Oefenvragen bij Hoofdstuk 6

1. Leg het gedachte experiment uit waarop de klassieke testtheorie gebaseerd is.
2. Wat is het bezwaar tegen de term ‘ware score’?
3. Wat wordt bedoeld met de opmerking dat meetfouten tautologisch gedefinieerd zijn?
4. Waarom is het in de praktijk van het testen niet zo waarschijnlijk dat iedereen met dezelfde nauwkeurigheid gemeten wordt?

5 Gebruik de eigenschappen [6.8] en [6.12] om te bewijzen dat het klassieke testmodel geschreven kan worden in termen van afwijkingscores, zoals in de formule [6.13].

6. Gegeven zijn de scores van acht personen op een test. De betrouwbare scores zijn bekend (hypothetisch geval).

Proefpersoon	T	E	X	t	e	x
1	9		9			
2	8		9			
3	7		6			
4	6		6			
5	6		6			
6	5		4			
7	4		5			
8	3		3			

   a. Bereken de meetfouten.

   b. Bereken de gemiddelden van T, E en X.

   c. Bereken de afwijkingsscores van t, e en x.

   d. Ga na dat S(T, E) = 0

   e. Geldt S (X, E) = 0? Verklaar het resultaat.

   f. Ga na dat $S_x^2 = S_T^2 + S_E^2$

7. Leg uit wat wordt bedoeld met het onderscheid in:
   a. Systematische en toevallige scorecomponent.
   b. Bedoelde en onbedoelde scorecomponent.
8. Maak gebruik van de gegevens uit opdracht 6.
   a. Bereken $r_{XX'}$.
   b. Vermenigvuldig de meetfouten met een factor 2 en bereken nogmaals $r_{XX'}$.
   c. Vermenigvuldig de betrouwbare scores met een factor 2 en bereken $r_{XX'}$.
   d. Wat valt op bij de uitkomsten van opgave b en c?

9. Gegeven zijn de afwijkingsscores van acht personen op twee paralleltests X en X', alsmede de betrouwbare scores op de eerste test in afwijkingsscorevorm.

Proefpersoon	t	e	x	t'	e'	x'
1	3	0	3			4
2	2	1	3			2
3	1	-1	0			1
4	0	0	0			-1
5	0	0	0			-1
6	-1	-1	-2			-1
7	-2	1	-1			-2
8	-3	0	-3			-2

   a. Bereken de betrouwbare scores t' en de meetfouten e'.

   b. Ga na dat S(e',t) = S(e',t') = S(e',e) = S(e',x) = 0. En tevens dat S(e',x') > 0.

   c. Ga na, dat door alle termen apart uit te rekenen inclusief r(X,X'), dat:

   $R_{XX'} = S_t^2/S_x^2 = S_{t'}^2/S_{x'}^2$.

   d. Ga na dat $S_e^2 = S_{e'}^2$.

10. Bedenk een test voor woordenschat tweemaal vijf items, waarbij de twee vijftallen op inhoudelijke gronden zo goed mogelijk 'parallel' gekozen worden. Probeer dit ook te doen voor twee drietallen van items waarmee de houding ten opzichte van abortus wordt onderzocht. Beschrijf ook wat bij deze twee opdrachten opvalt.
11. Als dezelfde vragenlijst voor functioneren in de klas na een jaar voor de tweede maal aan dezelfde representatieve steekproef van kinderen wordt voorgelegd, levert de correlatie tussen de twee series testscores dan een schatting van de betrouwbaarheid op? Licht het antwoord toe.
12. Leg uit wat een ondergrens voor de betrouwbaarheid is. Geef tevens aan wanneer zo'n ondergrens nuttig kan zijn.
13. Waarom is de term 'interne consistentie' misleidend?
14. Een aantal studenten is gezakt voor een tentamen. Zonder zich beter voorbereid te hebben dan de eerste keer, doen ze mee aan de herhaling. Toch slagen enkele studenten nu wel. Kan nu geconcludeerd worden dat de herhaling gemakkelijker was dan het eerste tentamen?
15. Gegeven is voor een studietoets met vijftig items van het goed/fout-type dat $r_{XX'}$= 0.92. X=32.6 en S(X)=4.1
    a. Schat volgens de lineaire regressiemethode de betrouwbare score van Martijn, die dertig items goed had.
    b. Bepaal of Martijns score significant verschilt van de aftestgrens van 35. Neem hiervoor aan dat schattingsfouten normaal verdeeld zijn, en toets op 5%-significantieniveau.
16. Maak de volgende opgaven.
    a. Bereken de standaardmeetfout van X bij de gegevens uit opdracht 6.
    b. Bereken een 90%-betrouwbaarheidsinterval bij opdracht a.
17. Leg uit hoe de relatief grote onnauwkeurigheid van testscores gecompenseerd wordt door een grotere testlengte.

