dat door informatie over de botdichtheid (vrij dicht wat samengaat met ‘ouder’) deze kans verhoogd is naar 0.5.

- 3.
- a. Je zou natuurlijk heel flauw kunnen zijn en zeggen dat we met deze data alleen iets kunnen zeggen over mogelijke verschillen in botdichtheid tussen jongens en meisjes van 17 jaar... Maar goed, eerst moeten we nagaan of we te maken hebben met twee onafhankelijke steekproeven of dat het om gepaarde observaties gaat. Er is geen verband tussen de jongen op regel 1 en het meisje op regel 1. Oke, ze hebben allebei de laagste score van hun sexe, maar dat maakt deze observaties nog niet tot gepaard. Dat verband hebben we zelf gecreëerd door de scores (per geslacht) van laag naar hoog te ordenen! Het zijn dus twee onafhankelijke steekproeven! (NB het zouden gepaarde observaties zijn wanneer we gebruikt zouden hebben gemaakt van twee-eiige tweelingen of wanneer we jongens en meisjes gematched zouden hebben op geboortedag/week o.i.d).
- Vervolgens moeten we bepalen of het om een gerichte of een ongerichte hypothese gaat. Aangezien we vooraf geen idee hebben of jongens of juist meisjes een grotere botdichtheid hebben, hebben we te maken met een ongerichte hypothese. Dus:
- $H_0: \mu_{\text{jongens}} - \mu_{\text{meisjes}} = 0$

$$H_1: \mu_{\text{jongens}} - \mu_{\text{meisjes}} \neq 0$$

We kunnen nu het 95% betrouwbaarheidsinterval bepalen of de p-waarde; voor beiden geldt:  $df = 7 + 7 - 2 = 12$  en dus  $t_{0.025} = 2.18$ ;  $SE(\bar{X}_{\text{jongens}} - \bar{X}_{\text{meisjes}}) = s_p \times \sqrt{(1/n_{\text{jongens}} + 1/n_{\text{meisjes}})} = 2.19 \times \sqrt{(1/7 + 1/7)} \approx 1.17$

$$95\% \text{- BHI: } (\bar{X}_{\text{jongens}} - \bar{X}_{\text{meisjes}}) \pm t_{0.025} \times SE(\bar{X}_{\text{jongens}} - \bar{X}_{\text{meisjes}}) = (10.05 - 8.6) \pm 2.18 \times 1.17 = 1.45 \pm 2.55 = [-1.1; 4].$$

$$\text{Toets: } t = ((\bar{X}_{\text{jongens}} - \bar{X}_{\text{meisjes}}) - 0) / SE(\bar{X}_{\text{jongens}} - \bar{X}_{\text{meisjes}}) = 1.45 / 1.17 \approx 1.24$$

Aangezien 0 in het BHI ligt en/of de gevonden t-waarde  $< t_{0.025}$  kunnen we  $H_0$  niet verwerpen. We hebben dus geen verschil tussen jongens en meisjes qua botdichtheid kunnen aantonen. Je kunt echter niet concluderen dat we bewezen hebben dat jongens en meisjes een gelijke botdichtheid hebben!

- b. Het gevolg zal zijn dat de spreiding in zowel de populatie jongens als de populatie meisjes zal toenemen: de verschillen zullen groter worden aangezien er nu ook verschillen die het gevolg zijn van leeftijdsverschillen bijkomen. Wanneer de spreiding groter wordt in beide populaties, wordt ook de  $s_p$  groter en dus de  $SE(\bar{X}_{\text{jongens}} - \bar{X}_{\text{meisjes}})$  en dus het BHI breder en de p-waarde groter. Dit laatste is wel op voorwaarde dat het gemiddelde verschil gelijk blijft (en dat gebeurt wanneer de gemiddelde leeftijd van de jongens onder 25 jaar niet afwijkt van die van de meisjes en wanneer er geen interactie is tussen leeftijd en geslacht voor wat betreft de invloed op botdichtheid).

4.

