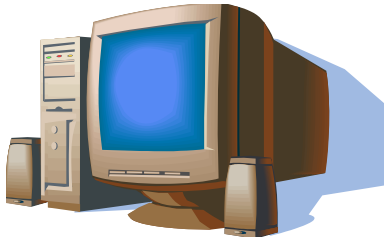
Onderzoeksmethoden: Statistiek 1: Beschrijvende statistiek

Peter de Waal
(gebaseerd op slides Marjan van den Akker, Peter de Waal)

Departement Informatica
Beta-faculteit, Universiteit Utrecht



0039475652058465426184950502876164759503 0...

- Joepie, ons computerprogramma levert output...
- Joepie, we hebben gegevens uit onze enquete...

Q: Wat doen we hiermee?

Output gegevens

1. Valideren
2. Ordenen:
   1. Tabellen
   2. Grafieken
   3. 'Statistieken'
3. Mogelijke conclusie definieren:
   1. Relaties en verschillen
   2. Gebaseerd op je onderzoeksvraag, maar eventueel andere interessante fenomenen.
4. Hypotheses toetsen en analyseren mbv. Statistiek.

Kansrekening en statistiek in de informatica

- Randomized algorithms
- Data-mining
- Bayesiaanse netwerken voor medische diagnose
- Planning met verstoringen
- Modellen voor bewegende karakters in spellen
- Testen computer-games

Materiaal

- Nel Verhoeven. Statistiek in stappen. Boom Lemma Uitgevers, 2013. ISBN 978 90 5931 9639.
- Gedeeltelijk gebaseerd op slides van Wetenschappelijke Onderzoeksmethoden (INKU Bachelor)

Wat is statistiek?

- '*Leer* en *methode* om door middel van cijfers inzicht te krijgen in massale verschijnselen, .. (van Dale)
- 'De wetenschap, de methodiek en de techniek van het verzamelen, bewerken, interpreteren en presenteren van gegevens. (Wikipedia)

Kansrekening en Statistiek

- Kansrekening:
  - Theoretische basis: Hoofdstuk 4
- Statistiek:
  - Theoretische basis: Hoofdstuk 4
  - Beschrijvende statistiek: Hoofdstuk 2 + 3
  - Toetsende statistiek: Hoofdstuk 5, 6, + 9

Vandaag: Beschrijvende statistiek

There are three kinds of lies: lies, damn lies, and statistics
(Mark Twain)

Enquetes

MAURICE DE HOND (11.09)	
VVD	36
PvdA	36
PVV	18
CDA	12
SP	20
D66	11
GL	4
CU	5
SGP	3
PvdD	3
50Plus	2

Onderzoek

NOS

Onderzoek 'vleeshufters' frauduleus

Vlees op de grillplaat van een gourmetstel
Foto: Flickr/woordenaar - CC 2.0 by-nc-sa

Toegevoegd: donderdag 8 sep 2011, 09:19
Update: donderdag 8 sep 2011, 10:25

Het onderzoek van een aantal hoogleraren waaruit zou blijken dat mensen "hufterig worden als ze aan vlees denken", berustte op fraude. Eén van die hoogleraren is professor Diederik Stapel van de Tilburg University, die gisteren op non-actief is gezet omdat hij met gegevens had gesjoemeld.

De twee andere hoogleraren die aan het vleesonderzoek hebben meegewerkt spreken van "een afgang van mega-formaat". Volgens hoogleraar Roos Vonk van de Radboud Universiteit in Nijmegen gebruikte Stapel verzonnen data. Op haar website schrijft Vonk: "Niet alleen was het onderzoek niet waardevrij, de resultaten waren compleet *fake*!".

"Ik wil dan ook iedereen tegen wie ik soortgelijke dingen heb gezegd mijn welgemeende excuses aanbieden. Ik wil ook duidelijk maken dat er vermoedelijk helemaal geen onderzoeksresultaten bestaan over de effecten van denken aan vlees en dat we hierover dus geen enkele uitspraak kunnen doen."

Bron: nos.nl

Onderzoek

Onderzoek naar gamen deugt standaard niet

#Games
Shooters zijn helemaal niet goed voor je ontwikkeling, bewijst onderzoek over gameonderzoek.

Redactie games
Amsterdam

Het beste onderzoek doet pijn, dat bewijst een studie onder leiding van de Amerikaanse onderzoeker Walter Boot maar weer eens. Is het al langer in de mode om game-onderzoeken af te kraken die zich richten op de negatieve kanten van gaming (verslaving, gewelddadigheid bij spelers, dat werk), Boot en collega's maken duidelijk dat van onderzoek naar positieve kanten van gaming ook weinig deugt.

