Statistiek
Hoorcollege 1, 09-11-2015

Kwantitatieve data-analyse

Data revolutie
Vandaag de dag speelt data analyse een rol in praktisch elke beslissing die gemaakt wordt door corporaties. Het is een significant corporate asset geworden.

Statistiek is het halen van informatie uit een gegevensset. Dit is belangrijk om betere beslissingen te kunnen maken en om interessante vragen te stellen.

Onderscheid maken tussen:
- **Beschrijvende statistiek** (descriptive statistics): data samenvatten en op een informatieve manier presenteren
- **Verklarende statistiek** (inferential statistics): gebruik maken van een steekproef uit populatie en hieruit conclusies trekken.

Statistische hoofdconcepten
Populatie: Een groep waarin je geïnteresseerd bent
Voorbeeld: alle eerstejaars BDK studenten
Sample: Een steekproef uit je populatie
Voorbeeld: Alle eerstejaars BDK studenten op de eerste rij.
Variabele: Het karakteristiek waarin je geïnteresseerd bent
Voorbeeld: Het inkomen van een Formule-1 rijder
Waarde: Alle mogelijke observaties van een variabele
Voorbeeld: Van €150.000 tot €30.000.000
Data: De daadwerkelijke geobserveerde waarde
Voorbeeld: Het geobserveerde jaarlijkse inkomen van 12 verschillende Formule-1 rijders in miljoenen euro's.

Typen data:
Kwantitatieve data: Je kunt er mee rekenen en het zijn reële nummers op de getallenlijn.
Voorbeeld:jaarlijkse inkomen van Formule-1 rijders
Ordinale data: Er bestaat een rangorde
Voorbeeld: gezondheid status (1= very good, 2 = good etc.).
Nominale waarde: Er bestaat geen rangorde
Voorbeeld: Burgerlijke staat (1=single, 2=getrouwd, 3=gescheiden, 4=weduwe).

Een **bar chart** wordt gebruikt om frequenties te beschrijven (histogram). Een **pie chart** wordt gebruikt om relatieve frequenties te beschrijven (cirkeldiagram).

Relatie tussen 2 variabelen:
- **Scatter diagram**: hoe dichter de punten op de lijn liggen, hoe sterker de relatie. Als alle punten op de lijn vallen noemen we dit deterministisch.
- **Lineaire relatie:** Als de meeste punten dicht bij een rechte lijn liggen

Basisprincipes voor grafieken en tabellen
Visualisatie van data is heel belangrijk
Het boek "The Visual Display of Quantitatieve Informatie" laat 7 basisprincipes van grafische uitmuntendheid zien:

- Geef de data weer
- Vermijd onduidelijkheid over wat er in moet staan
- Grafieken moeten goed leesbaar zijn, er moet in 1 oogopslag te zien zijn wat bedoeld wordt
- Integreer de tekst en de grafiek

Er kan ook gelogen worden met statistiek. Dit gebeurt door bijvoorbeeld een grafiek heel erg uit te zoemen waardoor het verschil heel groot lijkt. In werkelijkheid kan het verschil dan veel kleiner zijn.

Arithmetic mean (average): de som van de observaties gedeeld door het nummer van de observasties

- Population mean (gemiddelde): $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$, met N=populatiegrootte
- Sample mean (steekproefgemiddelde): $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{n}$, met n=steekproefgrootte

Let op de notatie! μ is een onbekende parameter en $\bar{x}$ een statistiek

Maten van centrale locatie

Mediaan: De middelste observatie
Modus: De observatie dat de hoogste frequentie heeft. De modus van een steekproef of populatie hoeft niet uniek te zijn.
Percentiel: De waarde voor welke P % minder dan de waarde is en (100-P)% groter is dan het waarde
Voorbeeld: als jou cijfer het 80th percentiel is, zit 80% er onder en 20% er boven.
Kwartiel: Het 25^e, 50^e en 75^e percentiel
Mediaan: Het 50^e percentiel/ het tweede kwartiel
Spreiding (range): De grootste observatie – de kleinste observatie
Interquartile range: 3^e kwartiel – 1^e kwartiel

Een **boxplot** bestaat uit eerste, tweede en derde kwartiel. Ook heb je '**snorharen'**, het maximum en het minimum. Punten dat buiten de snorharen liggen worden outliers genoemd. De maximale lengte van een snorhaar is 1,5 keer de interquartile range. Een voorbeeld van een boxplot is te vinden in de slides van hoorcollege 1 slide 38 en 39.

De range en interkwartiel range maken maar gebruik van 2 datapunten.
Variantie: De gemiddelde afwijking van het gemiddelde.

- **Populatie variantie**: $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$
- **Steekproefvariantie**: $s^2 = \frac{\sum_{i=1}^{n}(x_{i-\bar{x}})^2}{n-1}$

De afwijking kan positief en negatief zijn, daarom wordt er een kwadraat genomen.

Voor een voorbeeld van steekproefvariantie, zie de slide.

Chebyshev's ongelijkheid en empirical regel

In elke steekproef of populatie zijn bijna alle waarden dichtbij het centrum. Ten minste $1-(1/k^2)$ van de waarden zijn in de k standaard deviaties van het centrum, voor k>1.
Zie slide 44 voor de percentagse of the data in het interval.

Correlatie: het weergeven van de relatie tussen 2 waarden. Dit kan goed weergegeven worden bij gewicht en lengte. Het kan weergegeven worden in een scatterplot. Bij een rechte lijn is er geen relatie, bij een stijgende lijn een positieve relatie, en bij een dalende lijn een negatieve relatie.

- **Population covariance**: slide 49
- **Sample covariance**: slide 49