Statistiek voor het laboratorium

Teo Kleintjes

"Statistiek betekent dat je nooit hoeft te zeggen dat je het zeker weet"

C.J. Bradfield

Didactisch concept : Vervoort Boeken
Grafisch ontwerp: uwontwerp.nl Eindhoven
Versie juni-2018

isbn 9789464180077

Verantwoording

Dit onderdeel van de methode is bedoeld voor iedereen die zijn vaardigheden en basiskennis van de statistiek wil verbeteren. Voor bedrijven en instellingen is het leveren van *kwaliteit* aan de klanten van levensbelang. Daarom is de laatste jaren het vak *kwaliteitszorg* sterk ontwikkeld. Omdat kwaliteitszorg sterk op statistiek leunt, is het van belang dat je daar als moderne werknemer iets van weet en kunt gebruiken tijdens je werk.

In dit deel worden reflectievragen gebruikt om meer inzicht en overzicht te krijgen.

De vele interactieve oefeningen op internet en het gebruik van Excel en SPSS zijn bijzonder geschikt om de rekenvaardigheid onder de knie te krijgen.

De site www.vervoortboeken.nl is een belangrijke ondersteuning. Hier zijn hulpmiddelen te vinden zoals uitwerkingen, links naar internetsites, Excel-tools en extra uitleg via powerpoint-tools.
De links naar internet verwijzen door naar sites met simulaties of oefenmogelijkheden.

Heel veel succes!

Speciale dank gaat uit naar de collega's Claartje Eggermont, Nazlı Evlek, Liesbeth Coffeng, Franca van de Loo en Jos Vervoort voor hun onvoorwaardelijke feedback en steun.

Gebruikte iconen :

1.1	**Reflectievragen**
3.2	**Samenvatting voor aantekenschrift**
1.1	**Verwijzing naar internetsite**
01	**Excel-tool op de site** http://www.vervoortboeken.nl

Waarom de rood gekleurde woorden in de tekst?

Maak voor jezelf een register waarin je zelf de betekenis van deze woorden of begrippen beschrijft.
Neem 1 bladzijde per letter.

Inhoudsopgave

1.Precisie en juistheid

2.Meetwaarden verschillen. Hoe komt dat?

3.Spreiding van data (meetresultaten)

4.Uitschieters bepalen en afronden

5.Normaalverdeling

6.Van steekproef naar populatie

7.Kwaliteitszorg (controlekaarten)

8.Correlatie en regressie

9.Testen van meetresultaten

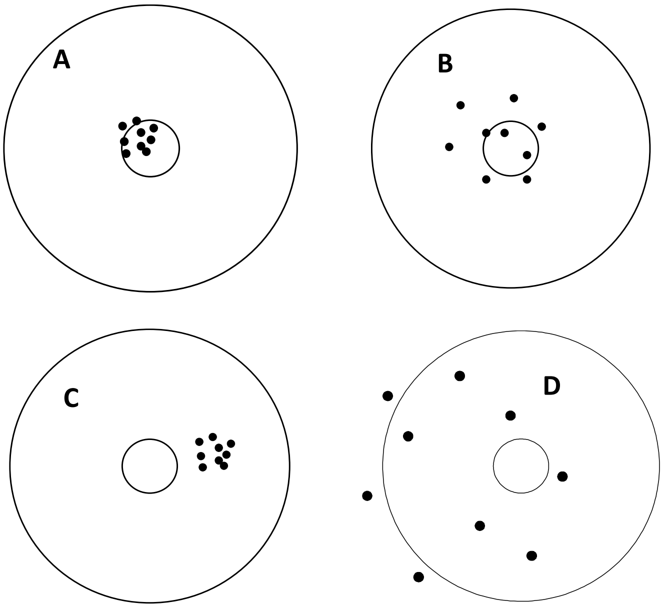
10.Extra oefeningen

Opgave 1.1

Blz. 8

Precisie en juistheid bij het schieten
Een voorbeeld van precies en juist vinden we bij schieten op een schietschijf.