18. Een test met veertig items en een betrouwbaarheid gelijk aan 0.50 wordt met tien ‘gelijkwaardige’ items uitgebreid. Wat is de betrouwbaarheid van de verlengde test?
19. Ga uit van een test met betrouwbaarheid van 0.25.
    a. Bereken de betrouwbaarheid voor het geval dat de test wordt verlengd met respectievelijk factor 2, 3, 4, 5 en 6.
    b. Welke conclusie kan uit de waargenomen trend worden getrokken?
20. Stel, een test bestaat uit zestig items, met als gevolg dat de testtijd erg lang is. de test heeft een betrouwbaarheid van 0.95. Hoeveel items mag ik weglaten, zodanig dat de betrouwbaarheid niet geringer wordt dan 0.85?
21. Leg uit waarom een testscore X nooit hoger met een variabele kan correleren dan met de betrouwbare score T.
22. Waarom zijn de verschilscores onbetrouwbaar? Van welke factoren is de betrouwbaarheid van verschilscores afhankelijk?
23. Leg uit waardoor een test in de populatie van tienjarige leerlingen onbetrouwbaarder is dan in de populatie van tien- en elfjarigen samen.
24. Waartoe dient de gestratificeerde alfacoëfficiënt?

Antwoorden Oefenvragen Hoofdstuk 6

1. Samengevat kan dit gezegd worden; herhaalbaarheid van metingen kan worden beoordeeld indien we een persoon vele malen dezelfde test onder gelijkblijvende condities voorleggen. Daarbij geldt dan dat de testprestaties bij verschillende afnemingen onafhankelijk van elkaar zijn; de persoon leert niet van afneming tot afneming en herinnert zich niets van vorige afnemingen. Bij iedere testsessie wordt als het ware weer opnieuw begonnen. De testsituatie is onveranderd gebleven en steeds zijn alle voor de meting relevante eigenschappen van de persoon van invloed op diens testprestatie. In deze situatie zijn er bij verschillende afnemingen factoren werkzaam die de testprestatie op onvoorspelbare wijze beïnvloeden. De klassieke testtheorie houdt zich bezig met het in kaart brengen van de relatieve inbreng van de over afnemingen onvoorspelbare invloeden op de testprestaties en de over afnemingen systematische werkzame eigenschappen van personen en testsituatie.
2. Ze kunnen aanleiding geven tot een platonische opvatting over datgene waar het symbool T voor staat. De term ‘ware’ of ‘true’ lijkt te refereren aan iets wat buiten de concrete testsituatie bestaat, in plaats van aan een gemiddelde, representatieve testprestatie.
3. Dat betekent dat het gebaseerd is op een cirkelredenering. De meetfout op replicatie j is dat deel van de geobserveerde testscore dat resteert wanneer de betrouwbare score ervan afgetrokken wordt. Ook hier dus geen referentie aan inhoudelijke, buiten de test en de testsituatie bestaande oorzaken van meetfouten.
4. Het is niet realistisch, als je bijvoorbeeld een kennistest neemt zal een persoon die veel weet anders scoren als een persoon die weinig weet.
5. [6.8] =

$$\bar{E} = \frac{1}{n}\sum_{i=1}^{n} E_i = 0.$$

Hierbij wordt verondersteld dat de gemiddelde meetfout in een populatie van n personen gelijk is aan nul.

[6.12] =

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} (T_i + E_i) = \bar{T}.$$

Volgens bovenstaande formule geldt dus dat in de populatie de gemiddelde geobserveerde score en de gemiddelde betrouwbare score gelijk zijn.

Wanneer je gebruik maakt van de eigenschappen van beide formules dan luidt het klassieke model in afwijkingsscorevorm:

[6.13] =

$$x_j = t_j + e_j.$$

6.

a. De meetfout bereken je door het verschil in T en X te bepalen.

