1 Meten en statistiek

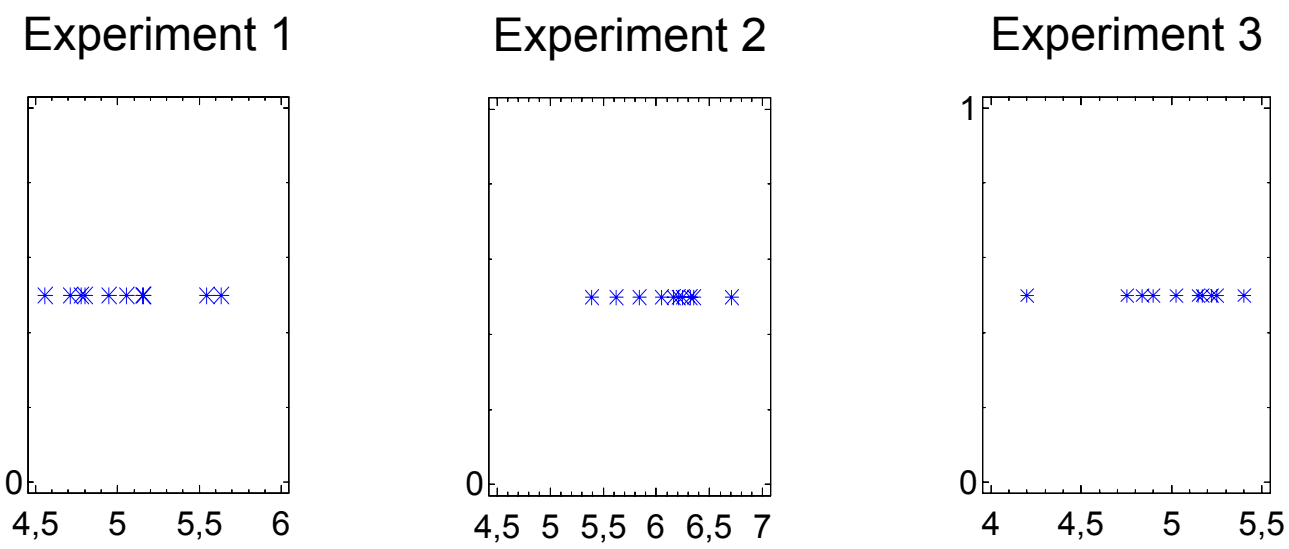
Bij het verrichten van metingen moeten we ons realiseren dat elke meting behept is met bepaalde onzekerheden of afwijkingen. Deze afwijkingen kunnen velerlei oorzaken hebben zoals afleesonzekerheden, onzekerheden van het apparaat en invloed van de omgevingstemperatuur. Als we bijvoorbeeld een titrimetrische bepaling doen, dan hebben we o.a. te maken met onnauwkeurigheden ten gevolge van het pipetteren, buretteren (aflezen begin- en eindstand) en het stellen van de titreervloeistof.
In dit hoofdstuk zullen we aangeven hoe we bij het analyseren en rapporteren van meetgegevens met deze afwijkingen rekening dienen te houden. Door een goede rapportage bereiken we dat onze inspanningen om een chemisch experiment te doen resulteren in goede en bruikbare gegevens. Statistische methoden spelen hierbij een belangrijke rol. In hoofdstuk 2 gaan we hier verder op in via het onderwerp spreidingsvoortplanting.

Kernbegrippen van dit hoofdstuk:

- afwijkingen
  - systematisch
  - toevallig
  - relatief
  - absoluut
- uitbijter
- toets van Dixon
- locatiekentallen
  - gemiddelde
  - mediaan
- spreidingskentallen
  - standaardafwijking
  - variantie
  - variatiecoëfficiënt CV
  - bereik
  - interquartielafstand IQR
  - gemiddelde absolute afwijking MAD
- grafische weergaven
- strooidiagram
- lijndiagram
- Box-and-Whisker plot
- density trace
- normale verdeling
- normal probability plot
- toets van Shapiro-Wilks
- betrouwbaarheidsintervallen voor gemiddelde en variantie
- hypothese toetsen

1.1 Soorten afwijkingen

Hieronder zijn in een figuur de waarden van metingen weergegeven van een drietal experimenten. Het betreft hier herhaalde metingen. De echte waarde is 5.



Figuur 1.1: Drie experimenten

We zien in experiment 1 dat de waarden redelijk netjes verspreid rond de werkelijke waarde 5 liggen. De gemiddelde waarde ligt niet ver van de echte waarde 5 af. In experiment 2 is ook sprake van een spreiding van de resultaten maar met dit verschil dat het gemiddelde hiervan sterk verschilt van de werkelijke waarde 5 (let op de verschillende schalen op de horizontale as). De waarnemingen lijken systematisch naar rechts verschoven te zijn. In experiment 3 zien we metingen die hetzelfde gedrag vertonen als in experiment 1, behalve één waarneming rond 4,2 die duidelijk afwijkt van de overige metingen. Om beter over zulke situaties te kunnen praten en daarna onderbouwde kwantitatieve analyses te kunnen uitvoeren, gaan we nu dieper in op verschillende soorten afwijkingen in meetgegevens.

We onderscheiden drie soorten afwijkingen: toevallige afwijkingen, systematische afwijkingen en uitbijters.

Toevallige of statistische afwijkingen (Engels: indeterminate of random errors)

Dit zijn afwijkingen die zowel positief als negatief kunnen zijn. Deze treden op bijvoorbeeld bij het aflezen van buret, pipet, balans, etc. Ook temperatuursfluctuaties hebben invloed op het volume van pipet en buret, op de viscositeit van een vloeistof en op de werking van een balans. Toevallige afwijkingen zijn dus altijd aanwezig, en we kunnen hiervoor dan ook niet direct corrigeren. Door een meting meerdere malen uit te voeren en dan de gemiddelde uitkomst te nemen, vallen positieve en negatieve toevallige afwijkingen geheel of gedeeltelijk weg. Op deze manier kunnen we het effect van toevallige afwijkingen op een statistische manier verminderen.

Systematische afwijkingen (Engels: determinate of systematic errors)

Dit zijn afwijkingen die zich grotendeels in één richting manifesteren, of steeds positief of steeds negatief. Een voorbeeld is een pipet van 25,00 ml die na ijking 24,90 ml blijkt te zijn. In experiment 2 is ook sprake van een of meer systematische afwijkingen, omdat de meetwaarden allemaal groter (veel) groter zijn dan de werkelijke waarde. In het algemeen kunnen systematische afwijkingen geëlimineerd worden door ijking van meetapparatuur.

Uitbijters (Engels: outliers)

Afwijkingen kunnen ook ontstaan als gevolg van slordigheden, zoals overschrijffouten, rekenfouten, afleesfouten, gebruik van verkeerde hoeveelheden, etc. Deze grove afwijkingen worden uitbijters genoemd. Experiment 3 in Figuur 1.1 is waarschijnlijk een uitbijter.