Dit kan niet met een 'normale' t-toets (voor gepaarde observaties) en ook niet met een rangtoets want we hebben geen uitkomsten. Dit moet dus met een (aanpassing van) mediaantoets. Een voorbeeld hiervan heb je gezien in de oefenopgaven, n.l. 16-3. Het is ook wel af te leiden. Wanneer er sprake is van een eerlijke beoordeling dan verwacht je dat in de helft van de gevallen de Chinese jongen als oudste zal worden betiteld en in de rest van de gevallen de Kongolese jongen. Mat andere woorden: indien er sprake is van een eerlijke beoordeling dan geldt: noem X het aantal keer dat de Kongolese jongen als oudste geduid wordt: we doen 12 experimenten met telkens een vaste kans op succes van 0.5: X is binomiaal verdeeld met  $\pi = 0.5$  en  $n = 12$ . De p-waarde is de kans op dit resultaat of een nog extremer, m.a.w.  $P(X \geq 9)$  als X binomiaal verdeeld is met  $\pi = 0.5$  en  $n = 12$ . Deze kans haal je direct uit tabel IIIc: 0.073. Nu zou je misschien geneigd zijn om te zeggen dat we hebben

aangetoond dat er sprake is van een niet eerlijke beoordeling; in de opgave stond immers dat we  $\alpha = 0.10$  moesten gebruiken en onze p-waarde is lager dan de  $\alpha$ . Maar vergeet niet dat dit een éézijdige p-waarde is. De tweezijdige is twee keer zo groot, dus 0.146, en deze is groter dan  $\alpha$ . We kunnen dan ook niet stellen dat we hebben aangetoond dat er sprake is van oneerlijke beoordeling. NB. We gebruiken een tweezijdige p-waarde aangezien we vooraf niet zeker weten welke groep als jongste beoordeeld zal worden.

5.

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

a. Om te toetsen of er sprake is van samenhang tussen de echte leeftijd en het oordeel van de beoordelaar maken we gebruik van de  $\chi^2$ -toets:

$H_0$ : Er is geen samenhang tussen de echte leeftijd en het oordeel van de beoordelaar

$H_1$ : Er is wel samenhang tussen de echte leeftijd en het oordeel van de beoordelaar

Nu kunnen we, uitgaande van  $H_0$ , de verwachte frequenties berekenen. 50% is minderjarig, dus je verwacht (indien er geen samenhang is) dat 50% van de als minderjarig beoordeelde ook echt minderjarig zal zijn etc. Dit levert als verwachte frequenties:

Verwacht (E)	Echte leeftijd		
Oordeel	Minderjarig	Meerderjarig	Totaal
“Minderjarig”	25	25	50
“Meerderjarig”	25	25	50
Totaal	50	50	100

O	E	$(O-E)^2/E$
30	25	1
20	25	1
20	25	1
20	25	1

De toetsingsgrootte is de som van de laatste kolom en volgt als  $H_0$  waar is een  $\chi^2$  verdeling met  $(2 - 1) \times (2 - 1) = 1$  vrijheidsgraad:  $\chi^2 = 4$ ,  $df = 1$  tabel VII: p-waarde  $< 0.05$ . De

nulhypothese kan dus verworpen worden; We hebben samenhang tussen de echte leeftijd en het oordeel van de beoordelaar aangetoond.

b. Wil zo'n oordeel zin dan is de samenhang die we net hebben aangetoond een eerste vereiste. Zou er geen samenhang zijn tussen het oordeel en de echte leeftijd dan zou het oordeel even goed zijn als het opgooien van een munt. Maar samenhang alleen is niet genoeg: het moet eigenlijk wel een flinke samenhang zijn. De mate van samenhang kan je bepalen met behulp van de juiste correlatiecoëfficiënt; in dit geval is dat de  $\phi$ -coëfficiënt:  $\sqrt{\chi^2/n} = \sqrt{4/100} = 0.2$ . Dit is natuurlijk niet zo hoog als je weet dat  $\phi$  maximaal 1 kan zijn. Je hoeft echter geen te berekenen (ik had al op het college gezegd dat deze maat niet tot de verplichte tentamenstof behoort), je kunt ook gewoon kijken. Zowel bij de minderjarigen als de meerderjarigen is 40% van de oordelen fout; ook overall is dus 40% van de oordelen fout. Een aangezien bij willekeurig gokken ongeveer 50% fout zou zijn is het oordeel dus niet echt goed te noemen 6.

Er waren 40 observaties ( $n = 40$ ). 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.