Vier schutters hebben schoten afgevuurd. Zie volgende pagina.



Welke kwalificatie hoort bij welk plaatje ?

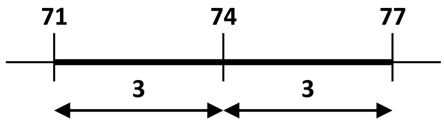
1. precies en juist
2. niet precies en onjuist
3. precies en onjuist
4. niet precies en juist

Geef in het antwoord aan welk cijfer bij welke letter hoort.

Opgave 1.3

Hoe precies is een meting bij een bepaalde meetmethode?
Bij een apparaat of methode is de precisie bekend.

Voorbeeld
De politie heeft mijn snelheid gecontroleerd en 74 km/h gemeten. De meetmethode heeft een nauwkeurigheid van 3 km/h.
Dat betekent dat de meetresultaten maximaal 3 km/h kunnen afwijken.
Het resultaat noteer je als: $\boldsymbol{v = 74 \pm 3}$ **km/h**.
In plaats van nauwkeurigheid kun je beter spreken van **onnauwkeurigheid**.De meting met onnauwkeurigheid kun je tekenen als een **interval** op een getallenlijn:



De meting heeft dan een **absolute onnauwkeurigheid** van 3 km/h. Je kunt dit ook relatief weergeven dus in procenten, die noem je dan de **relatieve onnauwkeurigheid**.

relatieve onnauwkeurigheid

$$\text{relatieve onnauwkeurigheid} = \frac{\text{absolute onnauwkeurigheid}}{\text{gemeten waarde}}$$

Blz. 10

$$\text{relatieve onnauwkeurigheid} = \frac{3}{74} \times 100\ \% = 4{,}1\ \%$$

twee notaties van onnauwkeurigheid

met absolute onnauwkeurigheid	v = 74 ± 3 km/h
met relatieve onnauwkeurigheid	v = 74 km/h ± 4,1 %

Van een andere wegmisbruiker wordt de snelheid bepaald op 135 km/h.

a Geef de juiste notatie van deze meting met absolute en relatieve onnauwkeurigheid.

b Maak ook een tekening van het interval.

Opgave 2.13 Verschilmeting

Met een buret heb je twee volumes afgelezen

Buretaflezing	
V_{begin} (mL)	35,18
V_{eind} (mL)	11,56
afleesonnauwkeurigheid (mL)	0,02
instrumentonnauwkeurigheid (mL)	0,05

Blz. 24

a Bereken het toegevoegde volume vloeistof.

b De instrumentonnauwkeurigheid van 0,05 is een systematische onnauwkeurigheid. Wat betekent dat?

c Stel dat de instrumentonnauwkeurigheid +0,04 mL is, wat betekent dat dan voor het toegevoegde volume vloeistof?

d Bereken de onnauwkeurigheid in het resultaat.

e Noteer de volgende regel:

verschilmeting en onnauwkeurigheid

Bij een verschilmeting met één instrument hoef je alleen rekening te houden met de afleesonnauwkeurigheid

Opgave 3.5

Blz. 32

3.1

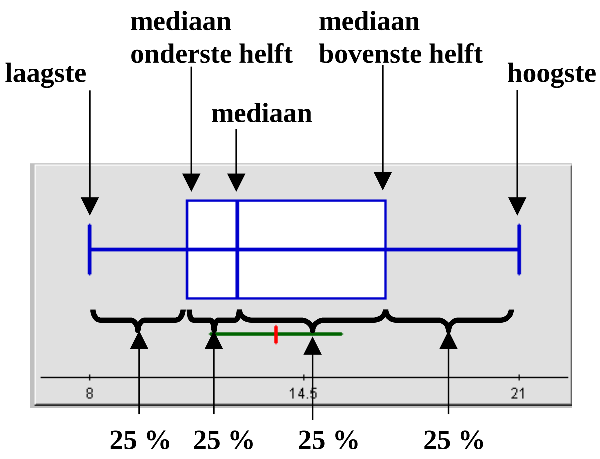
Boxplot

Een **boxplot** is een eenvoudige manier om de verdeling van data (waarnemingen, meetwaarden) weer te geven. Een voorbeeld (zie tabel rechts):