1.2 Kentallen en grafische weergaven

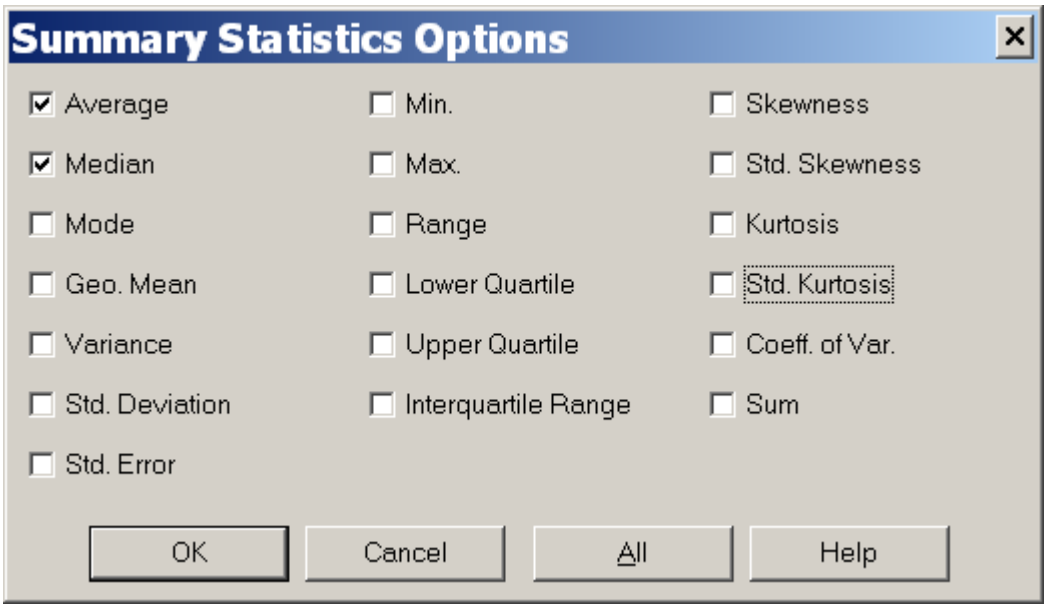
In deze paragraaf gaan we een kwantitatieve onderbouwing geven van de begrippen uit de vorige paragraaf. We gaan er van uit dat we n metingen $x_1, \cdots, x_n$ hebben uitgevoerd, terwijl de werkelijke waarde x_t is. Deze waarde kan onbekend zijn of bekend zijn vanuit theoretische overwegingen of andere, zeer nauwkeurige experimenten.

Het **gemiddelde** $\bar{x}$ is gedefinieerd als $\frac{1}{n}\sum_{i=1}^{n} x_i$.

A. Nauwkeurigheid (Engels accuracy)
De nauwkeurigheid geeft aan hoe dicht de metingen liggen bij de werkelijke of geaccepteerde waarde x_t als we toevallige afwijkingen buiten beschouwing laten. De nauwkeurigheid wordt bepaald door systematische afwijkingen. In experiment 2 van Figuur 1.1 zien we dat er ten gevolge van systematische afwijkingen een grote onnauwkeurigheid in de metingen zit. Om de afwijking van één meting of een serie metingen ten opzichte van de werkelijke waarde aan te geven, gebruiken we:

absolute afwijking E: $e_i = x_i - x_t$ (E kan positief of negatief zijn) of $e = \bar{x} - x_t$

Hierbij is e_i de absolute afwijking van meting i, terwijl e de absolute afwijking van het gemiddelde van de n metingen is. Merk op dat door het gemiddelde te beschouwen het effect van toevallige afwijkingen verminderd wordt. Het gemiddelde is dus een goede indicatie van de locatie van de metingen. Men spreekt van een locatiekental. Merk op dat als er een uitbijter is bij een klein aantal metingen, het gemiddelde een sterk vertekend beeld kan geven. In zulke gevallen zijn andere locatiekentallen zoals de **mediaan** (de middelste waarneming) of **getrimde gemiddelden** (het gemiddelde van de waarnemingen die overblijven als een bepaald percentage van de buitenste waarnemingen aan beide kanten wordt weggelaten). In StatGraphics zijn de meeste kentallen te vinden via de menukeuze **Describe, Numeric Data, One-Variable Analysis**. Aan de linkerkant wordt dan o.a. een veld Summary Statistics geopend. Door met de rechtermuis in dit veld te klikken springt er een venster open waarmee ingesteld kan worden welke kentallen weergegeven worden:



Getrimde gemiddelden zijn te vinden in StatGraphics via de menukeuze **Describe, Numeric Data, Outlier Identification**.
Alléén de absolute afwijking opgeven heeft vaak weinig zin omdat de meetwaarde er niet in verdisconteerd is. Een afwijking van 1 mm bij een meting van de afstand tussen Eindhoven en Maastricht heeft minder consequenties dan een afwijking van 1 mm bij het meten van de elektrodeafstand van een bougie. Daarom is de zogenaamde relatieve afwijking een betere maat voor de nauwkeurigheid van een bepaalde meting.

relatieve afwijking e_{rel} : $e_{\text{rel},i} = \dfrac{x_i - x_t}{x_t}$ of $e_{\text{rel}} = \dfrac{\overline{x} - x_t}{x_t}$

Hierbij is $e_{\text{rel},i}$ de relatieve afwijking van meting i, terwijl e_{rel} slaat op de relatieve afwijking van het gemiddelde.

B. **Precisie** (Engels precision)

De precisie van een serie metingen geeft de mate van overeenkomst aan tussen de meetwaarden onderling. Deze mate van overeenkomst of spreiding wordt bepaald door toevallige afwijkingen (zie Figuur 1.1). Om de precisie van metingen uit te drukken zijn in de statistiek een aantal kentallen beschikbaar.

1. Standaardafwijking s (Engels: sample standard deviation)

De standaardafwijking van n metingen of steekproefstandaardafwijking geeft de spreiding van de data weer door een gemiddelde kwadratische afwijking van het gemiddelde. Dit is de wiskundig optimale manier om spreiding weer te geven als er geen uitbijters zijn. De formule voor de berekening van de standaardafwijking s luidt:

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{\left(\sum_{i=1}^{n} x_i^{\,2}\right) - n\overline{x}^2}{n-1}}\,.$$

De eerste vorm is de officiële definitie die het beste de intuïtie weergeeft. De tweede vorm, die volgt uit de eerste door een kleine berekening, is geschikter voor berekeningen vanwege het vermijden van afrondfouten. De factor *n-1* is nodig om wiskundige redenen die in dit college niet ter zake doen. Pas op voor verwarring met het Engelse begrip **standard error** dat gedefinieerd is als $s_{\overline{x}} = s_x / \sqrt{n}$. Dit begrip geeft de spreiding van het gemiddelde aan. We komen hier later op terug. Soms wordt ook wel de **variatiecoëfficiënt** $CV = s_x / \overline{x}$ gebruikt. De CV

kan ook in procenten worden uitgedrukt. Deze grootheid kan handig zijn om de precisie van experimenten in verschillende meetbereiken eerlijk met elkaar te vergelijken. Een verder voordeel van de CV is dat deze grootheid dimensieloos is.

