

Methoden van het Wetenschappelijk Onderzoek

Zin en onzin van statistiek

Statistiek komt ernstig over... of niet

- “Deze tandpasta helpt tegen caries in 1 op 2 gevallen.”
 - Het werd slechts geprobeerd op 4 personen.
- “De gemiddelde jonge informaticus verdient 1550 €.”
 - Dit is een sample, men heeft echt niet alle jonge informaticussen bevraagd. Mensen liegen vaak over hun loon. En wat is de spreiding?
- “Windows XP is stabielere dan andere besturingssystemen.”
 - Hoe meet je dat? Kan je dat sowieso objectief meten? Spelen hier geen “religieuze” kwesties mee? Wie waren je respondenten?
- “*Knack* wordt meer gelezen dan *Dag Allemaal*.”
 - En toch wordt *Dag Allemaal* meer verkocht? Zouden respondenten soms liegen over wat ze lezen?
- “De effectiviteit van dit product werd wetenschappelijk getest.”
 - Let op, er staat niet “wetenschappelijk *bewezen*”; waarschijnlijk helpt het product dus helemaal niet.

Waarom toch statistiek?

- Dient om onderzoek kwantitatief te onderbouwen.
- Dient om gegevens te “comprimeren” en zo een overzicht te bewaren.
- Twee soorten statistiek
 - Descriptieve statistiek
 - Inferentiële statistiek

“Statistiek is vervelend”

- Toch is statistiek de enige juiste manier om wetenschappelijk onderzoek correct te rapporteren.
- Een basiskennis van statistiek laat ook toe om wetenschappelijk onderzoek kritisch te **evalueren** en te **vergelijken**.

Het verzamelen van data

- Misschien wel het belangrijkste is het verzamelen van je data
 - **hier worden de meeste fouten (bewust of onbewust) gemaakt!**
- Voorbeeld
 - De NMBS die de stiptheid van de treinen nagaat.
 - De onveiligheidspoll van de federale regering.
 - Enquête naar seksuele ervaringen bij jongeren.
 - Enquête naar hygiëne.
 - Vaderschaptests.
- Die noemt men in de statistiek een “**bias**”.
- Wees bewust van bias, en vermeld steeds bias bij je statistiek!

Verzamelen van data dmv. enquête

- Hoe wordt de data vergaard?
- Hoeveel deelnemers zijn er/welk percentage van de populatie is dit?
- Exclusie criterium.
- Respons.
- Vertekent de niet-respons de resultaten?
- Hoe is de vragenlijst opgesteld?
- Betrouwbaarheid van de vragenlijst?
- Wie zijn de interviewers?
- Opleiding en opvolging van de interviewers?
- Uniformiteit van de omstandigheden van het interview?

Verzamelen van data

- Invloed van vraagstelling op resultaten:
 - Vindt u het goed dat u als werknemer op uw inkomen belast wordt om zo mensen te betalen die thuis zitten en niet werken ?
 - Vindt u dat de overheid mensen die werk willen maar geen vinden, moet ondersteunen?

Gegevens verzamelen: methodes

- Willekeurig verzamelen
 - Elk geval heeft een gelijke kans om opgenomen te worden in de statistiek.
 - Nodig: een welomlijnde populatie, een lijst met alle gevallen en een random generator.
- Systematisch verzamelen
 - Het eerste geval is willekeurig, de rest wordt volgens een bepaalde procedure gekozen. Bv. Start bij een willekeurig rolnr. en daarna increment met 10.
 - Vergroot niet echt een fout, behalve bij een vertekende populatie.
- Proportioneel gestratificeerd verzamelen
 - Per groep in de populatie neem je een aantal gevallen, gelijk aan de verhouding in de populatie.
- Disproportioneel gestratificeerd verzamelen.

Hoeveel samples?

- Hoeveel samples heb je nodig?
- Er is een vuistregel: **meer is beter**.
 - Natuurlijk is het verzamelen van gegevens beperkt door tijd, geld en ruimte.
- Vermeld altijd het aantal samples.
 - Dat is altijd het vergeten nummertje... Want zo kan je zelfs aantonen dat de kans dat je kop werpt 80% is.

Rapporteren van statistiek

- Het “gemiddelde”, rekenkundig gemiddelde, mediaan, of modus?