- sorteer de 14 gegevens van laag naar hoog
  8, 9, 9, 11, 11, 12, 12, 13, 15, 16, 17, 17, 20, 21
- bepaal de mediaan; in dit geval 12,5
- bepaal de mediaan van de onderste helft: 11
- bepaal de mediaan van de bovenste helft 17
- maak met deze waarden een tekening zoals in het voorbeeld
  gemaakt met:
  http://nlvm.usu.edu/en/nav/frames_asid_200_g_3_t_5.html

	Data
1	11
2	12
3	13
4	16
5	9
6	11
7	17
8	12
9	17
10	20
11	8
12	9
13	21
14	15

<none>



De mediaan van de onderste helft wordt ook wel het **eerste** (of onderste) **kwartiel** K_1 genoemd. De mediaan van de bovenste helft is dan het **derde** (bovenste) **kwartiel** K_3.

a Maak nu zelf een boxplot (met de hand of met de applicatie op de website) van de examenscores uit de vorige opgave.

b Is hier sprake van een scheve verdeling?

Zoals we later zullen zien kun je een boxplot ook gebruiken om uitschieters te bepalen.

bij een steekproef delen we niet door *n* maar door *n*-1

- bereken het gemiddelde van de kwadraten:

$$\frac{\sum (x_i - \overline{x})^2}{n-1} = \frac{418}{8} = 52,25$$

Het getal dat we nu gevonden hebben wordt **variantie** genoemd.

$$\text{variantie } = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

Blz. 39

- neem nu hier de wortel uit:

$$\sigma_{n-1} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} = \sqrt{52,25} = 7,23$$

06

- dit getal is een maat voor de gemiddelde afwijking van de waarnemingen t.o.v. het gemiddelde en heet de **standaarddeviatie** of **standaardafwijking** (het symbool σ heet ook **sigma**), het is een goede maat voor de precisie van de meting.

De standaarddeviatie is, net zoals de eerder gebruikte spreiding, een absolute onnauwkeurigheid en dus een maat voor de precisie Als we ook de relatieve onnauwkeurigheid willen weten, rekenen we de **variatiecoëfficiënt** uit:

$$\text{variatieco ëfficiënt } = \frac{\sigma_{n-1}}{\overline{x}} \cdot 100\ \%$$

c Bereken de variatiecoëfficiënt van meting A.
Waarom geeft deze waarde een goed beeld van de precisie van de meting?

d Maak eenzelfde tabel voor meting B en bereken de standaarddeviatie en variatiecoëfficiënt van deze meetserie.

e Vergelijk de precisie van meting A en meting B.

Samenvattend, bij een steekproef geldt dus:

variantie van een steekproef

$$\text{variantie } = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

standaarddeviatie van een steekproef

$$\sigma_{n-1} = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$$

variatiecoëfficiënt van een steekproef

$$\text{variatiecoëfficiënt } = \frac{\sigma_{n-1}}{\overline{x}} \cdot 100\,\%$$

