

Samenvatting Onderzoekstechnieken

TIN 2 - HoGent

Lorenz Verschingel

13 april 2015

1 Het onderzoeksproces

1.1 De wetenschappelijke methode

Aan de hand van **empirisch onderzoek** zijn we geïnteresseerd in volgende zaken:

1. Exploratie
2. Beschrijving
3. Voorspelling
4. Controle

Het onderzoeksproces verloop normaal gezien als volgt:

1. Formuleren
Wat is de onderzoeksvraag
2. Exacte informatie behoefte definiëren
Welke specifieke vragen moeten we stellen
3. Uitvoeren onderzoek
Enquêtes, simulaties. . .
4. Verwerken gegevens
Statistische software
5. Analyseren gegevens
Uitvoeren statistische methodes
6. Conclusie schrijven
Schrijven onderzoeksverslag

1.2 Basisconcepten in onderzoek

1.2.1 Variabelen en waarden

Een variabele is een eigenschap van een object waardoor we objecten van elkaar kunnen onderscheiden.

Een waarde is een specifieke eigenschap, een invullen voor een variabele.

1.2.2 Meetniveaus

De kwalitatieve schalen zijn:

1. Nominaal: Categorieën geslacht, ras, land. . .
2. Ordinaal: Volgorde militaire rang, opleidingsniveau. . .

De kwantitatieve schalen zijn:

1. Interval: Meting: nulpunt is onbelangrijk graden Celsius
2. Ratio: Meting: t.o.v. absoluut nulpunt meter, Joule, kilogram

1.2.3 Verbanden tussen variabelen

Er is een verband tussen variabelen als hun waarde systematisch veranderen.

Men is vooral op zoek naar oorzakelijke verbanden:

- Frustratie leidt tot agressie
- Alcohol leidt tot minder oplettendheid

De oorzaak is hierbij de onafhankelijke variabele.

Het gevolg is de afhankelijke variabele.

Hierbij moet men wel opletten. Een verband tussen variabelen duidt niet noodzakelijk op een oorzakelijk verband.

2 Analyse van 1 variabele

2.1 Beschrijvende statistiek

2.1.1 Centrummaten

Het **gemiddelde** is de som van alle waarden gedeeld door het aantal waarden.

Om de **mediaan** te vinden, sorteert men de waarden en kiest men dan het middelste nummer. Bij een even aantal getallen neemt men het gemiddelde van de twee middelste.

De **modus** is het vaakst voorkomende getal in een reeks getallen. Als men niet onmiddellijk de modus kan aflezen kan men gebruik maken van ranges. Deze ranges zijn dan modale klassen.

2.1.2 Spreidingsmaten

Het **bereik** van een reeks getallen is de absolute waarde van het verschil tussen het grootste en het kleinste getal in de reeks: $|x_{min} - x_{max}|$

De **kwartielen** van een gesorteerde reeks getallen zijn de waarden die de lijst in vier gelijke delen verdeelt. Elk deel vormt dus een kwart van de dataset. Men spreekt van een eerste, tweede en derde kwartiel genoteerd als respectievelijk Q_1 , Q_2 , Q_3 . Hierbij is Q_2 de mediaan.

De **variantie** is het gemiddelde gekwadrateerde verschil tussen de elementen van de dataset en zijn gemiddelde: $\sigma^2 = \frac{1}{n} \sum_i^n (\mu - x_i)^2$

De **standaardafwijking** is hde vierkantswortel van de variantie.

2.2 Eenvoudige grafieken

2.2.1 Cirkeldiagram

Voordelen:

- Met percentages rond 20% kan men makkelijk verduidelijken t.o.v. de volledige dataset.

Nadelen:

- Vergelijking op basis van de hoek.
- De figuur wordt onduidelijk als er veel categorieën zijn.

Men gebruik best zo weinig mogelijk een cirkeldiagram.

2.2.2 Staafdiagram

Voordelen:

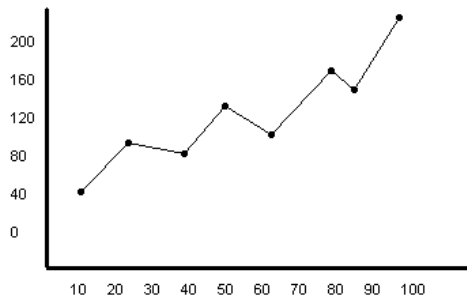
- Categorieën zijn makkelijk te vergelijken.
- Per categorie zijn meerdere staven mogelijk.

2.2.3 Boxplot

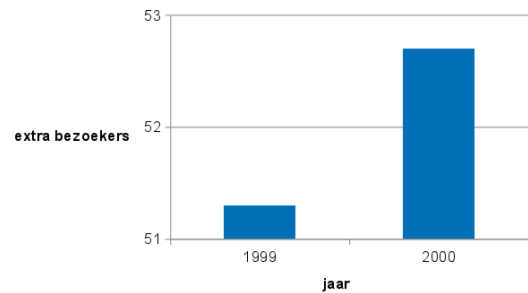
Voordelen:

- Snelle manier om data te inspecteren en verschillende datasets te vergelijken.