- Gemiddelde (mean):
$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

- Mediaan: sorteren, er liggen evenveel waarden onder als boven de mediaan.
- Modus: waarde welke het meest voorkomt (de populairste waarde).



\$45,000



\$15,000



\$10,000



← **ARITHMETICAL AVERAGE**

\$5,700



\$5,000



\$3,700



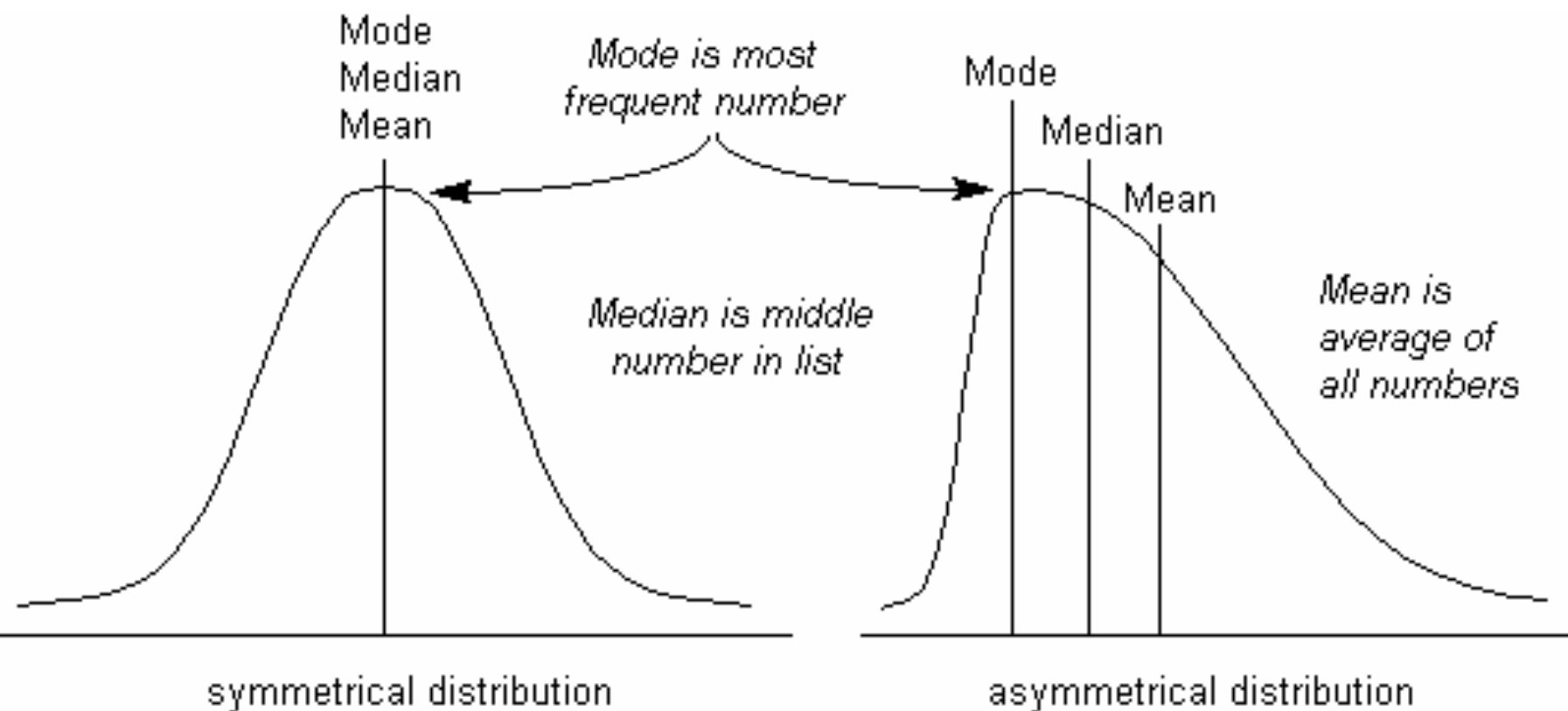
← **MEDIAN** (the one in the middle)
(12 above him, 12 below)

\$3,000



← **MODE**
(occurs most frequently)

\$2,000



Maat van dispersie

- Als je gemiddelden rapporteert, moet er steeds een maat van dispersie bij!
- Een goede maat van dispersie houdt rekening met
 - Alle gegevens.
 - Beschrijft de gemiddelde afwijking van de scores ten opzichte van het gemiddelde.
 - Neemt toe als de heterogeniteit van de gegevens toeneemt.