Boot, assistent-professor aan de psychologiefaculteit van de Florida State University, schrijft in wetenschapsblad *Frontiers in Psychology* dat onderzoek naar positieve effecten van gamen stelselmatig onbetrouwbaar is. 'Het is een hype', zegt Boot. 'In werkelijkheid is er maar weinig onweerlegbaar bewijs dat games bijdragen aan het ontwikkelen van cognitieve vaardigheden.'

Boot en zijn team wijzen op onderzoek dat de afgelopen tien jaar is gedaan, en waarvan de uitkomsten door (game)media gretig zijn opgepikt. Schietspellen zouden je reactiesnelheid vergroten, door samenspelen in grote online games zouden mensen ook buiten de games socialer zijn – dat werk.

Maar, helaas pindakaas. De studies hebben allemaal een wankele methodologische basis, en zijn daarmee wetenschappelijk waardeloos, schrijft Boot. Hij wijst erop dat in game-onderzoeken weliswaar wordt gemeten hoe een groep hard-core gamers presteert in vergelijking met niet-gamers, maar dat dat weinig zegt als er geen random groep gamers wordt gevolgd. Onderzoekers werven hun kandidaten op campussen en vragen 'expert gamers' deel te nemen aan het onderzoek. Nadeel daarvan is dat een groep mensen wordt aangesproken met uitzonderlijke gamevaardigheden. Daardoor kan een grotere reactiesnelheid in andere situaties dan niet aan het gamen kan worden opgehangen: het is net zo goed mogelijk dat snelle games van nature snelle jongens aanspreken. Exit onderzoek.

Bron: De Pers, 19-09-2011

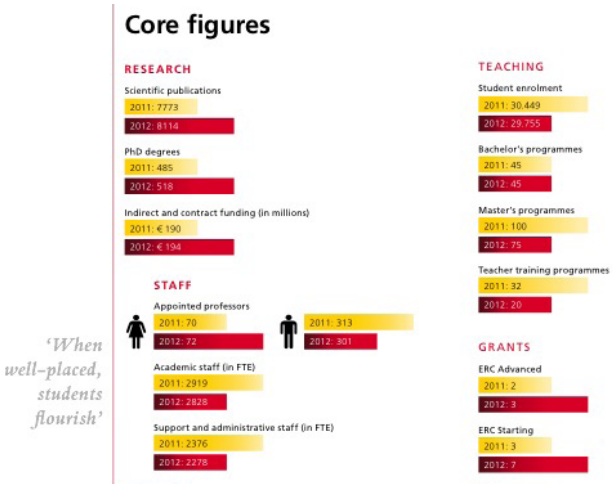
Definities

Ingeschreven studenten: **176**
Meegedaan aan tentamen: **123** (70%)
Geslaagd: **94** (76%, 53% van het totaal)
Number of replies: **35** (20%)

0: 0 x
1: 0 x
2: 0 x
3: 0 x
4: 14 x
5: 15 x
6: 19 x
7: 22 x
8: 24 x
9: 19 x
10: 10 x

n=123, mean=7.01, stddev=1.79

Vertekenende plaatsje



Bron: UU Jaarbeeld 2012

Statistiek kan veel. . .

Maar wees nauwkeurig!

Manieren om fouten te maken:

- Garbage data
- Slechte steekproef
- Wisselende of onduidelijke definities
- Vertekenende plaatjes
- Verkeerde gevolgtrekkingen
- . . .

DATA VERVALSEN IS FRAUDE

Definities: Populatie en steekproef

- **Populatie**: verzameling van alle personen, objecten of gebeurtenissen waar een vraagstelling of onderzoek betrekking op heeft
- **Steekproef**: selectie van elementen uit de populatie
- **Variabele**: te meten/bepalen karakteristiek van persoon, object. . .

Vraag: Hoe vaak gaan Utrechtse informatica-studenten uit?