Opgave 4.1

Blz. 46

Uitschieters: de **Dixons-test** of **Q-test**

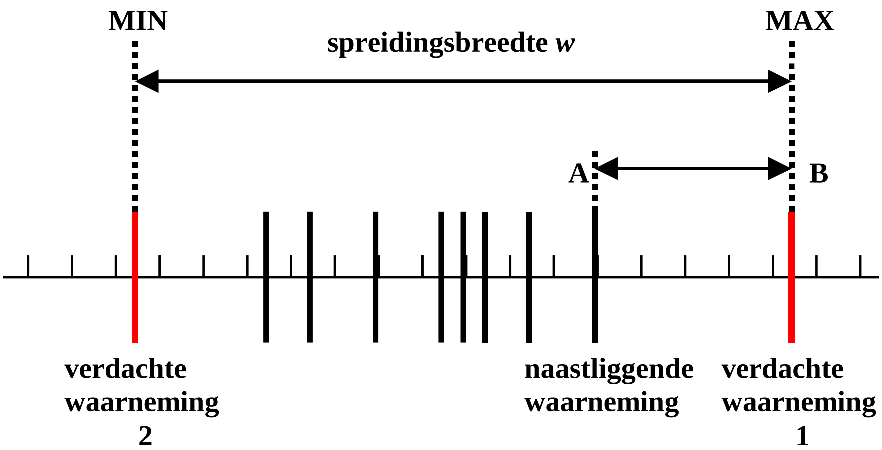
Van afvalwater is 10 maal het loodgehalte bepaald:

Loodgehalte in afvalwater (mg/L)				
2,4	2,1	2,3	1,7	2,2
2,5	2,4	3,7	2,4	2,3

a Zijn hier mogelijk **uitschieters** aanwezig, denk je?

b Zet de waarden op een getallenlijn.. Helpt dit om een uitschieter aan te wijzen?

Ons gevoel kan ons bedriegen bij statistiek, daarom zijn er regels afgesproken.
Een uitschieter is een minimum of maximum meetwaarde die opvallend ver van de naastliggende af ligt, zie tekening



De Dixons-test maakt het mogelijk om een enkele uitschieter op te sporen. Als er uitschieters in groepjes voorkomen moeten we een boxplot inzetten.
Bij de Dixons-test kijkje naar de verhouding tussen twee lijnstukken: AB en de spreidingsbreedte *w*.
De meest verdachte waarneming is 3,7

c Bereken de lengte van AB, dus:
AB = verdachte waarneming 1 – naastliggende waarneming

d Bereken *w*.

e Bereken:

testwaarde Dixons-test of Q-test

$$Q_{test} = \frac{|\text{verdachte waarneming} - \text{naastliggende waarneming}|}{w}$$

de rechte strepen in

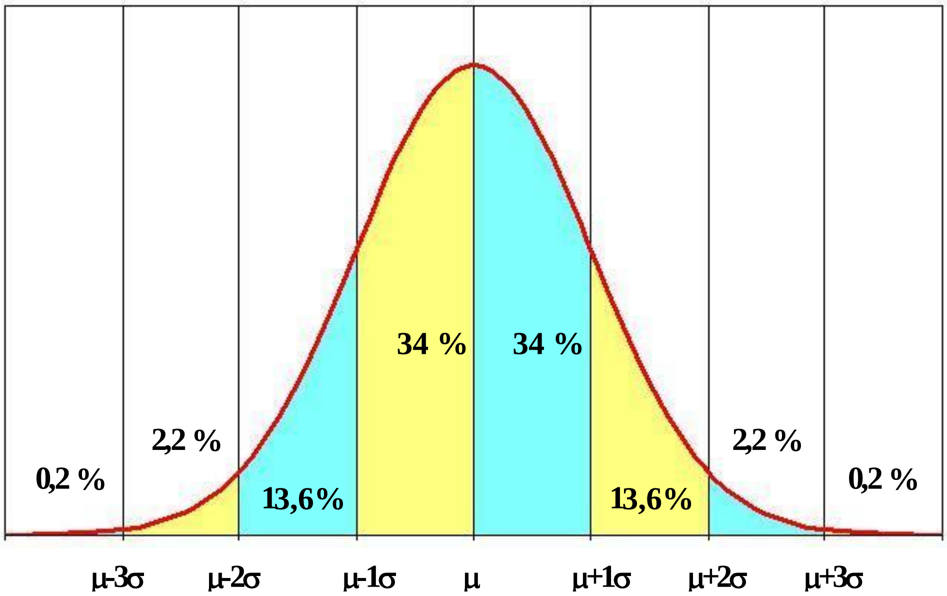
|A-B| betekenen dat de uitkomst altijd positief