• Uitersten van de gegevens.

• Gemiddelde afwijking:

$$MD \equiv \frac{1}{N} \sum_{i=1}^N |x_i - \bar{x}|$$

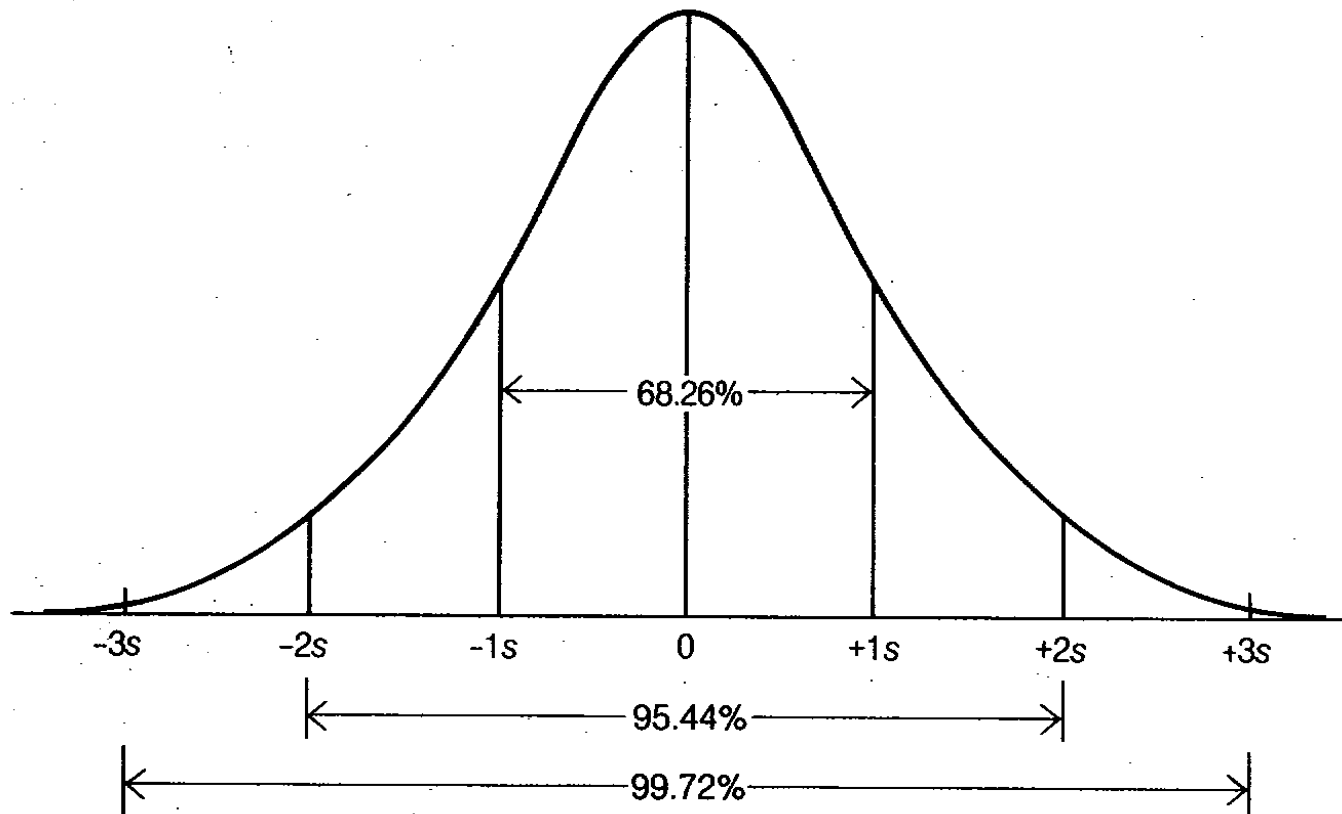
• Standaard variatie:

$$s_N^2 \equiv \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

• Standaard afwijking

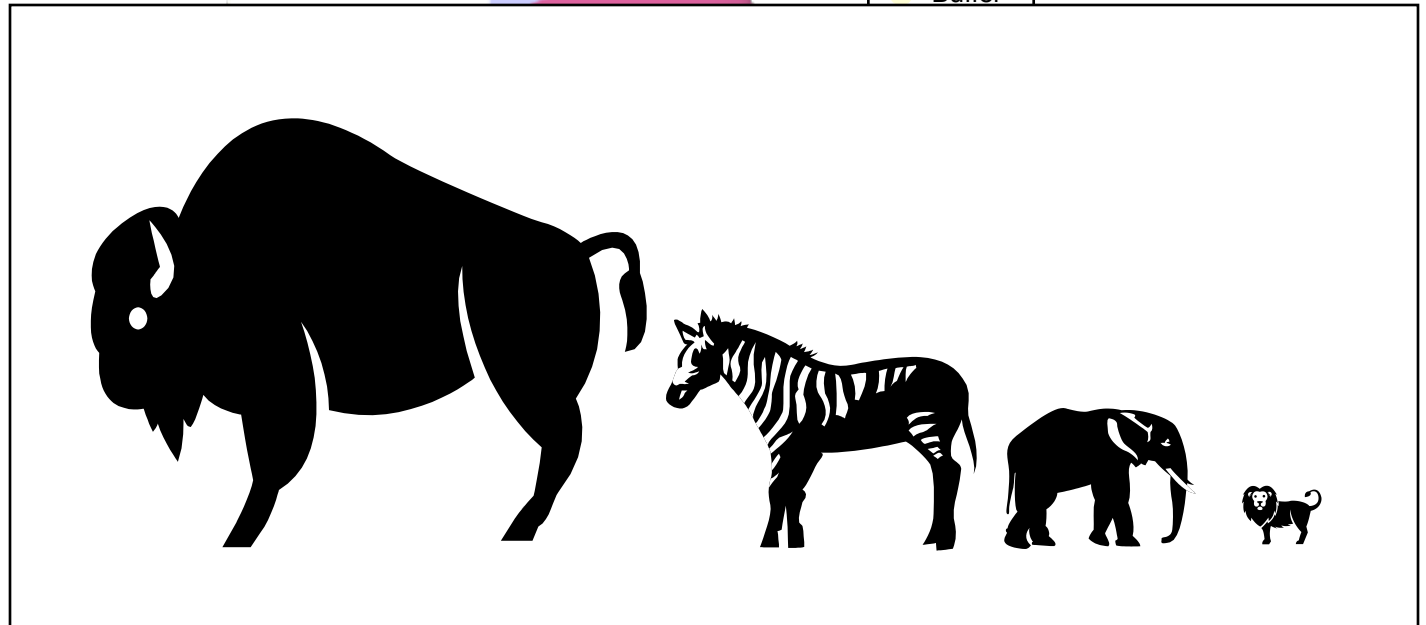
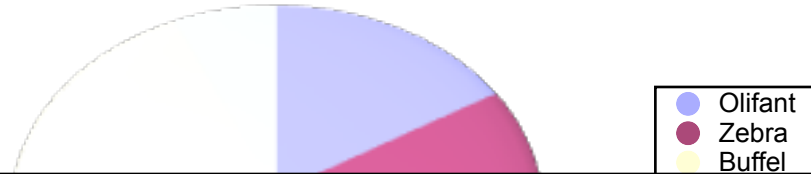
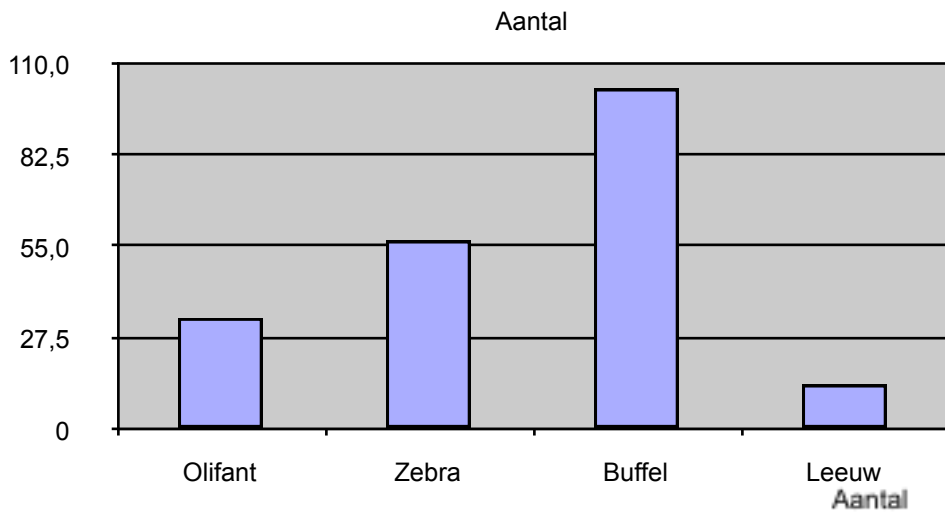
$$s_N = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Normaalverdeling