2. Variantie *v*

De (steekproef)variantie is het kwadraat van de standaardafwijking.

$$v_x = {s_x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} {x_i}^2\right) - n\bar{x}^2\right)$$

De variantie is in de wiskunde vaak wat makkelijker te hanteren dan de standaardafwijking. Een nadeel van de variantie is dat de eenheid van variantie het kwadraat is van de eenheid van de oorspronkelijke waarnemingen.

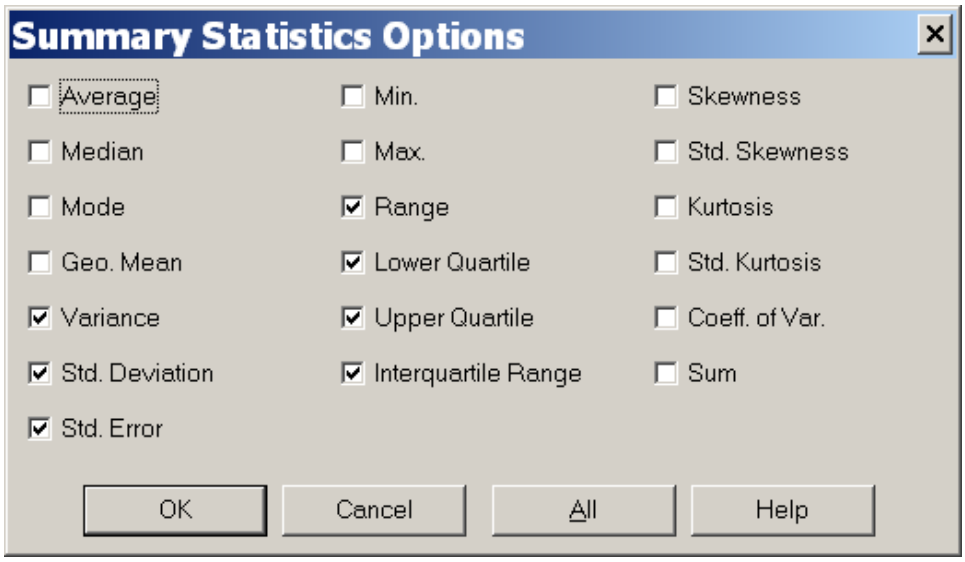
3. Bereik *r* (Engels range)

Het bereik *r* is het verschil tussen de grootste en de kleinste meetwaarde. Dit kental was vooral populair toen er nog geen rekenmachines waren, omdat het gemakkelijk uit te rekenen is. Als het aantal metingen kleiner dan 10 is, voldoet het bereik redelijk goed t.o.v. de standaardafwijking. Als er meer dan 10 metingen zijn, dan is de kans op uitbijters zo groot dat het bereik de spreiding systematisch gaat overschatten.

De bovenstaande 3 kentallen zijn alle gevoelig voor uitbijters. Om de invloed van uitbijters te verminderen, heeft men in de statistiek alternatieve spreidingskentallen ingevoerd. De belangrijkste twee volgen hieronder. Voordeel van deze kentallen is hun kleinere gevoeligheid voor uitbijters waardoor ze uitstekend geschikt zijn om gebruikt te worden in data-analyse. Nadeel is dat de wiskundige theorie zeer moeilijk is, waardoor er geen eenvoudige formules zijn voor gebruik in betrouwbaarheidsintervallen of toetsen (zie later in dit hoofdstuk). Men is hierdoor aangewezen op het gebruik van statistische software.

4. Interquartielafstand *IQR* (Engels interquartile range)

Om dit kental te kunnen definiëren, hebben we eerst een ander kental nodig. Het α%-quantiel is een getal zodanig dat α% van de metingen kleiner is. Er is een precieze definitie die de problemen bij kleine aantallen waarnemingen opvangt door interpolatie. Merk op dat het 50%-quantiel niet anders is dan de bij het onderdeel nauwkeurigheid genoemde mediaan. Het 25%-quantiel wordt ook wel 1e quartiel genoemd, terwijl het 75%-quantiel het derde quartiel genoemd wordt. In StatGraphics worden de namen lower and upper quartile gebruikt. Nu kunnen we ook de naam interquartielafstand begrijpen: *IQR* = 3e quartiel – 1e quartiel.



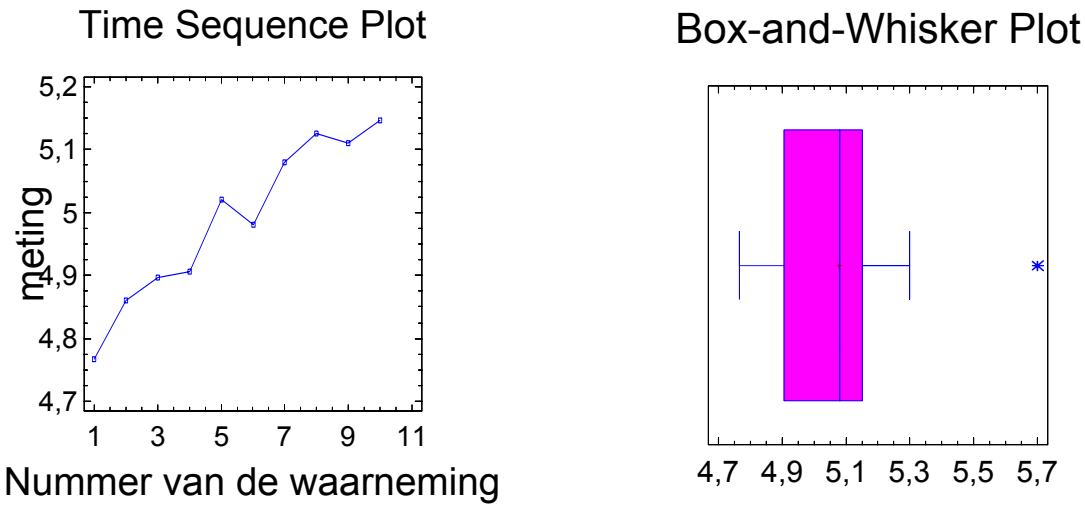
5. Gemiddelde absolute afwijking (Engels: Mean Absolute Deviation *MAD*)

Deze naam laat zien dat (Engelstalige) statistici gevoel voor humor hebben. Dit kental wordt gedefinieerd door $MAD = \frac{1}{n-1}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$. In StatGraphics is het te vinden via de menukeuze **Describe, Numeric Data, Outlier Identification.** De gedachte achter dit kental is dat alle afwijkingen even zwaar gewogen worden. Dit in tegenstelling tot de steekproefstandaardafwijking en de steekproefvariantie, waarbij door het kwadrateren afwijkingen kleiner dan 1 nog kleiner worden en afwijkingen groter dan 1 extra zwaar meegenomen worden. Het laatste heeft natuurlijk tot gevolg dat de steekproefstandaardafwijking en de steekproefvariantie gevoelig zijn voor eventuele uitbijters. De *MAD* is minder gevoelig voor uitbijters. Helaas is de *MAD* wiskundig lastig te hanteren, waardoor men op het gebruik van statistische software is aangewezen.