- Formule 15-2 (blz. 475) geeft  $r = 268.96/(\sqrt{(296.31 \times 276.5)}) \approx 0.94$ .
- Formule 11-5 (blz. 362) geeft  $b = 268.96/296.31 \approx 0.91$ . Voor  $a$  geldt:  $a = \bar{Y} - 0.91 \bar{X} = 1.748$ . De regressievergelijking wordt dus:  $Y = 1.748 + 0.91X (+ e)$ . Let op dat het in dit geval echt zinnig is om de echte leeftijd te voorspellen. Zo willen we het meetinstrument namelijk gaan gebruiken!
- Vul voor  $X = 19.5$  in: voorspelde  $Y = 1.748 + 0.91 \times 19.5 \approx 19.5$ .
- Er wordt je gevraagd om een toets te doen:  $X = 19.5$  en dus voorspelde  $Y = 19.5$ . Kunnen we nu  $H_0: Y \leq 17$ , of liever  $H_0: Y < 18$  (aangezien je als je 17.99 bent nog steeds minderjarig bent!) verwerpen ten gunste van  $H_1: Y \geq 18$ . NB dit is dus een gerichte toets!!! Dat kan met een BHI (maar denk er dan aan dat je voor een  $\alpha$  van 0.05 een 90% BHI berekent; zie de opmerking in het proeftentamen), of met een toets. Ik kies hier voor de toets:  $t = (\text{gevonden waarde} - \text{waarde in } H_0)/SE$ . De SE haal je uit het BHI:  $s\sqrt{(1/n + (X_0 - \bar{X})^2/\Sigma x^2 + 1)} = 0.956\sqrt{(1/40 + (19.5 - 17.2)^2/296.31 + 1)} = 0.976$  (NB zoals je ziet is dit bijna de residuele variantie; daarom het deze ook wel Standard error of the estimate). De toets wordt dus:  $t = (19.5 - 18)/0.976 \approx 1.54$  Deze waarde is kleiner dan de (éenzijdige!) grenswaarde van 1.70 (of 1.68 al naargelang je bij  $df = 30$  of  $df = 40$  kijkt; kijk echter sowieso bij een  $\alpha = 0.05$  want het gaat om een éenzijdige toets). M.a.w. we

hebben niet met 95% zekerheid aangetoond dat een persoon met een testscore van 19.5 daadwerkelijk ook meerderjarig is. Je snapt nu al wel dat zo'n test niet erg bruikbaar is. Ondanks de hele hoge correlatie tussen voorspelde en echte leeftijd kunnen we bij vrij hoge voorspelde leeftijden nog steeds niet met enige zekerheid stellen dat de persoon meerderjarig is!

### CASE

- a. De correlaties haal je uit de tabel Correlations en deze zijn voor BOTX en LEEFTYD 0.845 en voor BOTY en LEEFTYD 0.847. De significanties haal je uit de rij eronder. Let op dit zijn éézijdige p-waardes. Die zou je dus nog moeten verdubbelen. Maar je ziet zo als dat deze p-waardes zelfs naar verdubbeling zo klein zijn dat ze altijd onder de 0.05 ( $= \alpha$ ) liggen en dat de correlaties dus significant van 0 afwijken (ze zijn significant groter dan 0).
- b. De regressiegewichten haal je uit Coefficients en deze zijn  $-1.394$  voor BOTX en  $2.755$  voor BOTY uit de kolom sig lees je de (tweezijdige) p-waardes af en deze zijn beide hoger dan 0,05 ( $\alpha$ ). M.a.w. we kunnen niet concluderen dat de regressiegewichten van 0 afwijken.
- c. Dat lijkt tegenstrijdig: de correlaties zijn significant (en groot) maar de regressiegewichten zijn dat niet. Bovendien is het regressiegewicht van BOTX negatief terwijl de correlatie duidelijk positief is! De verklaring hiervoor is (multi)colineariteit (zie ook boek blz 501 en verder)! BOTX en BOTY hangen sterk samen ( $r = 0.999$ ); dat is ook niet zo raar aangezien het twee bepalingen van hetzelfde zijn. Als twee prediktoren sterk samenhangen en je stopt ze samen in een regressiemodel dan kun je de geschatte waarden voor de regressiegewichten niet echt serieus meer nemen. De oplossing is om maar één van de botdichtheidsmetingen in je model op te nemen: degene die het makkelijkst (goedkoopst) te bepalen is want geen van beide heeft op statistische gronden de voorkeur boven de ander.
- d. Gebruik model 2 en kijk in de tabel Coefficients. Dat heeft minder prediktoren maar alle prediktoren zijn significant en het model pas niet veel slechter dan model 1:  

$$\text{LEEFTYD} = 4.593 + 1.414\text{BOTY} + 0.06826\text{KRAAK} - 1.296\text{RING}.$$
- e. Die staat onder Model Summary: de multiple correlatie  $R = 0.933$
- f. Een schatting hiervoor is normaal  $(1 - R^2) * 100\%$ . We weten echter dat de R doorgaans te hoog geschat wordt. Daarom kun je beter gebruik van de adjusted  $R^2$  maken en dan kom je aan:  $(1 - 0.866) * 100\% = 13.4\%$