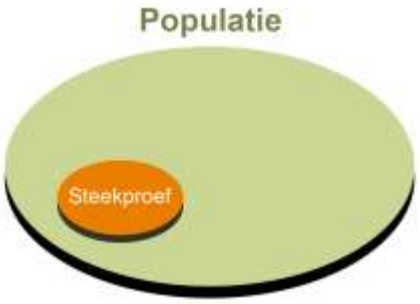
1. Beschrijvende statistiek

In de beschrijvende statistiek zal men gegevens overzichtelijk weergeven in tabellen of grafische voorstellingen en samenvatten d.m.v. enkele kengetallen.

1.1. Populatie en steekproef

- Een statistisch onderzoek start met het verzamelen van gegevens. De totale groep waarvoor we bepaalde eigenschappen willen te weten komen, noemen we de populatie. Vaak is de populatie te groot waardoor het fysiek onmogelijk of te duur is om de informatie te verzamelen voor alle elementen van de populatie. Dan kiest men voor een kleinere groep, namelijk een deelverzameling van de populatie. Die deelverzameling van de populatie waarvoor de informatie effectief verzameld wordt, noemen we de steekproef.



  Voorbeeld
  In een onderzoek naar kijkcijfers of politieke voorkeuren kunnen bezwaarlijk alle Belgen boven de achttien worden betrokken. Dat is ook niet nodig: een representatieve steekproef van 1000 tot 2000 personen geeft in de meeste gevallen voldoende valide en betrouwbare resultaten.

  Voorbeeld: zie bijlage: politieke peilingen

- Een belangrijke taak voor de statistiek is om uit een steekproef informatie af te leiden voor de volledige populatie. Daartoe is het wenselijk dat de steekproef representatief, betrouwbaar en aselect is.

  - Een representatieve steekproef is een afspiegeling van de populatie waaruit hij is getrokken. De waarnemingen in de steekproef representeren (vertegenwoordigen) de eigenschappen van alle elementen in de hele populatie. Als dit het geval is, zijn de resultaten van het steekproefonderzoek generaliseerbaar naar de populatie. Het onderzoek is dan voor wat betreft de steekproeftrekking valide.

  - Behalve representatief moet een steekproef ook betrouwbaar zijn. Een steekproef is betrouwbaar, als bij herhaling van de steekproeftrekking (globaal) dezelfde resultaten worden verkregen. De variabiliteit van de uitkomsten in achtereenvolgende steekproeven hangt onder andere af van de variabiliteit in de populatie en van de betrouwbaarheid van de waarnemingsmethode. Voor wat de steekproeftrekking betreft wordt de betrouwbaarheid van een steekproef bepaald door de steekproefomvang.

- Een steekproef (of deelsteekproef bij samengestelde steekproeven) behoort aselect te zijn, dat wil zeggen, alle elementen in de (deel)populatie moeten een gelijke kans hebben om in de steekproef terecht te komen.
- Veel zal hierbij afhangen van de respons, hoeveel van de elementen van de steekproef zullen reageren!

1.2. Variabele

Door waarnemen wordt een eigenschap van een object gemeten of een kenmerk van een object vastgesteld. Eigenschappen of kenmerken worden ook variabelen genoemd. Een variabele kan diverse waarden aannemen.

Voorbeeld

Populatie: alle 17-jarigen in Vlaanderen
steekproef: De 17-jarigen in deze klasgroep
mogelijke variabelen: lengte, schoenmaat, kleur van broek, politieke voorkeur, geboortemaand, lievelingsvak, gewicht,…

Variabelen worden ingedeeld naar type variabele en meetniveau op basis van de waarden die zij kunnen aannemen.

1.2.1. Type variabele

- Het belangrijkste onderscheid naar type is dat tussen continue en discrete variabelen.

  - Een continue variabele kan in een bepaald interval iedere waarde aannemen.
  - Een discrete variabele heeft slechts gehele getallen of klassen als mogelijke waarden.

- Een andere indeling is die in kwalitatieve en kwantitatieve variabelen.

  - De waarden van kwantitatieve variabelen worden verkregen door te tellen, te meten, te wegen,… Met de waarden van kwantitatieve variabelen kan worden gerekend.
  - Een kwalitatieve variabele heeft een beperkt aantal mogelijke waarden. Met de waarden van kwalitatieve variabelen kan niet worden gerekend.

Voorbeeld

- Meting van tijd en lengte zijn continue variabelen, een telling is een discrete variabele evenals de bepaling van de bloedgroepen A, B, AB en 0.
- De meting van tijd en lengte en een telling zijn kwantitatieve variabelen, de bepaling van de bloedgroepen is een kwalitatieve variabele.

1.2.2. Meetniveau

De indeling van de variabelen naar meetniveau is vooral gericht op de *bewerkingen* die met de waarden van de variabelen mogen worden uitgevoerd.

Variabelen kunnen op vier niveaus worden gemeten: op *nominaal*, *ordinaal*, *interval*- en *rationiveau*.

We zeggen: "deze variabele is op nominaal niveau gemeten" of "het meetniveau van deze variabele is nominaal" of "deze variabele heeft een nominaal meetniveau."

- Op nominaal niveau gemeten variabelen zijn kwalitatieve variabelen, waarvan de categorieën niet in een vaste of zinvolle volgorde zijn te plaatsen.

  Voorbeeld

  'Bloedgroep' met als waarden A, B, AB en O of 'geslacht' met als waarden 'mannelijk' en 'vrouwelijk'.

  Deze kwalitatieve variabelen zijn discreet. Je kunt aan de waarden getallen koppelen, zoals 'mannelijk' = 0 en 'vrouwelijk' = 1. De getallen geven hier geen volgorde weer, ze dienen alleen ter onderscheiding van de twee niveaus van de variabele 'geslacht'.

  Het gemiddelde van een op nominaal niveau gemeten variabele heeft geen betekenis, evenmin zijn verschillen en verhoudingen hier betekenisvol.

- Op ordinaal niveau gemeten variabelen zijn kwalitatieve variabelen, waarvan de categorieën wel in een vaste en zinvolle volgorde zijn te plaatsen.