Alleen kentallen van een serie metingen bestuderen is niet aan te raden. Kentallen dienen samen met grafische weergaven bestudeerd te worden. In de praktijk is het zelfs aan te raden data eerst grafisch te bekijken. In dit dictaat hebben we eerst kentallen behandeld, omdat één van de grafische weergaven niet te begrijpen valt zonder kennis van kentallen. Een bekende uitspraak in dit verband is “Eén plaatje zegt meer dan duizend woorden”. We bekijken de volgende grafische weergaven:

- strooidiagram
- lijndiagram of tijdreeks
- Box-and-Whiskerplot

Het **strooidiagram** hebben wel al gezien in Figuur 1.1. Met behulp hiervan kunnen we al snel een eerste indruk krijgen. Let altijd wel op de schaal. Computerprogramma’s passen vaak de schaal automatisch zodanig aan, dat het plaatje het hele scherm vult. Kleine afwijkingen kunnen dan heel groot lijken! In een **lijndiagram** of **tijdreeksdiagram** worden de waarnemingen uitgezet in de volgorde waarin ze gemeten zijn. Dit kan soms nuttige informatie geven. Zo hebben sommige meetapparaten last van opwarmingverschijnselen (een voorbeeld hiervan zijn elektrische massabalansen). De waarnemingen in de tijd laten dan een duidelijke trend zien (zie bijvoorbeeld Figuur 1.2) In StatGraphics kunnen we beide diagrammen vinden via het menu **Plot, Scatterplots, Univariate Plot.**



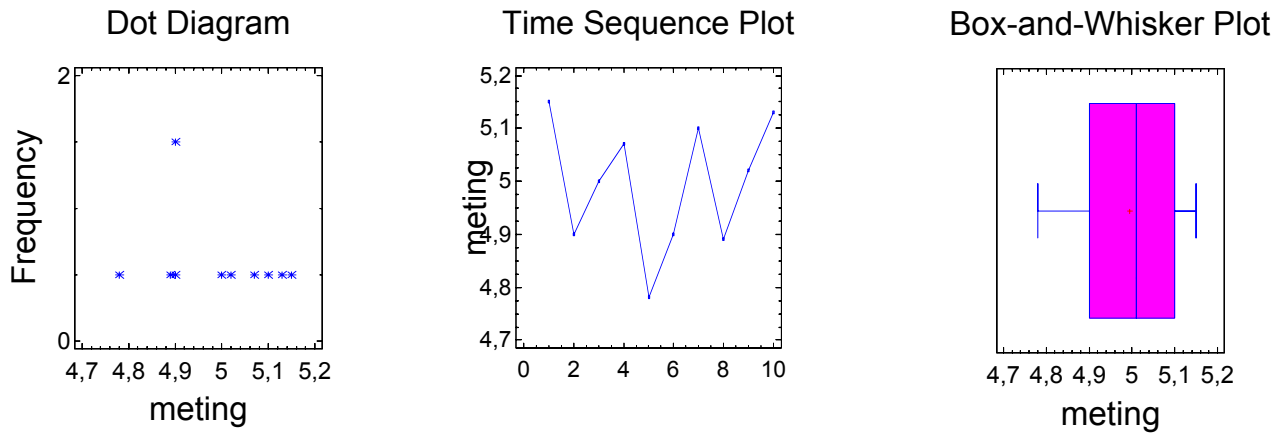
Figuur 1.2: Tijdreeksdiagram en Box-and-Whiskerplot

De Box-and-Whiskerplot (in het Nederlands soms snorrendoos genoemd) is een eenvoudig, maar effectief hulpmiddel om snel quartielen en uitbijters van waarnemingen te visualiseren. De linkerkant van de doos is het eerste quartiel, de rechterkant van de doos het derde quantiel. De streep tussen beide quartielen is de mediaan (het tweede quartiel). Ligt de mediaan niet in het midden van de doos, dan is dit een aanwijzing dat de data niet symmetrisch verdeeld zijn. Omgekeerd mag men niet meteen concluderen dat de waarnemingen symmetrisch verdeeld zijn als de mediaan in het midden van de doos ligt. De Box-and-Whiskerplot bevat ook twee horizontale lijnen die beginnen bij het eerste resp. derde quartiel. De lengte van deze lijnen is 1 ½ keer de interquartielafstand. Waarnemingen die hierbuiten vallen zijn uitbijters. In Figuur 1.2 zien we een uitbijter met een waarde van ongeveer 5,7. Verder kunnen we zien dat er veel waarnemingen liggen tussen 5,1 en 5,15.
In StatGraphics kunnen deze plots gemaakt worden via het menu Plot, Exploratory Plots, Box-and-Whisker Plot of via Describe, Numeric Data, One-variable Analysis.

We illustreren nu de grafische weergaven en de behandelde kentallen aan de hand van een getallenvoorbeeld. Stel we hebben de volgende 10 waarnemingen:

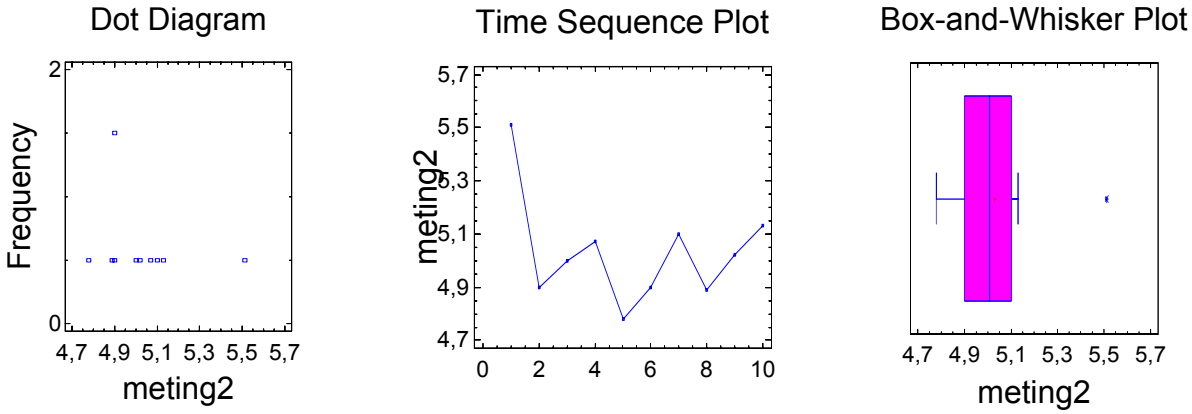
5,15	4,90	5,00	5,07	4,78	4,90	5,10	4,89	5,02	5,13

Hieronder volgen grafische weergaven van deze 10 waarnemingen, gevolgd door een overzicht van de belangrijkste kentallen. De dataset is opgeslagen onder de name simpeleda.sf3. Het betreft hier de kolom meting.