Opgave 5.2

Blz. 57

Vuistregels normaalverdeling

In onderstaande figuur zijn alle percentages aangegeven die bij een normaalverdeling horen bij de gebieden die 1, 2 of 3 standaarddeviaties van het gemiddelde liggen.



a Hoeveel % ligt tussen $\mu\boldsymbol{-1}\sigma$ en $\mu\boldsymbol{+1}\sigma$?

b Hoeveel % ligt tussen $\mu\boldsymbol{-2}\sigma$ en $\mu\boldsymbol{+2}\sigma$?

c Hoeveel % ligt tussen $\mu\boldsymbol{-3}\sigma$ en $\mu\boldsymbol{+3}\sigma$?

d Neem de figuur over en zet deze getallen er ook bij.

De getallen krijgen meer betekenis met een voorbeeld.
Het gemiddelde gewicht van de Nederlandse man is 82,4 kg. Van de mannen weegt 2,4 % minder dan 54,0 kg.

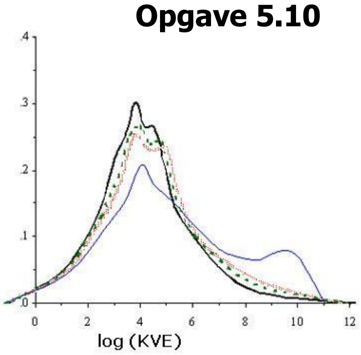
e Bereken de standaarddeviatie van het gewicht. Maak als hulpmiddel een schets van de normaalverdeling met getallen.

f Teken een normaalverdeling zoals in de figuur met alle juiste gewichten.

g Vul in: 2,4 % van alle Nederlandse mannen weegt meer dan ……. kg (tip: kijk goed in de figuur).

h Hoeveel % van de mannen heeft een gewicht tussen 96,6 kg en 110,8 kg?

Opgave 5.10



Standaard normaalverdeling en microbiologische metingen

De uitslagen van microbiologische meetmethoden zijn niet normaal verdeeld maar kennen een zogeheten Poissonverdeling. Dat komt door de enorme onzekerheid die optreedt bij dit soort metingen, met name bij zeer lage waarden, dus bij grote verdunningen.
Het is niet nodig om hier verder op in te gaan, want het blijkt dat de logaritmische waarde (^{10}logwaarde) van de uitslagen wel normaal verdeeld is, zie figuur links.
Van een monster is een aantal van 100 KVE geteld bij een verdunning van 10^2.

a Bereken het gehalte KVE in het monster.

De onnauwkeurigheid (standaarddeviatie) van de meetmethode is 0,15 (uitgedrukt in logwaarde)

b Bereken de 2σ grenzen van het gehalte in $^{10}\log$ (KVE).
c Bereken de 2σ grenzen van het gehalte in KVE.
d Hoe zie je aan de uitkomst van c dat de metingen in KVE niet normaal verdeeld zijn?

Opgave 6.3 Kan het niet met wat minder steekproeven?

Blz. 74

250 keer? vette shit !!

Het hoofdstuk begon met het probleem van het zoutgehalte van een oplossing dat in triplo bepaald was met als uitslag $\overline{x}$ = 15,1 mg/L en σ_{n-1} = 0,2 mg/L. Kunnen we daar ook de standaardfout bij gebruiken?

Het is weinig aantrekkelijk om deze steekproef 10 keer te herhalen, laat staan om per steekproef 25 samples te gebruiken. Dat zou neerkomen op 250 analyses voor één uitslag. De chef lab zou niet blij zijn.