  Voorbeeld

  Een stelling in een enquête, zoals 'speelfilms op tv onderbreken voor het uitzenden van reclamespots dient verboden te worden', met als antwoordcategorieën: 'geheel mee eens', 'mee eens', 'geen mening', 'niet mee eens' en 'geheel niet mee eens'. Deze antwoordcategorieën worden vaak scores genoemd.

  Deze kwalitatieve variabelen zijn discreet. Je kunt getallen koppelen aan de responsen van 'geheel mee eens' = 5 tot 'geheel niet mee eens' = 1, maar je mag er niet mee rekenen.

  Het gemiddelde van een op ordinaal niveau gemeten variabele heeft geen betekenis, evenmin zijn verschillen en verhoudingen betekenisvol.

- Op intervalniveau gemeten variabelen zijn kwantitatieve variabelen, waarvan de schaal van waarden geen natuurlijk nulpunt heeft. Intervalvariabelen kunnen zowel continu als discreet zijn.

  Voorbeeld

  De continue variabele 'temperatuur in graden Celsius', met als nulpunt de temperatuur van smeltend ijs en mogelijke waarden van -273.13° C tot zeer veel graden boven nul. De discrete intervalvariabele IQ, gemeten op een schaal met als centrale waarde 100 punten en mogelijke waarden van 0 (in theorie) tot 200 (of hoger).

- Op rationiveau gemeten variabelen zijn kwantitatieve variabelen, waarvan de schaal van waarden een natuurlijk nulpunt heeft. Ratiovariabelen kunnen zowel continu als discreet zijn.

  Voorbeeld

  De continue variabele 'temperatuur in graden Kelvin', met mogelijke waarden tussen het absolute nulpunt (-273.13° C = 0° K) tot zeer veel graden daarboven is een voorbeeld van een continue variabele die op rationiveau is gemeten. Ook 'lengte in meter' en 'gewicht in kilogram' zijn voorbeelden van continue variabelen die op rationiveau gemeten zijn.

Overzicht van de toegestane bewerkingen en de onderlinge relaties tussen waarden van variabelen op verschillend meetniveau.

Toegestane bewerking	Nominaal	Ordinaal	Interval	Ratio
Tellingen	+	+	+	+
Percentages	+	+	+	+
Rangorden	-	+	+	+
Verschillen	-	-	+	+
Gemiddelden	-	-	+	+
Verhoudingen	-	-	-	+
	+ Toegestaan		- Niet toegestaan	

Van de waarden van op intervalniveau en rationiveau gemeten variabelen kunnen de verschillen worden vergeleken, bij op *ordinaal niveau* gemeten variabelen kan dat niet. Het verschil tussen 10°C en 20°C is even groot als het verschil tussen 50°C en 60°C, maar het verschil tussen de score 'geheel mee eens' en 'mee eens' is niet persé even groot als het verschil tussen de scores 'niet mee eens' en 'geheel niet mee eens'.

Van de waarden van op *rationiveau* gemeten variabelen kunnen de verhoudingen (de ratio's) op een zinvolle manier met elkaar vergeleken worden. Bij *intervalvariabelen* kan dat niet. Honderd vlooien zijn er twee keer zoveel als vijftig. Maar iemand met een IQ van 120 is niet twintig procent intelligenter dan iemand met een IQ van 100.

1.3. Tabellen en grafieken

1.3.1. Frequentietabel voor discrete gegevens

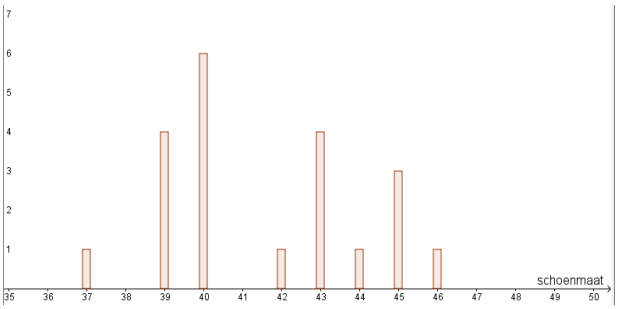
Een frequentietabel bevat een beknopte samenvatting van de uitkomsten van een onderzoek. In een frequentietabel wordt van iedere uitkomst de *frequentie*, de *proportie* of het percentage opgegeven. Afhankelijk van het doel worden soms ook *cumulatieve frequenties*, *proporties* of *percentages* in de tabel opgenomen.

Voorbeeld

Aan 21 leerlingen uit een klas van het IVde jaar werd hun schoenmaat gevraagd. De variabele schoenmaat is een kwantitatieve discrete variabele die op het intervalniveau gemeten is.

Variabele: mogelijke waarden		Absolute frequentie	Cumulatieve absolute frequentie	Relatieve frequentie (proportie)	Cumulatieve proportie
x_1	37	1	1	0,0476	0,0476
x_2	39	4	5	0,1905	0,2381
x_3	40	$6=f_3$	11	0,2857	0,5238
x_4	42	1	$12=caf_4$	0,0476	0,5714
x_5	43	4	16	0,1905	0,7619
x_6	44	1	17	0,0476	0,8095
x_7	45	3	20	0,1429	0,9524
x_8	46	1	21	0,0476	1
		n=21		1	

1.3.2 Staafdiagram voor discrete gegevens



1.3.3. Frequentietabel voor continue gegevens

Als er veel verschillende gegevens zijn, bijvoorbeeld als de gevens continu zijn, is het overzichtelijker de gegevens in klassen onder te brengen en de frequenties voor ieder van de klassen te bepalen.

Voorbeeld

In de tabel staan de rondetijden in seconden van een schaatser op de 10 km.

36.4	36.6	36.6	36.7	36.8
36.9	36.9	37.0	37.1	37.2
37.3	37.4	37.6	38.0	38.2
38.3	38.3	38.4	38.8	39.0
39.4	40.2	40.9	41.9	43.3

De rondetijden zijn in klassen ingedeeld en staan in de frequentietabel.

Klassen-grenzen	Frequentie	Klassen-midden	Klassen-breedte
36.0 - <38.0	13	37	2
38.0 - <40.0	8	39	2
40.0 - <42.0	3	41	2
42.0 - <44.0	1	43	2
Totaal	25		

Classificeren heeft tot doel de data overzichtelijker weer te geven. Het aantal klassen en de klassenbreedte bepalen de bruikbaarheid van de classificatie.