- Variabele: aantal uitgaansavonden per maand
- Populatie: alle studenten ingeschreven voor de opleiding Informatica Utrecht.
- Steekproef: Remco, Maxime, Jelle, Jeanine, Timo, Falco, Bram, Jona, Rutger

Definities: meetniveaus van variabelen

- Nominaal
- Ordinaal
- Interval
- Ratio

Meetniveau: nominaal

Nominaal meetniveau:

- Indeling in categorieen
- Indeling:
  - Uitsluitend (mutually exclusive)
  - Uitputtend (exhaustive)

Voorbeeld:

-  of 
- Informatica of Gametechnologie

Meetniveau: ordinaal

Ordinaal meetniveau:

- Indeling in rangorde

Voorbeeld: hoogste niveau van genoten vervolgopleiding:

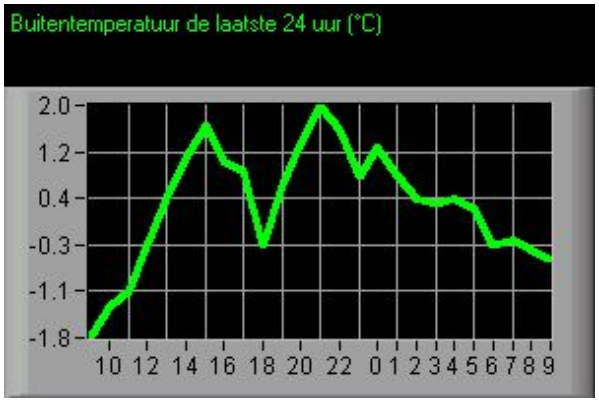
1. Middelbare school
2. HBO
3. Universiteit

Meetniveau: interval

Interval meetniveau:

- Rangorde, maar zegt ook iets over grootte van het verschil, afstand (geen natuurlijk nulpunt)

Voorbeeld: Temperatuur in Celsius



Meetniveau: ratio

Ratio meetniveau

- rangorde, zegt iets over afstand *en* over verhouding,
- Gevolg: Natuurlijk nulpunt, geen negatieve waarden.

Voorbeeld

- Lichaamslengte
- Gewicht
- Looptijd algoritme

Welke schaal?

Likert schaal:

- Op een bewering wordt gereageerd in termen van eens / oneens

Vond je het vak leuk? *Did you enjoy the course?*	**1**: niet / *no* 1 x **2**: 5 x **3**: 5 x **4**: 14 x **5**: wel / *yes* 10 x *n*=35, mean=3.77, stddev=1.1

Traveling Salesman probleem

- Achtergrond voor pizza-koeriers.
- Gegeven zijn N steden en hun onderlinge afstanden.
- Vind de kortste route waarbij je elke stad precies n keer bezoekt.

Variabelen:

- Looptijd algoritme
- Lengte route

Vraag:

- Welke schaal?
- Nulpunt?

Descriptieve maten uit steekproef

Steekproef één variabele: $X_1,\ X_2,\ X_3,\ \ldots,\ X_N$

- Verhoudingsmaten
- Centrummaten
- Spreidingsmaten

Steekproef twee variabelen: $X_1,\ X_2,\ X_3,\ \ldots,\ X_N$ en $Y_1,\ Y_2,\ Y_3,\ \ldots,\ Y_N$

- Relatiematen

Verhoudingsmaten

Absolute frequenties:

- 7 (van de 12)

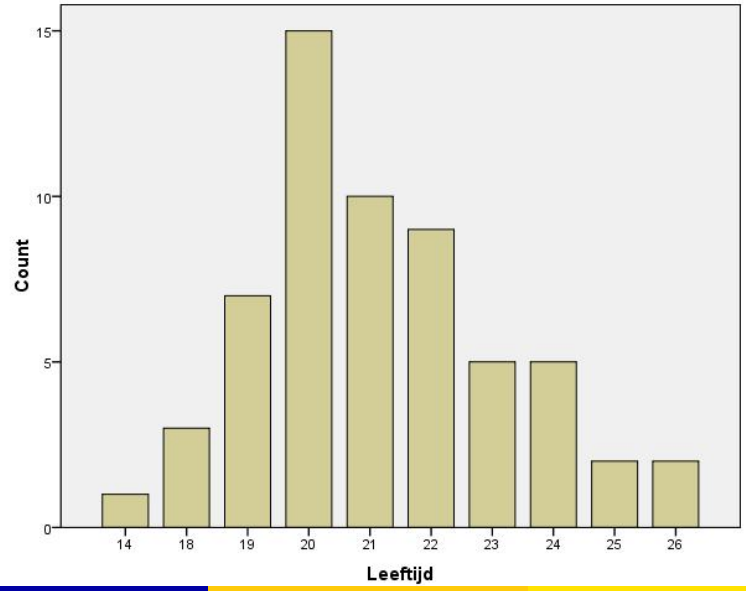
Relatieve frequenties:

- 3 op de 100
- 3%
- 0.03

Frequentietabel

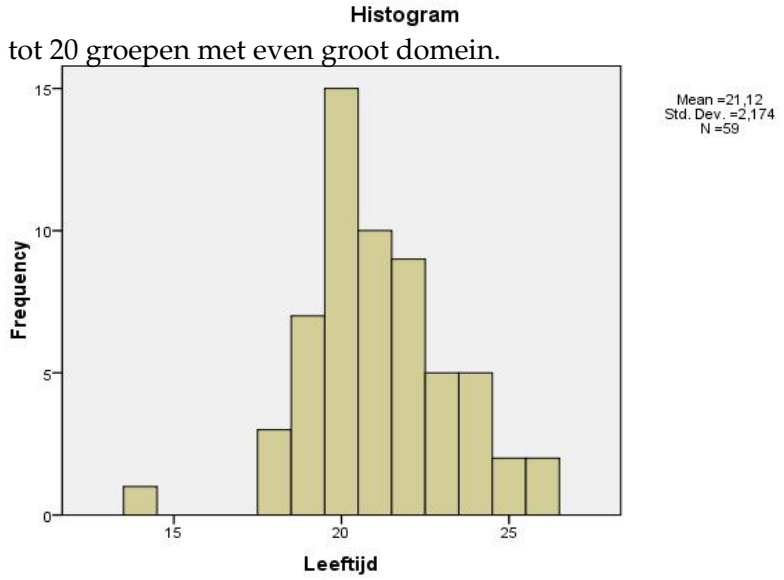
	Frequentie	Percentage	Cumul. perc.
4	14	11.38	11.38
5	15	12.20	23.58
6	19	15.45	39.02
7	22	17.89	56.91
8	24	19.51	76.42
9	19	15.45	91.87
10	10	8.13	100.00
totaal	123	100	100.00

Frequentieverdelingen: bar chart



Frequentieverdelingen: histogram

10 tot 20 groepen met even groot domein.



Scoreverdelingen: percentiel(score)

De score van het n_e percentiel (P_n) is de score waarbij tenminste n% in de verdeling lager of gelijk scoort, en tenminste 100-n% hoger of gelijk.

- Bijvoorbeeld $P_{90} = 189$ kan betekenen dat 90% van alle Nederlanders een lengte ≤ 189 cm heeft.
- Het meest gebruikte percentiel is de **mediaan** P_{50}: 50% van de observaties liggen links van de mediaan.
- Soms wordt ook nog gebruikt:
  - P_{25} (heet ook **eerste kwartiel**),
  - P_{75} (heet ook **derde kwartiel**).
- Pas op bij frequenties groter dan 1.

Frequentietabel

	Frequentie	Percentage	Cumul. perc.
4	14	11.38	11.38
5	15	12.20	23.58
6	19	15.45	39.02
7	22	17.89	56.91
8	24	19.51	76.42
9	19	15.45	91.87
10	10	8.13	100.00
totaal	123	100	100.00

Wat is de mediaan? 7

Wat is P_{25}? 6

Centrummaten: modus

Modus (“Eng: Mode”):
de waarde in de distributie die het meest voorkomt; de categorie met de hoogste frequentie

Ook mogelijk:

- Bimodaal (kameelverdeling)
- Multimodaal

Vond je het vak leuk? *Did you enjoy the course?*	**1**: niet / *no* 1 x **2**: 5 x **3**: 5 x **4**: 14 x **5**: wel / *yes* 10 x *n*=35, mean=3.77, stddev=1.1

Centrummaten: mediaan

- Het punt dat de waarnemingen door midden deelt, of
- De waarde die, in de ordening van laag naar hoog, hoort bij de middelste, of
- Het punt waarbij tenminste 50% lager of gelijk scoort, en tenminste 50% hoger of gelijk scoort.

Voorbeeld A: 1, 2, 3, 5, 6

Voorbeeld B: 1, 2, 3, 5, 6, 7

Sorteer van klein naar groot:

- Bij oneven aantal getallen: kies middelste
- Bij even aantal getallen: kies gemiddelde van middelste 2 (Excel)
- Of kies het hele interval ([3,5]) in Voorbeeld B als mediaan.

Centrummaten: gemiddelde

Indicatie van het *evenwichtspunt* van de meetwaarden.