Gelukkig hebben statistici overal een oplossing voor. Je kunt de formule

$$SE = \frac{\sigma_n}{\sqrt{n}}$$

ook gebruiken om de standaardfout te berekenen als de standaarddeviatie van de populatie bekend is. Dan moeten we wel eerst bedenken van de populatie van het meten van een monster voorstelt.

Vergelijk:
lengtemeting triplo $\overline{x}$ = 181,2 cm en σ_{n-1} = 7,1 cm
meting zoutgehalte $\overline{x}$ = 15,1 mg/L en σ_{n-1} = 0,2 mg/L

6.1

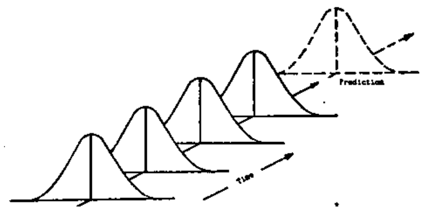
R1 Waardoor wordt de spreiding in de resultaten van de lengtemeting bepaald? Uit de onnauwkeurigheid van de meetmethode of de verdeling van de lengte over de populatie van Nederlandse mannen? Of allebei?
R2 Waardoor wordt de spreiding in de resultaten van de zoutmeting bepaald? Uit de onnauwkeurigheid van mijn meetmethode of de verdeling van het zoutgehalte over de populatie (het monster)? Of allebei?
R3 Wat stelt de populatie eigenlijk voor bij zo'n gehaltebepaling?

Theorie
Shewhart van Bell Telephone Laboratories heeft in 1924 gezegd dat er twee soorten variaties zijn met verschillende oorzaken:

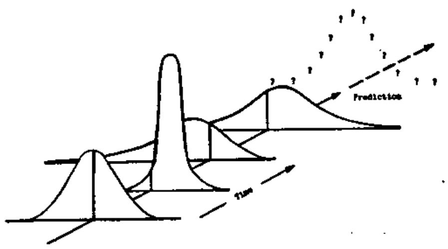
toevallige en aanwijsbare oorzaken

Blz. 86

Toevallige oorzaken
(Engels: random causes), aanwezig in ieder proces en niet te vermijden. Bijvoorbeeld: het trekken van rode ballen; we verwachten een normaalverdeling, ook in de toekomst: figuur rechts. Als er alleen toevallige oorzaken zijn is het proces onder controle (Eng. "in control").



Aanwijsbare, speciale oorzaken
(Engels: special causes), die niet gewenst zijn en vermeden kunnen worden. Als er zich speciale oorzaken voordoen is het proces niet onder controle (Eng. "out of control") en moet er ingegrepen worden.
Men spreekt ook wel van *onbeheerste kwaliteit.*



out of control
onbeheerste kwaliteit

Een proces is onder controle als we op basis van ervaringen in het verleden, binnen zekere grenzen kunnen voorspellen hoe het zich in de toekomst zal gedragen.

Walter A. Shewhart, 1931

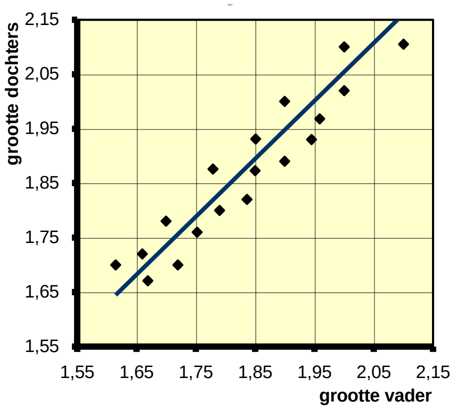
Hoe stellen we nu vast of de kwaliteit onbeheerst is of onbeheerst dreigt te raken? Het hulpmiddel hierbij vormen zogenoemde (proces)**controlekaarten**. In het Engels SPC charts: Statistical Process Control charts.
Hierin worden de eigenschappen (variabelen) van het proces weergegeven, die gemeten worden als functie van de tijd.