- Het aantal klassen is ongeveer $\sqrt{n}$ (n is het aantal waarnemingen). Minimaal moeten er 4 en maximaal mogen er 20 klassen zijn.
- Klassenbreedte = variatiebreedte / aantal klassen. De variatiebreedte is het verschil tussen de grootste en de kleinste gevonden uitkomst. De klassenbreedte mag naar eigen inzicht en behoefte aangepast worden.
- In de regel zijn klassen even breed. Als dit niet zo is, overweeg dan het gebruik van de frequentiedichtheid in plaats van de frequentie.
- Kies ronde getallen voor de klassengrenzen (zeker bij continue uitkomsten).
- Klassen moeten exclusief en uitputtend zijn. Exclusief wil zeggen, dat iedere uitkomst in slechts één klasse past. Uitputtend wil zeggen, dat iedere uitkomst in een klasse kan worden ingedeeld.

Wanneer uitkomsten in klassen van ongelijke breedte zijn ingedeeld kunnen de frequenties een vertekend beeld opleveren. Bij een variabele, die in ongelijke klassen is ingedeeld, is het daarom beter om met frequentiedichtheden te werken.

De frequentiedichtheid legt een relatie tussen de frequentie en de klassenbreedte.

$$\text{frequentie dichtheid} = \frac{\text{frequentie}}{\text{klassenbreedte}}$$

Voorbeeld

In een onderzoek naar jongerensterfte in Nederland is de leeftijd van overlijden in klassen ingedeeld, zie kolom 1 van de tabel. De indruk zou ten onrechte kunnen ontstaan, dat de twintigers het hoogste sterfterisico hebben. De leeftijdscategorieën hebben echter ongelijke *klassenbreedte*. Als we daarvoor corrigeren door de frequenties (kolom 2) te delen door de klassenbreedtes (kolom 3) dan krijgen we de frequentiedichtheden (kolom 4) in aantallen per jaar.

Leeftijd van overlijden	Frequentie	Klassenbreedte	Frequentie-dichtheid
0 - < 1	1220	1	1220
1 - < 5	610	4	152.5
5 - < 10	305	5	61
10 - < 20	915	10	91.5
20 - < 30	1525	10	152.5

Aan de dichtheden is te zien dat kinderen van 0 jaar oud het meeste risico lopen te overlijden.

1.3.4. Histogram voor continue gegevens

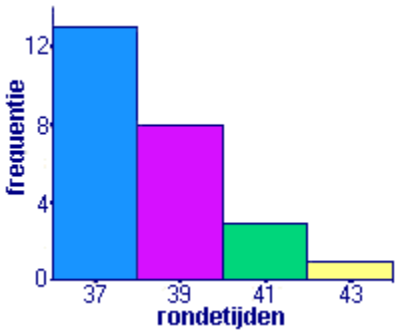
Een histogram wordt vaak gebruikt om kwantitatieve continue gegevens samen te vatten.

De gegevens worden ingedeeld in klassen (intervallen) die allemaal even breed gekozen worden (= intervallen van gelijke lengte). In het histogram staan op de horizontale as de klassen. Boven iedere klasse wordt een kolom getekend. De hoogte van de kolom komt overeen met de frequentie of de proportie (of het percentage) van de gegevens in die klasse. Om aan te geven dat de variabele continu is, grenzen de kolommen aan elkaar.

Een histogram is geschikt voor de weergave van de frequentieverdeling van variabelen die minimaal op intervalniveau gemeten zijn.

Voorbeeld

Het histogram van de frequentietabel van de rondetijden van een schaatser op de 10 km.



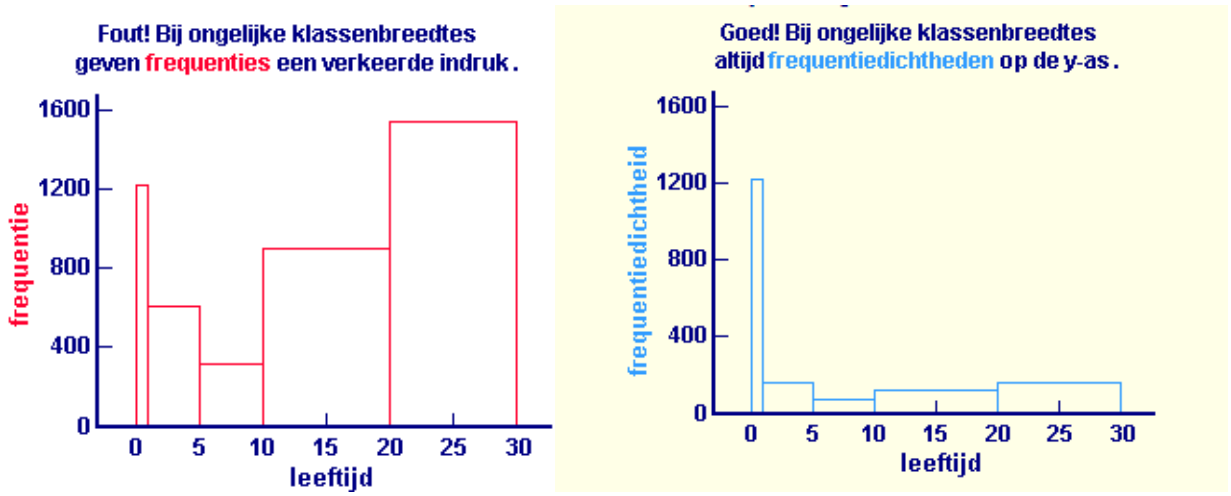
Als de waarden van een continue variabele zijn ingedeeld in klassen met ongelijke klassenbreedte, moet in een histogram de frequentiedichtheid op de verticale as worden uitgezet. De frequentie of de proportie zouden een vertekend beeld geven.

Voorbeeld

In een onderzoek naar jongerensterfte in Nederland is de leeftijd van overlijden in klassen ingedeeld, zie de kolommen 1 en 2 van de frequentietabel. De leeftijdscategorieën hebben ongelijke klassenbreedte. Corrigeren we daarvoor door de frequenties te delen door de klassenbreedtes (kolom 3) dan krijgen we de frequentiedichtheden (kolom 4) in aantallen per jaar.