Summary Statistics for meting	Summary Statistics for meting
Count = 10 Average = 4,994 Median = 5,01 Lower quartile = 4,9 Upper quartile = 5,1	Count = 10 Variance = 0,0149822 Standard deviation = 0,122402 Standard error = 0,0387069 Range = 0,37 Interquartile range = 0,2

Als we de waarneming 5,15 per ongeluk als 5,51 wordt genoteerd of waargenomen, dan krijgen we onderstaande plots en kentallen. Het is nuttig om te zien welke kentallen veranderen en hoe de grafische weergaven veranderen.



Figuur 1.3: Grafische weergaven van data met uitschieter 5,51

Summary Statistics for meting2	Summary Statistics for meting2
Count = 10 Average = 5,03 Median = 5,01 Lower quartile = 4,9 Upper quartile = 5,1	Count = 10 Variance = 0,0404222 Standard deviation = 0,201053 Standard error = 0,0635785 Range = 0,73 Interquartile range = 0,2

1.3 Kansrekening

We hebben gezien dat toevallige afwijkingen zowel positief als negatief kunnen zijn. Dat wordt namelijk bepaald door het toeval. De wiskundige theorie die zich bezig houdt met toeval heet kansrekening (Engels: probability theory). We hebben enige kennis van deze theorie nodig om kwantitatieve onderbouwingen kunnen te geven van de eerder behandelde begrippen. Laat X de uitkomst van een meting zijn. Een wiskundig model voor toevallige uitkomsten van een meting leggen we vast door de(cumulatieve) verdelingsfunctie van X te geven:

$$F(t) = P(X \le t).$$

In de kansrekening wordt X een stochast genoemd. Aangezien metingen in de chemie meestal continu zijn (binnen een bepaald bereik kan elke waarde aangenomen worden), geldt dat $P(X=t) = 0$ voor elke afzonderlijke waarde t. Dit verklaart bovengenoemde keuze voor de verdelingsfunctie om de uitkomsten van een stochast te beschrijven. In de praktijk is het vaak handig om naast de verdelingsfunctie ook de afgeleide te beschouwen. Deze afgeleide heet de **dichtheidsfunctie** (afgekort: **dichtheid**):

$$f(t) = \frac{d}{dt} F(t).$$

Indien men de dichtheid kent, kan de verdelingsfunctie terugvinden door te integreren: $F(t) = \int_{-\infty}^{t} f(x)dx$. Een grafische interpretatie is dat men kansen kan vinden als oppervlakte onder de dichtheid.

Er zijn veel kansverdelingen bekend. Een overzicht is te vinden in het Statistisch Compendium. Het blijkt echter dat in veel gevallen toevallige afwijkingen met een zogenaamde normale verdeling (ook wel Gaussverdeling genoemd) beschreven kunnen worden. De verklaring hiervoor is dat de som van een groot aantal toevallige afwijkingen zich, ongeacht de verdeling van deze afwijkingen, bijna gedraagt als een toevallige afwijking met een normale verdeling. De precieze wiskundige formulering van dit feit heet **Centrale Limietstelling**. Een mooie demonstratie van de Centrale Limietstelling is te zien op http://www.maths.soton.ac.uk/~sml/ma120/SamplingApplet.html . De dichtheid van een normale verdeling heeft een bekende klokvorm met als formule $f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(t-\mu)^2}{2\sigma^2}\right)$.

Merk op dat de normale verdeling afhangt van twee parameters. De parameter μ is de verwachting van de verdeling. Dit is een theoretisch gemiddelde waarde. Aangezien de dichtheid van een normale verdeling symmetrisch rond μ is, verwachten we evenveel en even grote waarden groter en kleiner dan μ. De parameter μ wordt om deze reden een locatieparameter genoemd. De parameter σ is een maat voor de spreiding. Om wiskundige redenen is het gebruikelijk σ^2 i.p.v. σ als parameter te beschouwen. Let bij het gebruik van software altijd op de gebruikte conventie om een normale verdeling te specificeren, m.a.w. wordt σ^2 of σ gebruikt. Een grote waarde van σ leidt tot een grote kans op uitkomsten die ver weg liggen van μ. Om beter vertrouwd te raken met deze begrippen, zijn de volgende Java applets beschikbaar:

http://www.win.tue.nl/~marko/statApplets/functionPlots.html en http://www-stat.stanford.edu/~naras/jsm/NormalDensity/NormalDensity.html .

Hoe moeten we nu zo'n kromme interpreteren? Het totale oppervlak binnen deze kromme is derhalve 1 of 100%. D.w.z. elke nieuwe meting valt met een waarschijnlijkheid van 100% in dit gebied. Hieronder volgen enkele waarden van mogelijke oppervlakken:
Het oppervlak binnen:

$\mu \pm 0{,}67\sigma$	is	0,500		
$\mu \pm 1{,}00\sigma$	is	0,683		
$\mu \pm 1{,}645\sigma$	is	0,975		
$\mu \pm 1{,}96\sigma$	is	0,950		
$\mu \pm 2{,}00\sigma$	is	0,954		
$\mu \pm 2{,}33\sigma$	is	0,980		
$\mu \pm 2{,}58\sigma$	is	0,990		
$\mu \pm 3{,}00\sigma$	is	0,997		

Tabel 1.1: Overschrijdingskansen bij een normale verdeling

Andere waarden kan men vinden in tabel 9.1 van het Statistisch Compendium. Hierbij dient men te weten dat de normale verdeling met $\mu=0$ en $\sigma^2=1$ de standaardnormale verdeling heet. De standaardnormale verdeling wordt vaak aangegeven met de letter Z. Als X normaal verdeeld is met parameters μ en σ^2, dan is $(X-\mu)/\sigma$ standaardnormaal verdeeld. De overgang van X naar $(X-\mu)/\sigma$ heet standaardiseren. Een applet die dit illustreert is te vinden op http://psych.colorado.edu/~mcclella/java/normal/normz.html .

Voorbeeld: stel de uitkomst X van een meting is normaal verdeeld met $\mu=3$ en $\sigma^2=4$. Wat is de kans dat de uitkomst van een meting kleiner is dan 6,2?