Proefpersoon	T	E	X	t	e	x
1	9	0	9	3	0	3
2	8	1	9	2	1	3
3	7	-1	6	1	-1	0
4	6	0	6	0	0	0
5	6	0	6	0	0	0
6	5	-1	4	-1	-1	-2
7	4	1	5	-2	1	-1
8	3	0	3	-3	0	-3

b. Gemiddelde van X is zes, gemiddelde van E is nul en gemiddelde van T is zes.

c. De afwijkingscore bereken je door te kijken hoeveel het getal afwijkt van het gemiddelde.

d. S(T,E) = 0, omdat meetfouten met geen enkele andere variabele correleren, alleen met X. Daarom is S(T,E) ten alle tijden nul. Je kan het controleren door de formule voor de covariantie van twee variabelen te gebruiken.

e. S(X,E) = 0 gaat niet op. Want S(X,E) is 0,5.

f. $$S^2 = \frac{\Sigma (x - \bar{x})^2}{n} = \frac{\Sigma x_i^2}{n}$$

Dat als voorbeeld uitgewerkt voor X.

9-6 = 3, vervolgens $3^2 = 9$

9-6 = 3, vervolgens $3^2 = 9$

6-6 = 0, vervolgens $0^2 = 0$

6-6 = 0, vervolgens $0^2 = 0$

6-6 = 0, vervolgens $0^2 = 0$

4-6 = -2, vervolgens $-2^2 = 4$

5-6 = -1, vervolgens $-1^2 = 1$

3-6 = -3, vervolgens $-3^2 = 9$

Al die waarden optellen; 9+9+4+1+9=32, en dat vervolgens delen voor n, dus door 8.

Voor X komt er dus de waarde 4 uit.

Dit kan je voor de waarden E en T op dezelfde manier uitwerken.

Dan komt er het volgende uit; 4 = 3,5 + 0,5. De formule is dus juist.

7.

a. Onderscheid tussen systematische en toevallige scorecomponent.

$X_{ij} = T_i + E_{ij.}$

Hierin is T het systematische deel en E het toevallige deel. De systematische scorecomponent is over onafhankelijke replicaties een constante. T wordt gedefinieerd als de gemiddelde, geobserveerde score die persoon i heeft behaald over een zeer groot aantal onafhankelijke replicaties van de test.

Het toevallige deel E in de formule varieert over replicaties daarentegen op een onvoorspelbare wijze, daarom is het subscript j wel toegevoegd. i staat voor de persoon en j staat voor de replicatie.

b. Onderscheid tussen bedoelde en onbedoelde scorecomponent.
Een voorbeeld, er is test voor rigiditeit. Deze test meet naast rigiditeit ook emotionaliteit, agressiviteit en woordbegrip. De rigiditeitstrek is de bedoelde scorecomponent. De onbedoelde scorecomponent zijn emotionaliteit, agressiviteit en woordbegrip. Ook de meetfout valt onder de onbedoelde scorecomponent.

8.

a. $r_{XX'} = S^2(T) / S^2(X)$.
Dus 3.5/4= 0.875

b. De variantie van E wordt na vermenigvuldigen met 2 gelijk aan 2. Je moet dan de formule $S_x^2 = S_T^2 + S_E^2$ opnieuw invullen. S_x^2 wordt nu 5,5. Dat invullen geeft; 3.5/5.5 = 0.64

c. De variantie van T wordt na vermenigvuldiging met 2 gelijk aan 14. Verder volg je dezelfde stappen als in antwoord b. er komt dan een betrouwbaarheid uit van 0.97.

d. Bij b: Als de variantie van *E* groter wordt ten opzichte van de variantie van *T*, dus als meetfouten naar verhouding een grotere invloed op de testprestaties hebben, dan gaat de betrouwbaarheid er op achteruit. Bij c: Hier zien we het tegengestelde effect op de betrouwbaarheid als de relatieve invloed van meetfouten juist kleiner wordt.

9.

a.

Proefpersoon	t	e	x	t'	e'	x'
1	3	0	3	3	1	4
2	2	1	3	2	0	2
3	1	-1	0	1	0	1
4	0	0	0	0	-1	-1
5	0	0	0	0	-1	-1
6	-1	-1	-2	-1	0	-1
7	-2	1	-1	-2	0	-2
8	-3	0	-3	-3	1	-2

b. Meetfouten (E) correleren met geen enkele andere variabele, alleen met X. Daarom is S(T,E) ten alle tijden nul. Je kan het controleren door de formule voor de covariantie van twee variabelen te gebruiken

c. De berekening gaat hetzelfde als in opdracht 6.
Er komt uit; $r_{XX'} = 28/32 = 28/32 = 0.875$.

d. De berekening gaat eveneens hetzelfde als in opdracht 6. De uitkomst is bij beiden 4.