Leeftijd van overlijden	Frequentie	Klassenbreedte	Frequentie-dichtheid
0 - < 1	1220	1	1220
1 - < 5	610	4	152.5
5 - < 10	305	5	61
10 - < 20	915	10	91.5
20 - < 30	1525	10	152.5

Met deze gegevens zijn de onderstaande histogrammen gemaakt. In het linkse histogram zijn op de y-as de frequenties uitgezet en in het rechtse histogram staan op de y-as de frequentiedichtheden. Het linkse histogram geeft een onjuist beeld van het sterfterisico op verschillende leeftijden. Het rechtse histogram laat terecht zien, dat het sterfterisico in het eerste levensjaar het grootst is.



Merk op, dat het *oppervlak* van een staaf recht evenredig is met het totaal aantal sterfgevallen in een bepaalde periode.

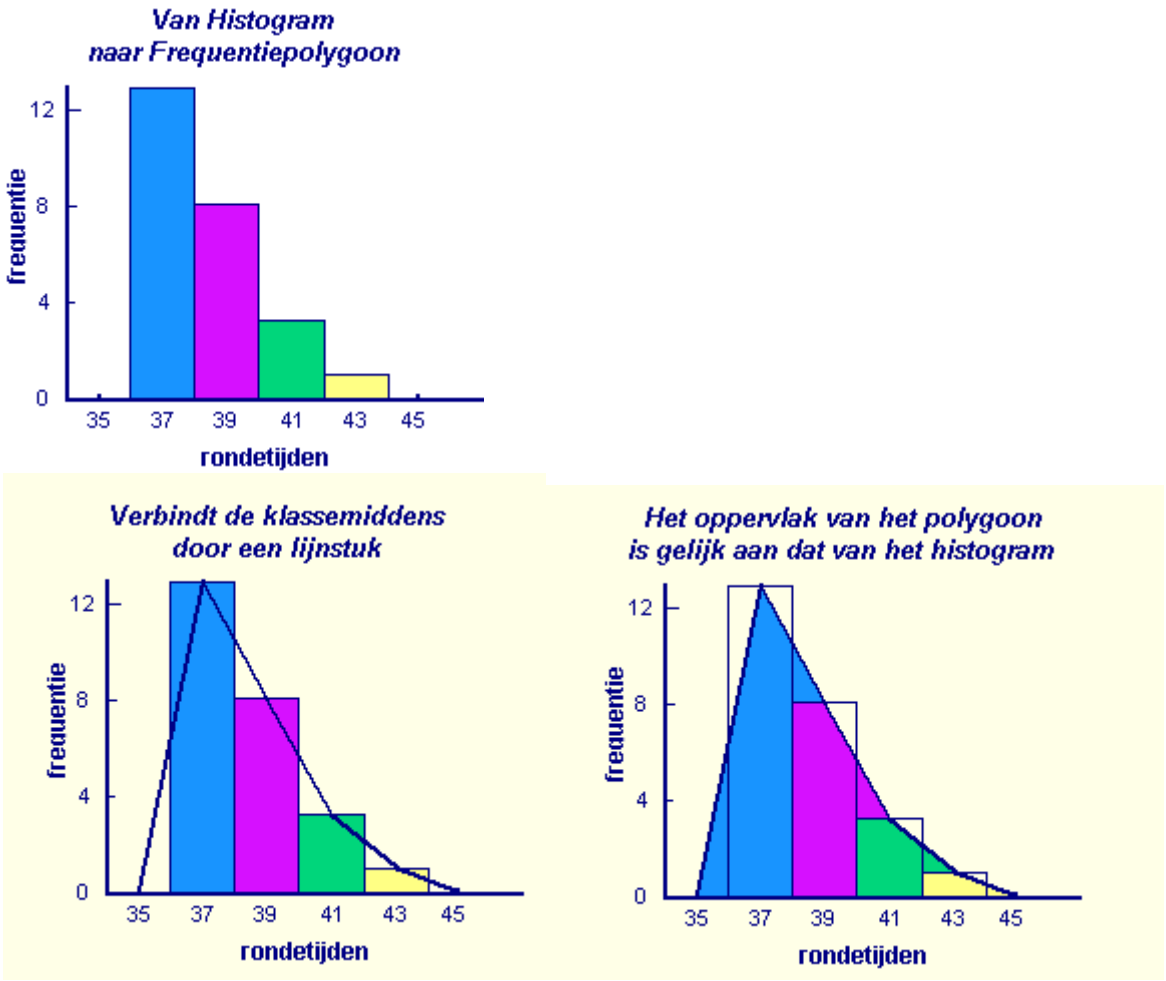
Ongelijke klassenbreedtes in een histogram of frequentietabel geven vaak aanleiding tot misverstanden. Deel de data daarom zo mogelijk in gelijke klassen in.

1.3.5 Frequentiepolygoon

Uitgaande van een histogram kan een frequentiepolygoon getekend worden. Op de verticale as staan de frequenties, op de horizontale as staan de klassenmiddens. De frequenties worden tegen de klassenmiddens uitgezet en de verkregen punten worden verbonden met een lijn.

Voorbeeld

De rondetijden van een schaatser op de 10 kilometer. Van ieder rondje is de tijd in seconden genoteerd. Hier vind je de frequentietabel en het histogram van de rondetijden.



De oppervlakte onder het polygoon moet gelijk zijn aan het totale oppervlak van de kolommen. Daarom wordt aan het begin en aan het einde een klasse met frequentie = 0 bijgetekend.

Bij ongelijke klassen is het gebruik van een polygoon niet correct. De oppervlakte onder de grafiek komt bij een polygoon met ongelijke klassen niet overeen met de oppervlakte onder het histogram.

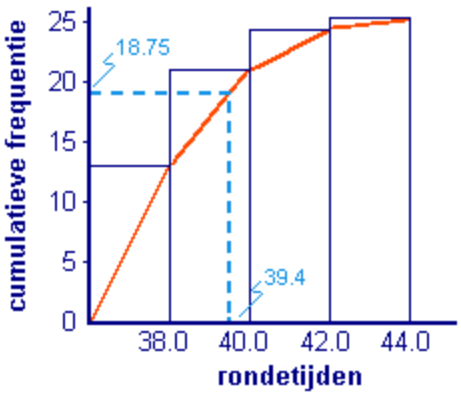
1.3.6 Cumulatief frequentiepolygoon

In een histogram worden de frequenties van de waarnemingen na classificeren in grafische vorm gepresenteerd. In een cumulatief frequentiediagram worden de cumulatieve frequenties in beeld gebracht, waarbij wordt uitgegaan van het overeenkomstige histogram.