- De som van alle waarden, gedeeld door het aantal waarden

Populatie:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Steekproef:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Student	Gevulde koeken
Jesse	2
Jasper	4
Jordy	4
Jorrit	3
Jarno	2
Gemiddelde	$\overline{X} = 3$

Als X_i heeft frequentie f_i, dan

$$\overline{X} = \frac{\sum f_i X_i}{n}$$

Test

Q: Op welk meetniveau kunnen de centrummaten toegepast worden?

- Mediaan
- Modus
- Gemiddelde

- Nominaal? (Modus)
- Ordinaal? (Modus, mediaan)
- Interval? (Modus, mediaan, gemiddelde)
- Ratio? (Modus, mediaan, gemiddelde)

Q: Welke centrummaat is gevoelig voor outliers (uitbijters)?

Voorbeelden

Voorbeeld A: 1 2 2 3 5 6 7 8 11

- Gemiddelde = ? 5
- Mediaan = ? 5

Voorbeeld B: 1 2 2 3 5 6 7 8 20

- Gemiddelde = ? 6
- Mediaan = ? 5

Spreidingsmaten

- Bereik
- Variantie
- Standaarddeviatie

Spreidingsmaat: bereik of "range"

- Hoogste waarde minus laagste waarde in een distributie
- Zegt niets over hoe het aantal scores verdeeld is binnen dat bereik.

Spreidingsmaten: *Populatie*variantie

- Gemiddelde kwadratische afwijking van het gemiddelde

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

- Standaarddeviatie

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Student	Gevulde koeken
Jesse	2
Jasper	4
Jordy	4
Jorrit	3
Jarno	2

Spreidingsmaten: *Steekproef*variantie

- *Schatting* voor populatievariantie σ^2
- Gemiddelde kwadratische afwijking van het gemiddelde

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

- Standaarddeviatie

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

- Vrijheidsgraden $df = n - 1$

Student	Gevulde koeken
Jesse	2
Jasper	4
Jordy	4
Jorrit	3
Jarno	2

$s^2 = \frac{(2-3)^2+(4-3)^2+(4-3)^2+(3-3)^2+(2-3)^2}{4} =$

$= \frac{4}{4} = 1$

Spreidingsmaten: Interquartile range

Herinnering:

- P_{25} = eerste kwartiel (Eng: quartile)
- P_{75} = derde kwartiel
- IQR = Interquartile range = $P_{75} - P_{25}$.

Relaties tussen twee of meer variabelen

Voor twee nominale variabelen:

Kruistabel:Tweedimensionaal frequentiediagram.

Voorbeeld

Gender vs. Smartphone

			Gender		
			Female	Male	Total
Smartphone	Yes	Count	380	2388	2768
		% within Smartphone	13.7%	86.3%	100.0%
Smartphone	No	Count	299	3333	3632
		% within Smartphone	8.2%	91.8%	100.0%
Total		Count	679	5721	6400
		% within Smartphone	10.6%	89.4%	100.0%

Relatiematen

- Twee variabelen: X en Y, met
  - Gemiddelden: $\overline{X}$ en $\overline{Y}$
  - Standaarddeviaties: s_X en s_Y
- Covariantie:

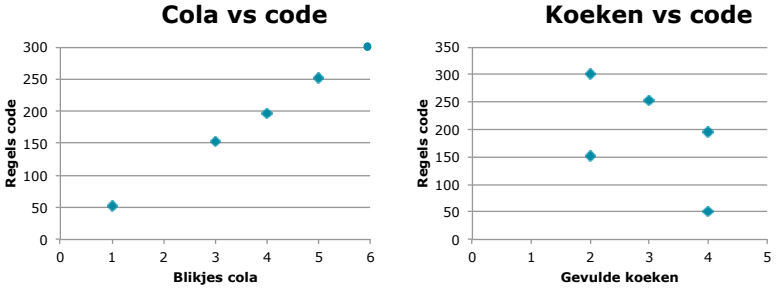
$$\text{cov}(x, y) = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{n - 1}$$

- Correlatie:

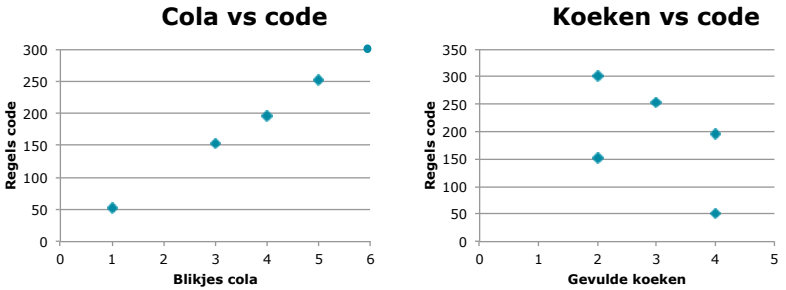
$$r = \frac{\text{cov}(x, y)}{s_X \cdot s_Y}, \ (-1 \leq r \leq 1)$$

