Samenvatting

Spreidingsmaten

Spreiding geeft aan hoe de data verdeeld zijn tussen de kleinste en de grootste waarde. Om deze spreiding aan te geven zijn er verschillende spreidingsmaten.

> De spreidingsbreedte, dit is het verschil tussen de hoogste en de laagste waarneming.

> De interkwartielafstand, dit is het verschil tussen het derde en het eerste kwartiel.

> De standaardafwijking of standaarddeviatie, deze spreidingsmaat is gebaseerd op de afwijkingen van de data ten opzichte van het gemiddelde en zegt dus iets over de spreiding rond het gemiddelde.

Normale verdeling en normaal verdeelde variabele

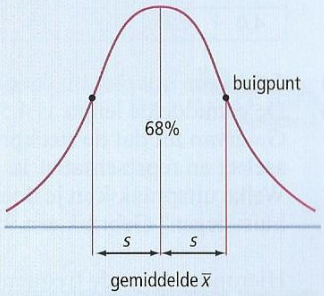
Een frequentieverdeling kun je vaak benaderen met een klokvormige grafiek.
De belangrijkste kenmerken van zo'n grafiek zijn:

> de grafiek is symmetrisch

> de symmetrie-as ligt precies bij de mediaan

> gemiddelde en mediaan vallen samen

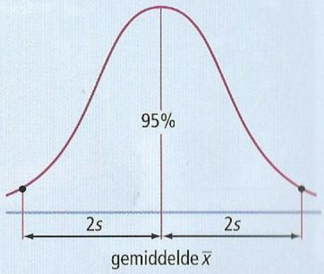
> De standaardafwijking s is de afstand van de symmetrie-as tot de buigpunten van de kromme.



Er gelden de volgende drie vuistregels:

1 van de data ligt ongeveer 68% tussen $\bar{x} - s$ en $\bar{x} + s$
2 van de data ligt ongeveer 95% tussen $\bar{x} - 2s$ en $\bar{x} + 2s$
3 van de data ligt vrijwel 100% tussen $\bar{x} - 3s$ en $\bar{x} + 3s$

Men spreekt in dit geval van een normale verdeling.
Je kunt ook zeggen dat de variabele normaal verdeeld is.



Populatiegemiddelde, steekproefgemiddelde, steekproevenverdeling

Bekijk de populatie: alle Nederlandse jongens van 21 jaar. Als je van elke jongen de lengte weet, dan kun je het gemiddelde berekenen. Dit gemiddelde heet het populatiegemiddelde. In de praktijk weet je niet van elke Nederlandse jongen van 21 de lengte en neem je een steekproef. De gemiddelde lengte in de steekproef heet het steekproefgemiddelde. Niet elke steekproef geeft hetzelfde steekproefgemiddelde. De variatie die daarbij optreedt kun je bestuderen door een heleboel steekproeven te simuleren, van elke gesimuleerde steekproef het gemiddelde te bepalen en te kijken naar de verdeling van deze steekproefgemiddelden. Je noemt dit de steekproevenverdeling van het gemiddelde. Een steekproevenverdeling kan benaderd worden met een normale verdeling.

Populatieproportie, steekproefproportie, steekproevenverdeling

In een populatie voldoet een bepaald deel van de elementen aan een zekere eigenschap. Dat deel heet de populatieproportie. In een steekproef heet het deel dat die eigenschap heeft de steekproefproportie. De populatieproportie en de steekproefproportie worden beide wordt geschreven als een getal tussen 0 en 1. De steekproefproportie is een schatter van de populatieproportie. Deze schatting is betrouwbaarder naarmate de steekproefomvang groter is.
De variatie in steekproefproporties wordt alleen zichtbaar wanneer je in theorie meerdere steekproeven simuleert. Je krijgt dan een steekproevenverdeling van een proportie.
Een steekproevenverdeling mag je, bij voldoende steekproefgrootte, benaderen met de normale verdeling.

Voorbeeld

In een steekproef onder 200 huishoudens blijken 34 huishoudens geen toegang tot internet te hebben.
De steekproefproportie bedraagt $\frac{34}{200} = 0{,}17$. Uitgaande van deze steekproefproportie krijg je een steekproevenverdeling met gemiddelde van 0,17 en standaarddeviatie van $\sqrt{\frac{0{,}17 \cdot 0{,}83}{200}} \approx 0{,}03$.

Betrouwbaarheidsinterval, foutenmarge, 95%-betrouwbaarheidsniveau

Het interval, waarbinnen de middelste 95% van alle steekproefproporties of steekproefgemiddelden ligt, heet het 95%-betrouwbaarheidsinterval. De halve lengte van dat interval is de foutenmarge.
Het 95%-betrouwbaarheidsinterval voor een populatieproportie is gelijk aan $\left[p - 2\sqrt{\frac{p(1-p)}{n}},\ p + 2\sqrt{\frac{p(1-p)}{n}}\right]$ met p de steekproefproportie en n de steekproefomvang.
Het 95%-betrouwbaarheidsinterval voor het populatiegemiddelde bereken je met de formule $\left[\bar{x} - 2 \cdot \frac{S}{\sqrt{n}},\ \bar{x} + 2 \cdot \frac{S}{\sqrt{n}}\right]$. Hierin is $\bar{x}$ het steekproefgemiddelde, S is de standaarddeviatie van de steekproef en n is de steekproefomvang.

Voorbeeld

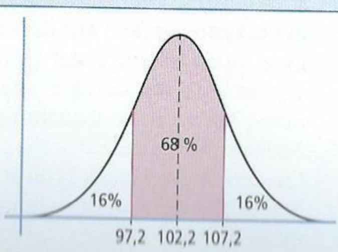
Het 95%-betrouwbaarheidsinterval van de steekproevenverdeling van het voorbeeld hierboven is $[0{,}17 - 2 \cdot 0{,}03; 0{,}17 + 2 \cdot 0{,}03]$ en dit is $[0{,}11; 0{,}23]$.
Je zegt: met 95% betrouwbaarheid heeft 17% van alle huishoudens geen toegang tot internet met een foutenmarge van 6%.

Je kunt de symmetrie en de vuistregels van de normale verdeling gebruiken

Als een variabele normaal verdeeld is, dan gebruik je de symmetrie en de vuistregels om uitspraken te doen over de variabele.

Voorbeeld

Van 2500 Elstar appels is het gewicht in gram genoteerd.
Het gemiddelde gewicht is 102,2 gram, de standaardafwijking bedraagt 5,0 gram. De verdeling van de variabele gewicht mag je benaderen met een normale verdeling.
Hoeveel appels wegen meer dan 97,2 gram?



Oplossing

Schets bij het gegeven gemiddelde en standaardafwijking een grafiek.
Omdat de variabele gewicht normaal verdeeld is, ligt 97,2 gram precies bij $\bar{x} - s$.
16% + 68% = 84% van de appels weegt meer dan 97,2 gram. Dat zijn dus $0{,}84 \cdot 2500 = 2100$ appels.