Voorbeeld
De rondetijden van een schaatser op de 10 km zijn in klassen met een klassenbreedte van 2 seconden ingedeeld en samengevat in onderstaande frequentietabel. Behalve de frequenties zijn daarin ook de cumulatieve frequenties opgenomen.

Klassen-grenzen	Frequentie	Cumulatieve frequentie
36.0 - <38.0	13	13
38.0 - <40.0	8	21
40.0 - <42.0	3	24
42.0 - <44.0	1	25
Totaal	25	

Om het cumulatieve frequentiediagram te tekenen worden in het histogram niet de frequenties, maar de cumulatieve frequenties uitgezet. Vervolgens worden de hoekpunten van de blokken zoals in het diagram aangegeven door een lijnstuk verbonden. De frequentie van de waarnemingen in een blok wordt pas bereikt aan het eind van het interval van het blok. Bijvoorbeeld de frequentie 13 wordt bereikt bij de rondetijd 38.0 en de cumulatieve frequentie 13 + 8 = 21 bij 40.0. Uitgaande van de veronderstelling, dat de waarnemingen in iedere klasse regelmatig verdeeld zijn, kan voor iedere rondetijd op de x-as de bijbehorende cumulatieve frequentie bij goede benadering op de y-as worden afgelezen.



Omgekeerd kan voor iedere cumulatieve frequentie de bijbehorende waarde van de rondetijd worden berekend. In een goede benadering kunnen zo de kwantielen van een frequentie- of kansverdeling worden geschat. In het voorbeeld is het 3de kwartiel ingetekend. Op de y-as is dat 0.75 * 25 = 18.75. De corresponderende waarde op de x-as berekenen we door lineaire interpolatie: (18.75 - 13) / (21 - 13) = (x0.75 - 38) / (40 - 38) →x0.75 = 38 + 2 (18.75 - 13) / 8 = 39.4.

1.4. Kengetallen

1.4.1. Centrummaten

Een centrummaat is een getal dat aangeeft rond welke centrale waarde de waarnemingen liggen.

Veel gebruikte centrummaten zijn de *modus*, de *mediaan* en het *gemiddelde.* De keuze van een centrummaat wordt bepaald door:

- de gewenste informatie: wil je bijvoorbeeld de meest vóórkomende of liever de middelste uitkomst;
- het meetniveau van de variabele;
- de aanwezigheid van eventuele uitbijters (of uitschieters) in de data.

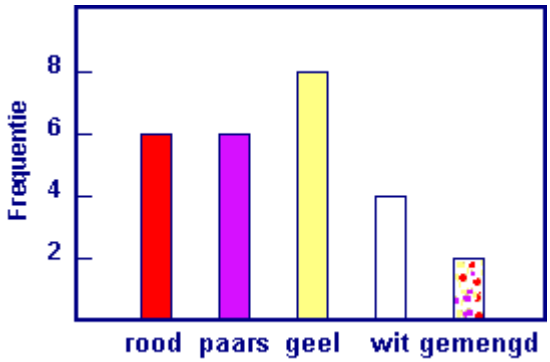
1.4.1.1. De modus

De modus is de waarde, die het meest vóórkomt, dus die met de hoogste frequentie. Bij een discrete variabele spreken we van de modus, bij een geclassificeerde continue variabele van de modale klasse. De modale klasse is de klasse met de hoogste frequentie of, bij ongelijke klassenbreedte, die met de hoogste frequentiedichtheid.

Voorbeeld

Een bloembollenverkoper heeft 26 soorten bloembollen in zijn assortiment. In de frequentietabel zijn de verschillende bollen gesorteerd op de kleur van de bloem. In deze tabel is geel de modus.

Kleur	Frequentie
rood	6
paars	6
geel	8
wit	4
gemengd	2
totaal	26



Een veel gemaakte fout is dat niet de waarneming met de hoogste frequentie als modus genoemd wordt (geel) maar de frequentie zelf (8).
De modus en de modale klasse zijn eenvoudig af te lezen uit de meeste grafieken: een staafdiagram voor discrete gegevens en een histogram voor continue, geclassificeerde gegevens.

1.4.1.2. De mediaan

Wanneer de gegevens in oplopende volgorde worden geplaatst, ontstaat een geordende getallenreeks. De mediaan is de middelste van deze naar grootte geordende getallen.

Voorbeeld

Negen personen hebben een test afgelegd. Het aantal fouten dat ieder van hen maakte is geteld.

0	0	1	1	1	2	2	3	4

De mediaan is het vijfde waarnemingsgetal, nl. 1.

Als het aantal gegevens even is, is de mediaan het gemiddelde van de middelste twee getallen.

Voorbeeld
Dit zijn de gegevens van 10 personen die de test deden.

0	0	1	1	1	2	2	3	4	6

De mediaan is het gemiddelde van het vijfde en zesde waarnemingsgetal: (1+2)/2=1,5.

De mediaan deelt de geordende gegevens in twee even grote groepen. Voor de mediaan geldt dat hoogstens 50% van de gegevens een waarde heeft die kleiner is dan die van de mediaan. Tegelijkertijd heeft hoogstens 50% van alle gegevens een waarde groter dan die van de mediaan. De mediaan is één van de meest gebruikte centrummaten.
Om de mediaan te kunnen bepalen moeten de gegevens gerangschikt kunnen worden. Op nominaal meetniveau is geen rangorde aan te brengen. Daarom kan de mediaan alleen berekend worden voor variabelen die minimaal op ordinaal meetniveau zijn gemeten.
Omdat de mediaan het middelste getal is, hebben extreem hoge of lage getallen weinig of geen invloed op de waarde van de mediaan. De mediaan wordt daarom een robuuste maat voor het centrum van een verdeling genoemd.

Berekening: Bereken M=(n+1)/2
Als M een geheel getal is dan is de mediaan het M-de getal in de rij.
Als M niet een geheel getal is dan neem je de twee dichtsbijzijde gehele getallen, deze twee getallen geven je dan de nummers van de getallen waar je het gemiddelde van uit moet rekenen.