10. Beide testen zelf bedenken. Het valt op dat het bij abortus veel moeilijker is, omdat je een mening uitvraagt. Dat kan niet aan de hand van twee drietallen van items. Je kan geen parallel vragen maken voor dat onderwerp.

11. Nee, de kinderen hebben in een jaar tijd dingen bijgeleerd. Daarom is de correlatie van de twee series scores geen goede schatting van de betrouwbaarheid.

12. Een ondergrens voor betrouwbaarheid is handig voor kleinere steekproeven, deze wijken door toeval vaak sterk van de populatie af. De maat alfa wordt gebruikt als ondergrens, het kan in dit geval heel goed zijn dat de waarde van alfa zo onnauwkeurig is geschat dat zij zelfs groter uitvalt dan de betrouwbaarheid. Daarom is een ondergrens zinvol.
13. Er zijn twee redenen waarom de opvatting nogal ongelukkig is. Ten eerste is alfa in veel gevallen een toenemende functie van het aantal items in de test. Een hoge betrouwbaarheid heeft dus alles te maken met de nauwkeurigheid van een meting, maar niet met wat de test mee. Interne consistentie zou onafhankelijk moeten zijn van het aantal items. Ten tweede kan alfa een hogere waarde hebben terwijl de test inhoudelijk in sterke mate heterogeen is.
14. Nee dat kan je met deze gegevens niet concluderen. Er kunnen andere factoren mee spelen, daarbij valt te denken aan de geestelijke toestand van de student, aan de omgeving en aan de layout van de test.
15.

a. De schatting van de betrouwbare score van Martijn kan berekent worden met onderstaande formule;

$$\hat{Y} = \frac{S(Y)}{S(X)} r(X,Y)[X - \bar{X}] + \bar{Y}.$$

S(Y) = S(X) = 4,1
r(X,Y) = 0.92
X = 30
gemiddelde van X = 32.6
$\bar{Y}$ = 32.6
Dit allemaal invullen in de formule geeft een score van 30.208

b. Daarvoor stel je een 95% betrouwbaarheidsinterval op.
De ondergrens is 28.03 en de bovengrens 32.39. Het getal 35 valt hier niet binnen.

16.

a. De formule van de standaardmeetfout is $S(E) = S(X)\sqrt{1 - r_{XX'}}$.
$R_{xx'} = S^2(T) / S^2(X) = 3.5/4 = 0.875$
$S^2(X) = 4$, dus S(X)=2.
Dit invullen in de formule geeft als antwoord 0.707 (afgerond).

b. T +/- 1.65*0.707

17. Door de testlengte te vergroten komt er een betrouwbaarder gemiddelde uit.
18. Door gebruik te maken van de volgende formule kom je op het antwoord.

$$r_{KK} = \frac{Kr_{XX'}}{1 + (K-1)r_{XX'}}.$$

r_{KK} is de betrouwbaarheid van de verkorte/verlengde test.

K is het aantal items verlengde test/aantal items oorspronkelijke test.

$r_{xx'}$ is de betrouwbaarheid van de gehele test.

Als je de getallen invult kom je uit op een betrouwbaarheid van 0.56 (afgerond).

19.

a. Gebruik dezelfde formule als in opgave 18. De antwoorden die er dan uitkomen zijn respectievelijk; .4, .5, .57 (afgerond), .625, .67 (afgerond)
b. De conclusie die hieruit te trekken valt is dat de betrouwbaarheid groter wordt wanneer de test groter wordt.

20. Om deze opdracht te beantwoorden moet de volgende vergelijking opgelost worden; 0.85 = (x/60 * 0.95) / (1+(x/60 -1) *0.95). Uit deze vergelijking komt dat er 42 items weggelaten kunnen worden.
21. Testscores zijn (tamelijk onnauwkeurige) schattingen van de betrouwbare score. Daarom kan een testscore nooit hoger met een variabele correleren dan met de betrouwbare score.
22. De verschilscores zijn onbetrouwbaar, omdat het afhankelijk is van de onbetrouwbaarheid van één of van beide testscores. Betrouwbaarheid van het verschil is lager naarmate de betrouwbaarheid van x1 en x2 afzonderlijk lager is. verder is de betrouwbaarheid van verschilscore gering als de samenhang tussen x1 en x2 sterk is.
23. Hoe groter de populatie waar de test zich over strekt, hoe betrouwbaarder het is.
24. De gestratificeerde alfacoëfficiënt kan gebruikt worden als vervanging van de paralleltest- of test-hertestmethode. Die zijn meer bewerkelijk.