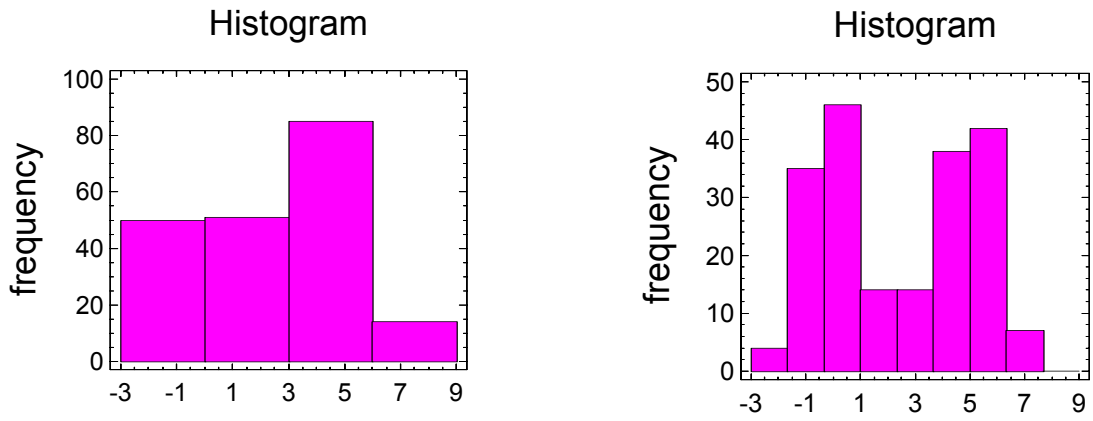
Oplossing: $P(X<6{,}2)=P\left(\frac{X-\mu}{\sigma}<\frac{6{,}2-\mu}{\sigma}\right)=P\left(Z<\frac{6{,}2-3}{2}\right)=P(Z<1{,}6)=0{,}9452.$

Dit soort berekeningen kan ook met StatGraphics uitgevoerd worden via het menu **Plot, Probability Distributions**. Bij **Tabular Options** moet men dan kijken bij Cumulative Distribution. Er kunnen 5 verschillende verdelingen tegelijk bekeken worden door in het venster linksboven met de rechtermuis te klikken en dan Pane Options te kiezen. Wie er meteen een mooi plaatje bij wil zien, kan terecht op http://psych.colorado.edu/~mcclella/java/normal/normz.html .

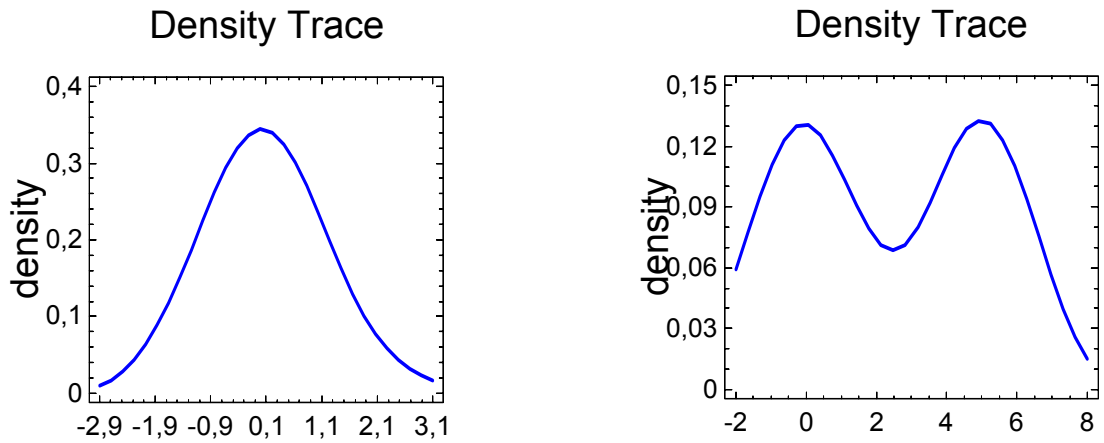
1.4 Toetsen van normaliteit

De normale verdeling neemt een belangrijke plaats in de statistiek. Vanwege de Centrale Limietstelling zijn waarnemingen vaak (bijna) normaal verdeeld. Anderzijds heeft de normale verdeling prettige wiskundige eigenschappen. Als gevolg hiervan zijn veel statistische methoden gebaseerd op de (soms impliciete) aanname dat de data normaal verdeeld zijn. Indien zulke methoden gebruikt worden voor data die niet normaal verdeeld is, dan is er geen enkele garantie dat de uitkomsten van deze statistische methoden betrouwbaar zijn. De verantwoordelijkheid voor het gebruik van statistische methoden ligt bij de gebruiker! Een andere praktische reden om normaliteit te onderzoeken is dat het vinden van afwijkingen van normaliteit inzicht in de experimenteeromstandigheden oplevert. Zo kan een afwijkingen van normaliteit veroorzaakt worden door het feit dat we onbedoeld twee foutenbronnen hebben i.p.v. één. Hier komen we nog op terug.

Het onderzoeken van normaliteit (d.w.z. onderzoeken of waarnemingen normaal verdeeld zijn), is een speciale vorm van de data-analyse die we in de voorgaande paragrafen hebben uitgevoerd. Ook hier is zowel een grafisch aspect als een getalsmatig aspect. Zoals we in paragraaf 1.3 gezien hebben is de dichtheid van de normale verdeling een symmetrische, klokvormige kromme. Om dit te controleren wordt vaak een histogram gemaakt. Een histogram wordt gemaakt door het bereik van de uitkomsten in een aantal even brede vakken (officiële naam: klassen, Engels: bins) te verdelen en dan te tellen hoeveel waarnemingen in elke klasse vallen. Het nadeel van deze methode is dat de vorm van een histogram sterk afhangt van de gekozen klassenbreedte. De onderstaande twee histogrammen zijn gemaakt van dezelfde data set. Wie zelf wil experimenteren met de invloed van de klassenbreedte op de vorm van het histogram, kan terecht op: http://www.stat.sc.edu/~west/javahtml/Histogram.html .



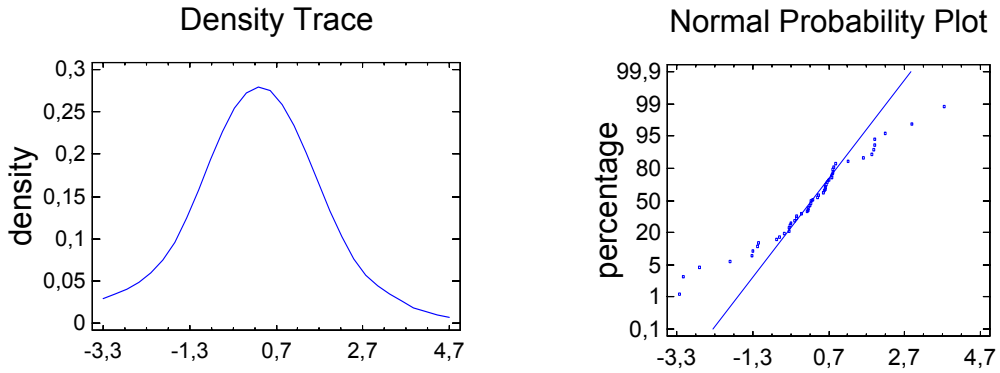
Figuur 1.4: Twee histogrammen van dezelfde dataset



Figuur 1.5 Density trace: normaal verdeelde data en tweetoppige data

Een betere grafische weergave is de zogenaamde **density trace** (ook wel naive density estimator genoemd), een soort glijdend histogram. M.a.w., bij een density trace is elk punt middelpunt van een klasse; de functiewaarde in dat punt is het aantal punten in die klasse gedeeld door het totaal aantal waarnemingen maal de klassenbreedte. Bij een histogram is er een vast aantal disjuncte klassen, bij een density trace zijn er dus oneindig veel elkaar overlappende klassen. In StatGraphics kan men een density trace vinden via **Describe, Distributions, Distribution Fitting (Uncensored Data)** of **Describe, Numeric Data, One-**

variable analysis . In Figuur 1.5 zien we links de density trace van normaal verdeelde data, terwijl we rechts zien dat er een tweetoppige verdeling is. Vaak duidt dit op meerdere foutenbronnen. In plaats van een tweetoppige kansverdeling te modelleren, is het praktischer eerst uit te zoeken of de experimentele omstandigheden wel kloppen. Wellicht zijn gegevens van twee verschillende experimenten abusievelijk samengevoegd. Een andere mogelijkheid is een onbedoelde foutenbron in de meetopstelling.