1.4.1.3. Het gemiddelde

Bij het gemiddelde van n getallen $x_1, x_2, \ldots, x_n$ zijn drie grootheden met elkaar verbonden:

- het *aantal* getallen, wat je algemeen noteert door n.
- de *som* van die n getallen, namelijk $x_1 + x_2 + \cdots + x_n$, wat je op een korte manier opschrijft als $\sum_{i=1}^{n} x_i$.
- het *gemiddelde* van die n getallen, wat gelijk is aan $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.

Het gemiddelde van een reeks getallen is dus gelijk aan de som van die getallen gedeeld door het aantal getallen.

Voorbeeld. Het gemiddelde van de leeftijden 10, 10, 11, 14 en 15 jaar is precies 12 jaar, want 10+10+11+14+15=60 en 60/5=12.

Soms worden de data gegeven in een frequentietabel (n getallen; p verschillende).

Uitkomst	Frequentie
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_p	f_p
Totaal	n

Dan kan het gemiddelde $\overline{x}$ als volgt berekend worden:

$$\overline{x} = \frac{\sum_{i=1}^{p} x_i f_i}{n}.$$

De berekening van het gemiddelde $\overline{x}$ kan gemakkelijk gebeuren door aan de frequentietabel een kolom toe te voegen die de waarden $x_i f_i$ bevat.

Voorbeeld

Hieronder zijn de leeftijden van 50 jongeren gegeven in een frequentietabel.

x_i	f_i	$x_i f_i$
13	12	156
14	10	140
15	14	210
16	7	112
17	3	51
18	4	72
Totaal	50	741

De gemiddelde leeftijd van deze jongeren is 14,82 jaar, want 156+140+210+112+51+72=741 en 741/50=14,82.

Het gemiddelde is de belangrijkste en meest gebruikte centrummaat. Van alle centrummaten is het gemiddelde het gevoeligst voor uitbijters, omdat bij de berekening van het gemiddelde *alle* uitkomsten meetellen en niet alleen de *middelste waarde* (mediaan) of de *meest vóórkomende waarden* (modus).
Het gemiddelde is bedoeld voor variabelen die gemeten zijn op minimaal intervalniveau.

1.4.2. Spreidingsmaten

Een spreidingsmaat drukt de mate van de spreiding van gegevens uit in een getal.
Veel gebruikte spreidingsmaten zijn de *variatiebreedte* of *range*, de *interkwartielafstand* en de *standaardafwijking* of de *variantie*.

1.4.2.1. De standaardafwijking en de variantie

De standaardafwijking is een maat voor de spreiding van getallen rond hun gemiddelde. Als getallen ver uiteen liggen, dan is de standaardafwijking groot. Als zij dicht tegen elkaar liggen, dan is de standaardafwijking klein.

We kijken eerst hoeveel elk getal afwijkt van het gemiddelde $\bar{x}$. Voor een getal x_i is die afwijking gelijk aan $x_i - \bar{x}$. We zouden eraan kunnen denken om een 'gemiddelde afwijking' te gebruiken:
$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$.
Deze 'gemiddelde afwijking' is niet bruikbaar als een maat voor de spreiding want de som van de afwijkingen is altijd gelijk aan nul. We lossen dit probleem op door alle afwijkingen te kwadrateren. Al die kwadratische verschillen worden dan samengeteld. Zo krijgen we $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Die som delen we door $n-1$. Delen door $n-1$ heeft een goede reden in de statistiek maar daar gaan we nu nog niet op in. Tenslotte trekken we uit dat resultaat de positieve vierkantswortel. Eerst kwadrateren en later de wortel trekken, zorgt ervoor dat de uitkomst terug in dezelfde eenheid kan geschreven worden als de eenheid van de oorspronkelijke metingen.

De <u>standaardafwijking</u> *s* van de getallen $x_1, \ldots, x_n$ is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

De <u>variantie</u> is het kwadraat van de standaardafwijking. De notatie hiervoor is s^2, zoals verwacht. De formule is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

De variantie heeft als <u>eenheid</u> het *kwadraat* van de oorspronkelijke meeteenheden. De standaardafwijking wordt in de oorspronkelijke eenheden uitgedrukt. Wanneer het gewicht van mensen bijvoorbeeld in kilogrammen gemeten wordt, dan is de variantie in kg^2, maar de standaardafwijking is in kg.

De standaardafwijking en de variantie zijn bedoeld voor variabelen die gemeten zijn op minimaal intervalniveau. De standaardafwijking en de variantie zijn <u>gevoelig voor uitbijters</u>.

Voorbeeld

De standaardafwijking van de leeftijden 10, 10, 11, 14 en 15 jaar kan in stappen als volgt worden berekend.
Bepaal eerst het gemiddelde $\bar{x}$ = (10 + 10 + 11 + 14 + 15) / 5 = 12 jaar.
Bereken voor elk getal x_i de afwijking van het gemiddelde: $x_i - \bar{x}$ (zie tabel).

Kwadrateer deze afwijkingen: $(x_i - \bar{x})^2$ (zie tabel).

Tel de gekwadrateerde afwijkingen op: $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 22$ (zie tabel).

Deel de som door ééntje minder dan het aantal waarnemingen (*n-1*). De uitkomst hiervan is de variantie (s²):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{22}{4} = 5{,}5.$$

Trek de wortel uit de variantie. De uitkomst is de standaardafwijking (s):

$$s = \sqrt{5{,}5} = 2{,}3.$$

Leeftijd, x_i	Afwijking van het gemiddelde, $x_i - \bar{x}$	$(x_i - \bar{x})^2$
10	10-12=-2	(-2)²=4
10	-2	4
11	-1	1
14	+2	4
15	+3	9
		$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 22$

Als de <u>data gegeven</u> zijn <u>in een frequentietabel</u> (*n* getallen; *p* verschillende) werken we zoals bij het gemiddelde met de frequenties.