Correlatie: voorbeeld

Student	Blikjes Cola	Gevulde koeken	Regels Code
Jesse	3	2	153
Jasper	4	4	196
Jordy	1	4	52
Jorrit	5	3	252
Jarno	6	2	301



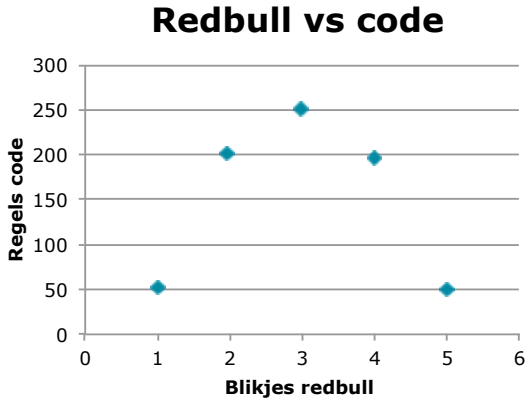
Correlatiematrix uit Excel



r	Cola	Koeken	Code
Cola	1		
Koeken	-0.5198	1	
Code	0.9995	-0.5398	1

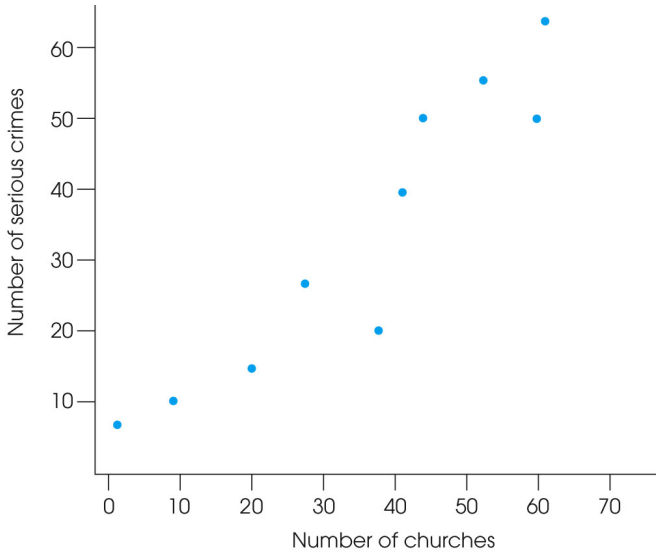
r dicht bij 1 of -1: lineair verband

Verband $\neq$ correlatie



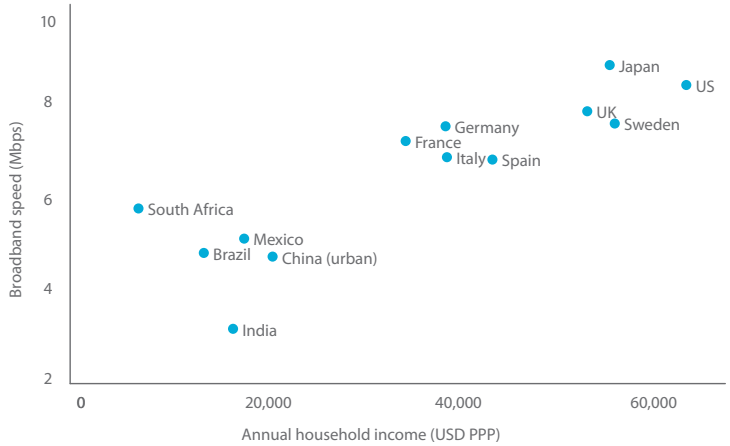
- "Kwadratisch" verband.
- Correlatie $r = 0$.

Correlatie $\neq$ Causaliteit



Correlatie $\neq$ Causaliteit (2)

Figure 2:The relationship between broadband speed and household income



Correlatie $\neq$ Causaliteit (2)

Samenvatting

- Variabelen
- Meetniveaus
- Beschrijvende statistiek

Volgende keer:

- Theoretische kansverdelingen
- Z-scores
- Normale verdeling
- Steekproefverdeling
- Centrale limietstelling