Figuur 1.6: schijnbaar normaal verdeelde data

Ook nuttig is de **quantile-quantile plot** (wat in dit geval overeenkomt met de bekende **normal probability plot**). Dit is een weergave van de verdelingsfunctie waarbij de assen zo uitgerekt zijn dat een data set die normaal verdeeld is, een vrijwel rechte lijn te zien geeft. Ook deze plot kan in StatGraphics gevonden worden via **Describe, Distributions, Distribution Fitting (Uncensored Data)** of **Describe, Numeric Data, One-variable analysis** of via **Describe, Numeric Data, Outlier Identification**. Een voorbeeld is te zien in Figuur 1.6 . De density trace laat een redelijk symmetrisch verdeelde klokkromme zien. De asymmetrie in dit soort gevallen komt meestal door een relatief kleine steekproef: de kans op grote waarden bij een normale verdeling is heel klein. De normal probability plot laat echter duidelijk zien dat de klokvormige kromme toch niet goed gevormd is om van een normale verdeling af te komen. Zowel bij grote als kleine waarden wijken de waarden in de normal probability plot af van de rechte lijn. De verklaring is dat er in beide staarten van de verdeling teveel waarnemingen zijn t.o.v. een normale verdeling. De data is dus niet normaal verdeeld.

Na een eerste grafische controle (die natuurlijk subjectief is) kan een objectieve controle uitgevoerd worden via een statistische toets. De toets van Shapiro-Wilks is een uitstekende toets. In StatGraphics is deze toets te vinden via **Describe, Distributions, Distribution Fitting (Uncensored Data)** of **Describe, Numeric Data, One-variable analysis** of **Describe, Numeric Data, Outlier Identification**. Men dient dat bij **Tabular Options** (het gele icoontje) de optie **Tests for Normality** aan te vinken. De overige toetsen zijn niet specifiek bedoeld om normaliteit mee te toetsen en dienen daarom niet gebruikt te worden. In het bijzonder is de tekst van de StatAdvisor in StatGraphics verwarrend. De toets werkt als volgt:

- is de *p*-waarde kleiner dan of gelijk aan 0,01, dan is de data niet normaal verdeeld
- is de *p*-waarde groter dan 0,01, dan is er geen reden om aan normaliteit van de data te twijfelen.

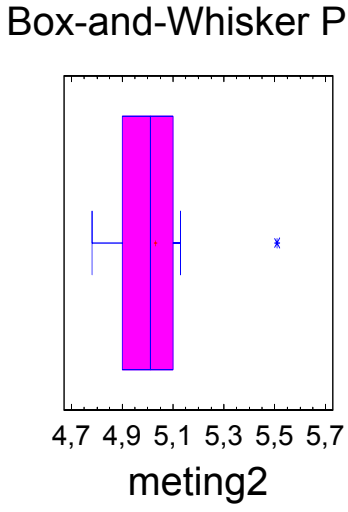
Meestal wordt bij toetsen niet 0,01 maar 0,05 als grens gebruikt. In dit geval is 0,01 aan te raden, omdat de statistische procedures die we in dit hoofdstuk gebruiken niet al te gevoelig zijn voor lichte afwijkingen van normaliteit. Door 0,01 te kiezen voorkomen dat we dat onno-

dig aan normaliteit getwijfeld wordt. Een typische uitdraai van een toets op normaliteit is te vinden in Figuur 1.7.

```
Tests for Normality for bimodalnormal

Computed Chi-Square goodness-of-fit statistic = 129,6
P-Value = 1,12133E-14

Shapiro-Wilks W statistic = 0,880953
P-Value = 0,0

Z score for skewness = 0,0255352
P-Value = 0,979622

Z score for kurtosis = 53,211
P-Value = 0,0
```

Figuur 1.7: uitdraai van normaliteit toetsen in StatGraphics

Normaliteit kan bij kleine aantallen waarnemingen verstoord worden door één enkele waarneming. Zo'n waarneming is vaak te zien in bovengenoemde weergaven. In de praktijk moet zo'n enkele waarneming onderzocht worden en weggelaten als er iets afwijkends geconstateerd wordt. Om objectief te toetsen of één of meerdere waarnemingen uitbijters zijn, kan men de **toets van Dixon** gebruiken.
In StatGraphics kan men de toets van Dixon vinden via **Describe, Numeric Data, Outlier Identification**. Het is belangrijk te beseffen dat deze toets gebaseerd is op de aanname dat de waarnemingen normaal verdeeld zijn (dit kan ook via dit menu door een extra optie aan te vinken bij **Tabular Options**. De toets van Dixon mag dus pas gebruikt worden, nadat we gecontroleerd hebben dat de data normaal verdeeld is. Als voorbeeld gebruiken we de data uit Figuur 1.3. Zowel de Box-and-Whisker plot als de toets van Dixon geven aan dat er een uitbijter is aan de rechterkant, d.w.z. een waarneming die te groot is vergeleken met de overige waarnemingen. De toets van Dixon kan gewoon gebruikt worden met een significantie van 5%. M.a.w., uitkomsten met een p-waarde kleiner dan 0,05 geven aanleiding tot de conclusie dat er één of meerdere uitbijters zijn.




```
Dixon's Test (assumes normality)
-----------------------------------------------------
                            Statistic       5% Test
1 outlier on right          0,612903        Significant
1 outlier on left           0,314286        Not sig.
2 outliers on right         0,66129         Significant
2 outliers on left          0,342857        Not sig.
1 outlier on either side    0,520548        Significant
```