Uitkomst	Frequentie
x_1	f_1
x_2	f_2
⋮	⋮
x_p	f_p
Totaal	n

De <u>variantie</u> s^2 kan dan als volgt berekend worden:

$$s^2 = \frac{\sum_{i=1}^{p}(x_i - \bar{x})^2 f_i}{n-1}.$$

De formule voor de standaardafwijking s is dan

$$s = \sqrt{\frac{\sum_{i=1}^{p}(x_i - \bar{x})^2 f_i}{n-1}}.$$

Voorbeeld

De leeftijden van 50 jongeren worden gegeven in onderstaande frequentietabel. Om de standaardafwijking van de leeftijden te berekenen, voegen we drie kolommen toe aan de frequentietabel: de afwijking van het gemiddelde $(x_i - \bar{x})$, de gekwadrateerde afwijking van het gemiddelde $(x_i - \bar{x})^2$ en $(x_i - \bar{x})^2 f_i$.
Uit het voorgaande weten we dat $\bar{x} = 14{,}82$.

x_i	f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 f_i$
13	12	13-14,82=-1,82	(-1,82)²=3,3124	(3,3124).12=39,7488
14	10	-0,82	0,6724	6,724
15	14	0,18	0,0324	0,4536
16	7	1,18	1,3924	9,7468
17	3	2,18	4,7524	14,2572
18	4	3,18	10,1124	40,4496
Totaal	50			111,38

De variantie van de leeftijden is dan

$$s^2 = \frac{\sum_{i=1}^{p}(x_i - \bar{x})^2 f_i}{n-1} = \frac{111{,}38}{49} = 2{,}27.$$

De standaardafwijking is

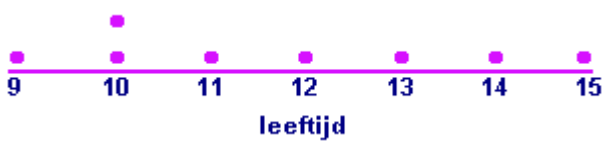
$$s = \sqrt{2{,}27} = 1{,}51.$$

1.4.2.2. Variatiebreedte of range

De variatiebreedte of range is het verschil tussen de hoogste en de laagste waarneming of tussen de hoogst en de laagst mogelijke waarden van een kwantitatieve variabele.

Voorbeeld

De variatiebreedte van de volgende leeftijden is 15 - 9 = 6 jaar.



De range geeft snel een eerste indruk van de spreiding in een verdeling. De standaardafwijking en de interkwartielafstand zijn betrouwbaarder spreidingsmaten, omdat de range erg gevoelig is voor toevallige uitbijters. Als het jongste kind in het voorbeeld 5 was geweest in plaats van 9 dan zou de range 15 - 5 = 10 jaar zijn geweest.
De range laat niet zien hoe de gegevens tussen de uiterste waarden verdeeld zijn. In het voorbeeld zijn de leeftijden gelijkmatig verdeeld. De verdeling had er echter ook zo uit kunnen zien.

1.4.2.3. Interkwartielafstand

De interkwartielafstand is het verschil tussen het eerste en het derde kwartiel.

Voorbeeld De rondetijden in seconden van een schaatser op de 10 km.

Het eerste kwartiel is 36.9 (25% van de uitkomsten is kleiner dan 36.9 en 75% van de uitkomsten is groter dan 36.9). Het derde kwartiel is 38.9. De interkwartielafstand is dus 38.9 - 36.9 = 2 seconden.

Om de kwartielen te bepalen moeten de gegevens eerst geordend worden. De interkwartielafstand is dus geschikt als spreidingsmaat voor variabelen, die op ordinaal of hoger niveau gemeten zijn.
De interkwartielafstand maakt alleen gebruik van het eerste en het derde kwartiel. Extreem hoge of lage getallen hebben geen invloed op de waarde van de kwartielen. De interkwartielafstand wordt dan ook een robuuste maat voor de spreiding van een verdeling genoemd. Bij scheve verdelingen verdient de interkwartielafstand daarom de voorkeur boven de standaardafwijking of de variatiebreedte.
Zie ook verder bij de boxplot.

1.4.2.4. Variatiecoëfficiënt

De standaardafwijking gedeeld door (de absolute waarde van) het gemiddelde, $s / |\bar{x}|$, wordt relatieve standaardafwijking of variatiecoëfficiënt genoemd. Omdat s en $\bar{x}$ in dezelfde eenheden worden gemeten, is de variatiecoëfficiënt dimensieloos. Soms wordt de variatiecoëfficiënt met 100 vermenigvuldigd en dan in procenten uitgedrukt.
De variatiecoëfficiënt kan gebruikt worden om de spreiding van variabelen te vergelijken, die op verschillende schalen zijn gemeten. De variatiecoëfficiënt is alleen zinvol te gebruiken, als de uitkomsten òf allemaal positief òf allemaal negatief zijn.

Voorbeeld

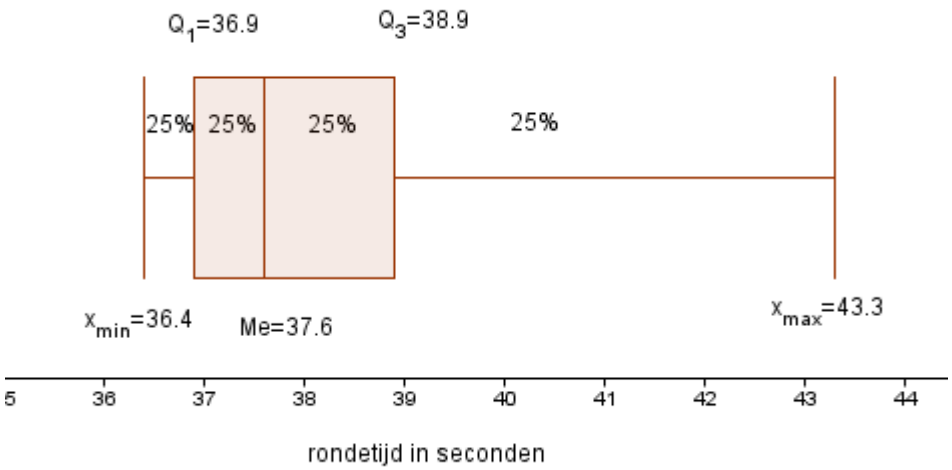
De standaardafwijking van het gehalte aan morfine in tabletten van nominaal 20 mg morfine bedraagt 0.8 mg. De standaardafwijking van het gehalte aan acetosal in tabletten van nominaal 500 mg is 20 mg. De variatiecoëfficiënt (in procenten) bedraagt voor de morfine 100 x 0.8 / 20 = 4% en voor de acetosal 100 x 20 / 500 = 4%. De (relatieve) spreiding tussen de doseervormen is dus even groot.