Opgave 8.1 **Wel of geen verband tussen de grootheden?**

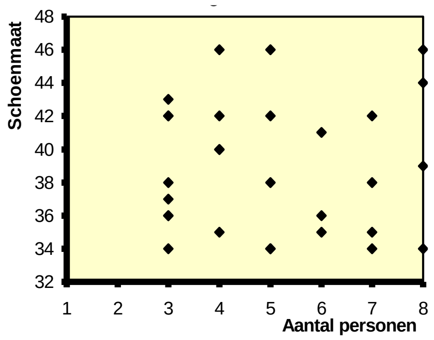
Bekijk de volgende voorbeelden:

Blz. 97

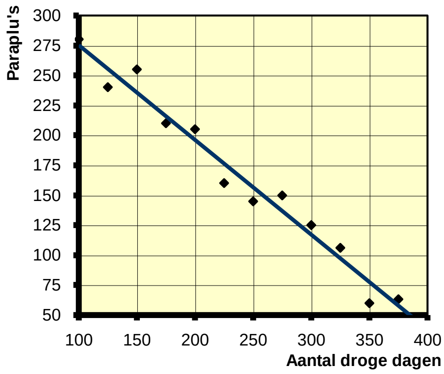
Hebben grote vaders ook grote zonen of grote dochters? Of is er helemaal geen verband?



Is er een verband tussen het aantal personen dat in een huis woont en de schoenmaat van de jongste bewoner?



Soms heb je ook een verband waarbij de ene grootheid groter wordt en de andere grootheid kleiner. Bijvoorbeeld: het aantal droge dagen (in een jaar) en het aantal verkochte paraplu's in een winkelketen.



a In twee gevallen is sprake van **correlatie** (afhankelijkheid) tussen de grootheden. Welke zijn dat?

b Bij één van de twee is de **correlatie positief**, bij de andere is de **correlatie negatief**. Welke zijn dat volgens jou?

Correlatie kan in een getal worden uitgedrukt: de **correlatiecoëfficiënt** symbool *r*. Dat is een getal tussen –1 en +1.

Opgave 9.1
μ

Testen van het uit de steekproef geschatte gemiddelde t.o.v.

Je gebruikt deze test in een lab bijvoorbeeld:
- als je iets wil vergelijken met een bekende of gewenste waarde;
- om de juistheid van een analysetechniek te beoordelen.

Hier nemen we als voorbeeld een zak cement waar 25 kg op staat. Bij kwaliteitscontrole in de fabriek wordt een steekproef genomen:

Gewicht van zakken cement (kg)				
23,5	24,8	23,9	25,8	24,6

Blz. 109

a Bereken het gemiddelde en de standaarddeviatie.

Voldoet de fabrikant hiermee aan de specificatie? Op het eerste gezicht lijkt dit van niet. Maar het kan *toeval* zijn dat er vier waarden onder 25 kg liggen. Bij herhaling van de steekproef zouden dan best alle 5 waarden boven de 25 kg kunnen liggen. Een test zal dit moeten uitwijzen. We volgen het stappenplan.

t- test van het uit de steekproef geschatte gemiddelde t.o.v. μ

1 De nulhypothese stellen:
H$_0$: "het gemiddeld gewicht verschilt **niet significant** van 25"
(dit is beter dan: "het gewicht voldoet aan de specificatie")
H$_0$: $\mu = 25$
2 Alternatieve hypothese
H$_1$: "het gemiddeld gewicht verschilt **significant** van 25"
H$_1$: $\mu \neq 25$ (geen voorkeur voor groter of kleiner).
3 Dus tweezijdig testen.
4 De testwaarde is: $t = \left|(\mu - \bar{x})\right| \cdot \dfrac{\sqrt{n}}{\sigma_{n\text{-}1}}$