Figuur 1.8: Box-and-Whisker plot en toets van Dixon voor data met uitbijter

1.5 Betrouwbaarheidsintervallen en toetsen

In deze paragraaf nemen we aan dat onze waarnemingen $X_1, \cdots, X_n$ normaal verdeeld zijn met verwachting μ en variantie σ^2. In paragraaf 1.2 zijn we de kentallen (steekproef)gemiddelde $\bar{x}$ en steekproefvariantie s^2 al tegen gekomen. Door middel van deze kentallen proberen we zo goed mogelijk de onbekende parameters μ en σ^2 te weten te komen. In de statistiek spreekt men van **schatters**. In de statistiek spreekt men van schatten als we een nauwkeurig vastgestelde procedure hebben om vanuit de data tot een getal te komen dat zo goed mogelijk een parameter van een kansverdeling benadert. Dit spraakgebruik verschilt dus van het dagelijkse spraakgebruik, waar schatten een veel vagere betekenis heeft. Men kan wiskundig bewijzen dat het gemiddelde en de steekproefvariantie zuivere schatters zijn van μ en σ^2, d.w.z. gemiddeld geven deze schatters de juiste uitkomst. De precieze uitkomst verschilt echter meestal per reeks waarnemingen! Zie bijvoorbeeld http://statwww.berkeley.edu/users/stark/Java/Ci.htm voor een simulatie van dit verschijnsel.
Het gemiddelde is een voorbeeld van een zogenaamde **puntschatter**. Dit is een schatter dat als uitkomst één enkel getal geeft. Een nadeel hiervan is dat het niet aangeeft hoe betrouwbaar deze uitkomst is. Een gemiddelde gebaseerd op 2 waarnemingen is natuurlijk veel onbetrouwbaarder dan een gemiddelde gebaseerd op 20 waarnemingen. Om dit verschil in betrouwbaarheid zichtbaar te maken, gebruikt met **betrouwbaarheidsintervallen**. We beginnen met een voorbeeld. Een 95 % betrouwbaarheidsinterval voor μ is een interval dat met 95% kans de echte, onbekende waarde van μ bevat. Als we waarnemingen hebben en het interval uitgerekend hebben, dan ligt de echte waarde μ natuurlijk of wel of niet in het interval. Met 95% kans wordt bedoeld dat voordat er waarnemingen zijn gedaan er 95% kans is dat dit interval de juiste waarde bevat. Net als hierboven is het weer nuttig om het interval als een procedure te zien die met een bepaalde kans een gewenst resultaat heeft.
Er bestaan expliciete formules voor allerlei betrouwbaarheidsintervallen. Hiervoor verwijzen we naar het Statistisch Compendium. We beperken ons hier tot een betrouwbaarheidsinterval voor μ gebaseerd op een normale verdeling. Als we nu aannemen dat de parameter σ^2 bekend is uit eerdere metingen, dan wordt een 100(1-α)%-betrouwbaarheidsinterval voor μ gegeven door:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

In het geval van een 95%-betrouwbaarheidsinterval is α=0,05 en $z_{\alpha/2} = 1{,}96$. Het is het nuttig om even stil te staan bij deze formule. We zien dat het gemiddelde gebruikt wordt als spil van het betrouwbaarheidsinterval en dat het interval symmetrisch is. Als we de betrouwbaarheid verhogen, dan verwachten we een breder interval. Immers, er moet een grotere kans zijn dat het interval de echte waarde μ bevat. Als we bijvoorbeeld een 99%-betrouwbaarheidsinterval wensen, dan moeten we $z_{0,005}$ = 2,57 gebruiken (zie het Statistisch Compendium of Tabel 1.1). Verder zien we dat vergroting van n (het aantal waarnemingen) leidt tot een smaller interval. Dit klopt met onze intuïtie: hoe meer waarnemingen, des te nauwkeuriger onze resultaten. Zoals eerder opgemerkt, komt dit doordat toevallige afwijkingen (gedeeltelijk) tegen elkaar wegvallen door te middelen. Omgekeerd kunnen we uitrekenen hoeveel waarnemingen we moeten doen om een betrouwbaarheidsinterval van een gegeven breedte te krijgen.We zullen in de volgende paragraaf zien dat $\sigma/\sqrt{n}$ de standaardafwijking is van het steekproefgemiddelde en dat het gemiddelde normaal verdeeld is met parameters μ en σ^2/n. M.a.w., de spreiding van de gemiddelde waarneming is $\sqrt{n}$ keer kleiner dan de spreiding van een individuele waarneming. Aangezien de wortelfunctie een kromming heeft bij 20, is het effect van vergroting van het aantal waarnemingen het sterkst bij reeksen waarnemingen met lengten

onder de 20. Het kental standard error uit paragraaf 1.2 geeft dus aan wat de spreiding van het gemiddelde is. Dit onderscheid wordt vaak vergeten met ernstige gevolgen. Pas dus op als iemand een waarde voor σ of *s* rapporteert.

Indien de variantie σ^2 onbekend is (wat meestal het geval is), zijn er andere formules voor de betrouwbaarheidsintervallen. Het gemiddelde heeft dan geen normale verdeling, maar een zogenaamde Student-*t*-verdeling. In dit college gaan we hier niet verder op in. Formules zijn te vinden in het Statistisch Compendium. Het betrouwbaarheidsinterval dat StatGraphics uitrekent via **Describe, One-Variable Analysis** is een betrouwbaarheidsinterval waarbij aangenomen wordt dat σ^2 onbekend is. Voor details verwijzen we naar het StatGraphics dictaat paragraaf 2.2. In paragraaf 2.4.1 van het StatGraphics dictaat staat hoe we StatGraphics kunnen laten uitrekenen wat het minimale aantal waarnemingen is om tot een betrouwbaarheidsinterval van een gegeven breedte te komen via **Describe, Sample Size Determination**.

Er zijn ook betrouwbaarheidsintervallen voor σ^2. Formules hiervoor zijn te vinden in het Statistisch Compendium. StatGraphics rekent deze intervallen uit via het menu **Describe, One-Variable Analysis**.

Als voorbeeld nemen we de data set uit paragraaf 1.2.

```
Summary Statistics for meting

Count = 10
Average = 4,994
Median = 5,01
Variance = 0,0149822
Standard deviation = 0,122402
Standard error = 0,0387069
Minimum = 4,78
Maximum = 5,15
Range = 0,37
Interquartile range = 0,2
Confidence Intervals for meting
------------------------------
95,0% confidence interval for mean: 4,994 +/- 0,0875612   [4,90644;5,08156]
95,0% confidence interval for standard deviation: [0,0841923;0,223458]
Confidence Intervals for meting
------------------------------
99,0% confidence interval for mean: 4,994 +/- 0,125791   [4,86821;5,11979]
99,0% confidence interval for standard deviation: [0,0756051;0,278784]
```

Tenslotte besteden we kort aandacht aan toetsen in relatie tot het vaststellen van systematische afwijkingen. Met behulp van een statistische toets kunnen we vaststellen of bijvoorbeeld de waarnemingen systematisch afwijken van de echte waarde 5. Dit kan in StatGraphics via het menu **Describe, One-Variable Analysis** door bij **Tabular Options** de optie **Hypothesis Tests** aan te vinken. Voor details verwijzen we naar paragraaf 2.2 van het StatGraphics dictaat. We merken op dat de *t*-toets net als de betrouwbaarheidsintervallen gebaseerd op normaliteit.

```
Hypothesis Tests for meting

Sample mean = 4,994
Sample median = 5,01
```

t-test

Null hypothesis: mean = 5,0
Alternative: not equal

Computed t statistic = -0,155011
P-Value = 0,880233

Do not reject the null hypothesis for alpha = 0,05.