1.5. De boxplot

Een boxplot is een grafiek, waarin de positie van de kleinste en de grootste waarnemingsgetallen, de mediaan en de twee overige kwartielen zijn gevisualiseerd.
De boxplot deelt de verdeling in vier stukken die ieder 25% van de gegevens bevatten. Het centrum, de spreiding en de scheefheid van de verdeling van de gegevens zijn goed te zien. Ook de interkwartielafstand is in een boxplot direct af te lezen.

Voorbeeld

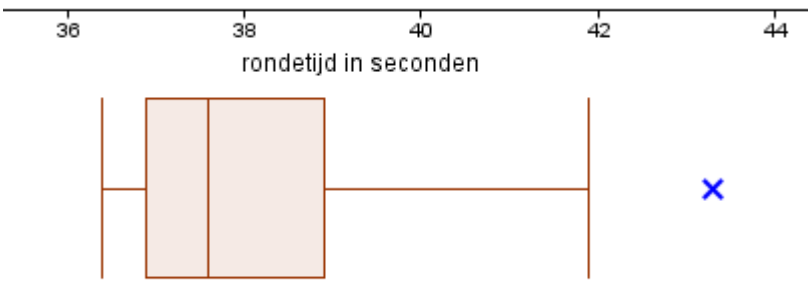
De rondetijden van een schaatser op de 10 km zijn in de boxplot samengevat.



Aangezien voor het maken van een boxplot de gegevens op volgorde gezet moeten worden, kan een boxplot alleen gemaakt worden voor variabelen op minimaal ordinaal meetniveau.
In een boxplot wordt de interkwartielafstand goed zichtbaar als het verschil tussen de waarden van de rechterzijde van de doos (het derde kwartiel) en die van de linkerzijde van de doos (het eerste kwartiel).

Uitbijters worden in een boxplot soms apart aangegeven. De stelen bevatten dan uiteraard minder dan 25% van alle gegevens. (Als vuistregel geldt dat de maximale steellengte 1,5 keer de interkwartielafstand is.)

De 'uitbijter' 43.3 seconden is in de boxplot apart aangegeven.


36
38
40
42
44
rondetijd in seconden

Checklist voor peilingen

1. Is duidelijk wie de **opdrachtgever** en/of **financier** van het onderzoek is? Dan kan worden vastgesteld of die belang heeft bij de uitkomsten. Dat is bijvoorbeeld het geval als het onderzoek wordt uitgevoerd in het kader van de promotie van een product, dienst of standpunt.
   - **Ja: Ga door naar 2.**
   - **Nee: Let op!** Er bestaat een risico dat de objectiviteit van het onderzoek niet is gegarandeerd.

2. Is er een **onderzoeksverantwoording** waarin precies staat aangegeven hoe het onderzoek is opgezet en uitgevoerd?
   - **Ja: Ga door naar 3.**
   - **Nee: Let op!** De betrouwbaarheid van het onderzoek kan niet worden vastgesteld.

3. Is duidelijk wat de **doelpopulatie** is? Dit is de groep die is onderzocht en waarop de conclusies van het onderzoek betrekking hebben.
   - **Ja: Ga door naar 4.**
   - **Nee: Let op!** De uitkomsten kunnen niet in de juiste context worden geïnterpreteerd.

4. Om de kwaliteit van de **vragenlijst** te kunnen beoordelen, moet in ieder geval zijn voldaan aan de volgende twee voorwaarden:
   - De volledige vragenlijst is opgenomen in de onderzoeksverantwoording;
   - De vragenlijst is voor de start van het onderzoek getest.

   Is aan deze voorwaarden voldaan?
   - **Ja: Ga door naar 5.**
   - **Nee: Let op!** De uitkomsten van het onderzoek kunnen onbetrouwbaar zijn.

5. Hoe is de **steekproef** getrokken? Is de steekproef geloot met een kanssteekproef waarin elke persoon in de doelgroep een positieve kans had om in de steekproef te komen? Die kansen moeten bij voorkeur gelijk zijn. In ieder geval moeten de kansen altijd kunnen worden berekend.
   - Geloot uit de hele groep. **Ga door naar 6.**
   - Geloot uit deel van de groep. Bijvoorbeeld alleen uit de Internetbezitters of alleen uit personen die in het telefoonboek staan. **Ga door naar 6**, maar besef dat de uitkomsten alleen betrekking hebben op dat deel van de groep.
   - Zelfselectie via Internet. **Let op!** Er zijn geen betrouwbare uitspraken mogelijk.
   - Een andere vorm van selectie met onbekende trekkingskansen, bijvoorbeeld een quota-steekproef. **Let op!** Er zijn geen betrouwbare uitspraken mogelijk.

6. Is de omvang van de gerealiseerde steekproef vermeld? Het gaat hier om het **aantal respondenten.**
   - **Ja: Ga door naar 7.**
   - **Nee: Let op!** De onzekerheidsmarges van de uitkomsten kunnen niet worden vastgesteld.

7. Is het **percentage respons** voldoende hoog, zeg hoger dan 50%?
   - **Ja: Ga door naar 8.**
   - **Nee: Let op!** Een lage respons kan leiden tot een grote mate van selectiviteit in het onderzoek en dus tot onjuiste uitkomsten.

8. Is een **correctie** (weging) uitgevoerd voor de opgetreden non-respons?
   - **Ja: Ga door naar 9.**
   - **Nee: Let op!** Non-respons leidt vaak tot een vertekening in de uitkomsten.

9. Worden **onzekerheidsmarges** in de uitkomsten vermeld? Dit is de afwijking die wordt veroorzaakt doordat niet iedereen in de steekproef zit, maar slechts een selectie.
   - **Ja.** Merk op dat in die marges niet de vertekening ten gevolg van non-respons en eventuele andere effecten (bijvoorbeeld geheugeneffecten) kunnen worden meegenomen. De onzekerheid kan dus nog groter zijn.
   - **Nee. Let op!** Het is dan lastig om de uitkomsten op hun juiste waarde te schatten. Echte effecten kunnen niet worden onderscheiden van de 'ruis' van de steekproef.