modulus strepen: uitkomst > 0

b Bereken deze *t*–waarde.
5 We willen een betrouwbaarheid van 95 %.
c Wat betekent dat bij tweezijdig toetsen? Hoeveel % links en rechts?
6 Zoek de kritische waarde op en vergelijk die met de testwaarde.
d Hoeveel vrijheidsgraden zijn er? Zoek de kritische waarde $t_{kritisch}$ op in de *t*–tabel.
e Vergelijk die met de testwaarde: is $t_{berekend} > t_{kritisch}$?
Zo **ja**, dan wordt de nulhypothese **aangenomen**.
Zo **nee**, dan is hij **verworpen** en moet de alternatieve hypothese dus waar zijn.
f Wordt de nulhypothese verworpen?
7 Conclusie trekken.
g Welke conclusie kunnen we trekken uit de test? In woorden graag!!

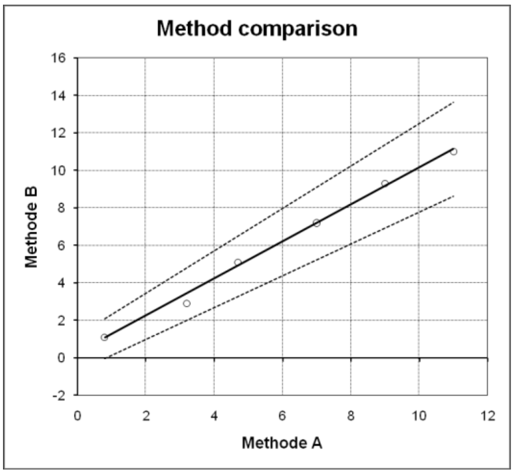
14

19

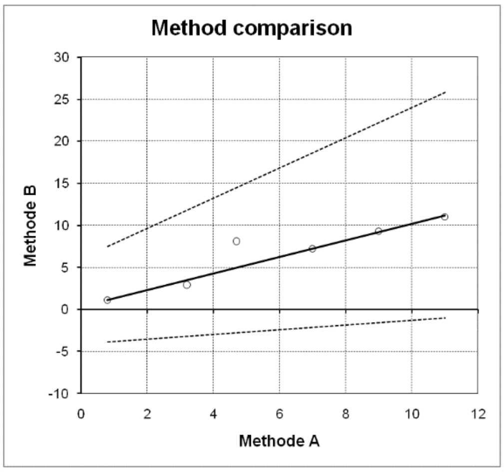
10.1

Met Excel kun je geen Passing en Bablok regressie doen. Een aanvulling op Excel waarmee dat wel kan is *Analyse-it*, maar dat moet gekocht worden. Er is wel een demoversie te downloaden.
Op de Duitse website: www.acomed-statistik.de kun je een Excel werkblad downloaden waarmee je Passing en Bablok regressie kunt uitvoeren.
Het resultaat van bovenstaand probleem zie je onder:

Blz. 124



Zonder uitschieter
$y = 0{,}984 \cdot x + 0{,}313$



Met uitschieter
$y = 0{,}984 \cdot x + 0{,}313$

a Waar zit het verschil nu in de uitslag? Kijk in de grafiek!

De software geeft ook de betrouwbaarheidsintervallen:

Resultaat van PB-Regressie	
helling	0,9839
ondergrens 95%-BI	0,8500
bovengrens 95%-BI	1,1316
snijpunt	0,3129
ondergrens 95%-BI	-0,7211
bovengrens 95%-BI	1,1775

Zonder uitschieter

Resultaat van PB-Regressie	
helling	0,9839
ondergrens 95%-BI	0,2791
bovengrens 95%-BI	1,7949
snijpunt	0,3129
ondergrens 95%-BI	-4,1038
bovengrens 95%-BI	6,0174

Met uitschieter

b Zijn de methoden vergelijkbaar? Dus: liggen de ideale waarden in de gevonden betrouwbaarheidsintervallen?