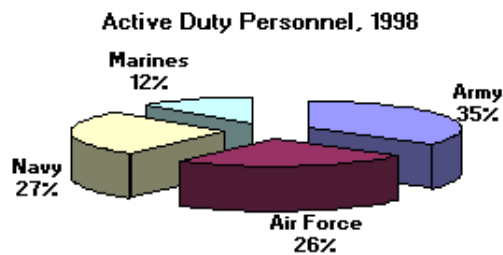
2.3 Interpretatie van grafieken



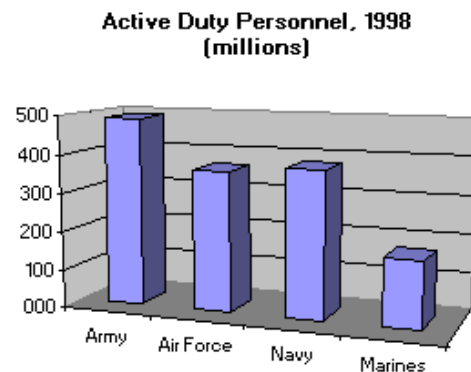
(a) Data-ambigüiteit



(b) Data distortion



(c) Data distraction 1



(d) Data distraction 2

Figuur 1: Valkuilen bij het interpreteren van grafieken

2.3.1 Data-ambigüiteit

Data-ambigüiteit betekent vergeten aan te duiden wat de data betekent. Zie figuur 1a.

Enkele tips om dit te voorkomen:

- Benoem de assen
- Geef een duidelijk titel
- Benoem de meeteenheid (en evt. de grootorde)
- Voeg een bijschrift toe met uitleg over de grafiek

2.3.2 Data distortion

Data distortion betekend dat men verkeerde conclusies kan trekken uit een grafische voorstelling. Zie figuur 1b: merk hierbij op dat de as niet op nul begint en er maar 3 waarden worden weergegeven.

2.3.3 Data distraction

Dit betekent dat de grafiek te veel toeters en bellen bevat. Men moet de *inkt to data ratio* beperken. De figuren 1c en 1d zijn hier voorbeelden van.

3 Analyse van 2 variabelen

3.1 Bivariatie analyse

3.2 Kruistabellen en Cramér's V

Een kruistabel wordt als volgt opgesteld:

1. Percenteer: Deel de cel door het kolomtotaal
2. Bepaal de schatter $e = \frac{\text{kolomtotaal} \times \text{rijtotaal}}{n}$ n: totaal aantal
3. Bepaal het verschil $cel - e$
4. Kwadrateren en normeren $(cel = \frac{\text{verschil}^2}{e})$

$$\chi^2 = \sum \frac{(a-e)^2}{e}$$

$$V = \sqrt{\frac{\chi^2}{n \times (k-1)}} \text{ met: } n = \text{totaal en } k = \min(\#rijen, \#kolommen)$$

Cramér's V is een maat die aanduidt hoe sterk de samenhang is tussen twee nominale variabelen. Dit getal ligt altijd tussen 0 en 1.

Waarde	Interpretatie
0	Geen samenhang
0.1	zwakke samenhang
0.25	redelijk sterke samenhang
0.5	sterke samenhang
0.75	zeer sterke samenhang
1	volledige samenhang

Tabel 1: Interpretatie van Cramér's V

3.3 Regressie

Bij regressie gaan we proberen een consistente en systematische koppeling tussen variabelen te vinden.

- **Niet-monotoon:** aanwezigheid (of afwezigheid) van de ene variabele gerelateerd aan de aanwezigheid (of afwezigheid) van een andere variabele.
- **Monotoon:** algemene richting van de samenhang tussen de twee variabelen kan aangeduid worden.

3.3.1 Lineaire regressie

Een lineair verband is een rechte lijnige samenhang tussen een onafhankelijke en afhankelijke variabele, waarbij kennis van de onafhankelijke variabele kennis over de afhankelijke variabele geeft.

KLEINSTE KWADRANTEN METHODE

We proberen een rechte te vinden van de vorm $y = \beta_0 + \beta_1 x$

Hierbij geldt:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Kwadraten omdat kleine verschillen minder in rekening moeten gebracht worden dan grote verschillen.

Je kan altijd een kleinste kwadratenmethode uitvoeren ==; Dit is daarom wel niet altijd goed.

3.4 Correlatiecoëfficiënt en determinatiecoëfficiënt

De Pearson correlatiecoëfficiënt is een maat voor de sterkte van de lineaire samenhang tussen x en y.

De determinatiecoëfficiënt verklaart het percentage van de variantie van de waargenomen waarden t.o.v. de regressierechte.

3.4.1 Covariantie

$$\text{cov}(x, y) = \sum_i \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$R = \frac{\text{cov}(x, y)}{\sqrt{\frac{(x - \bar{x})^2}{n} \times \frac{(y - \bar{y})^2}{n}}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \times \sqrt{\sum (y - \bar{y})^2}}$$