Grafieken

- Grafieken leren je veel in een oogopslag.
- Let echter op met het soort grafieken, sommigen zijn onbewust misleidend.

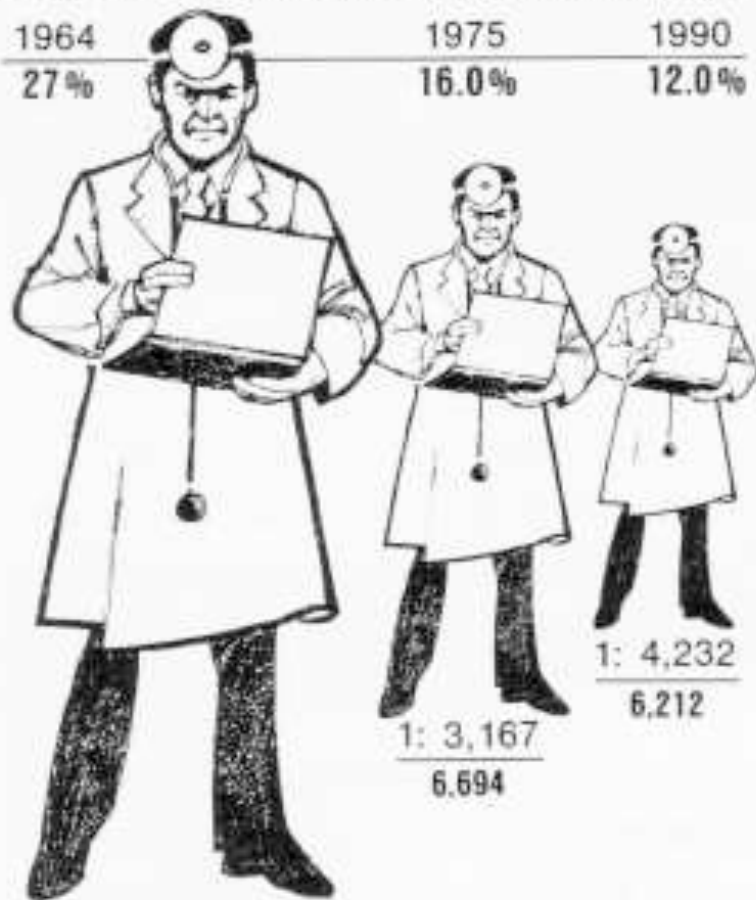


THE SHRINKING FAMILY DOCTOR

In California

Percentage of Doctors Devoted Solely to Family Practice

1964	1975	1990
27%	16.0%	12.0%



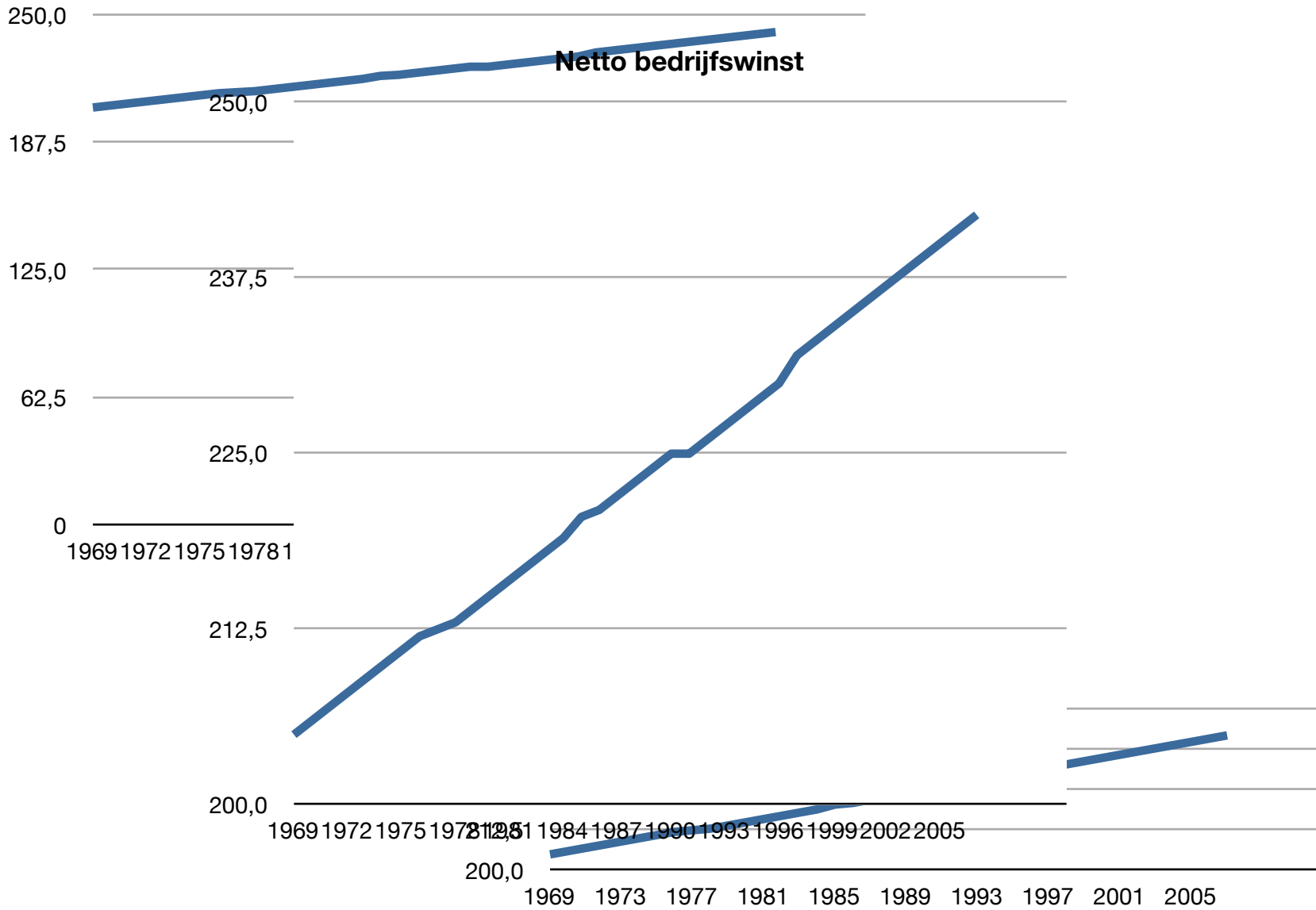
1: 3,167
6,694

1: 4,232
6,212

1: 2,247 RATIO TO POPULATION
8,023 Doctors

Los Angeles Times, August 5, 1979, p. 3.

Netto bedrijfswinst



Interpreteren van statistiek

- Statistiek reduceert data tot enkele bevattelijke getallen.
- Besluiten trekken uit een paar getallen is gevaarlijk.
 - Gemiddeld gezin heeft 2,4 kinderen.
- Opletten met uitspraken
 - Vaak wordt statistiek erg kortzichtig geïnterpreteerd.
 - Krantenartikelen, pop polls, prognoses, ...
- Het rekenkundig gemiddelde wordt al te vaak als “normaal” beschouwd. Als wat afwijkt van het rekenkundig gemiddelde is dan abnormaal.
 - Bv. Ontwikkeling van kinderen.
- Laat je niet intimideren door impressionante namen: “Universiteit X” of “Onderzoek Y” kunnen zich ook vergissen.

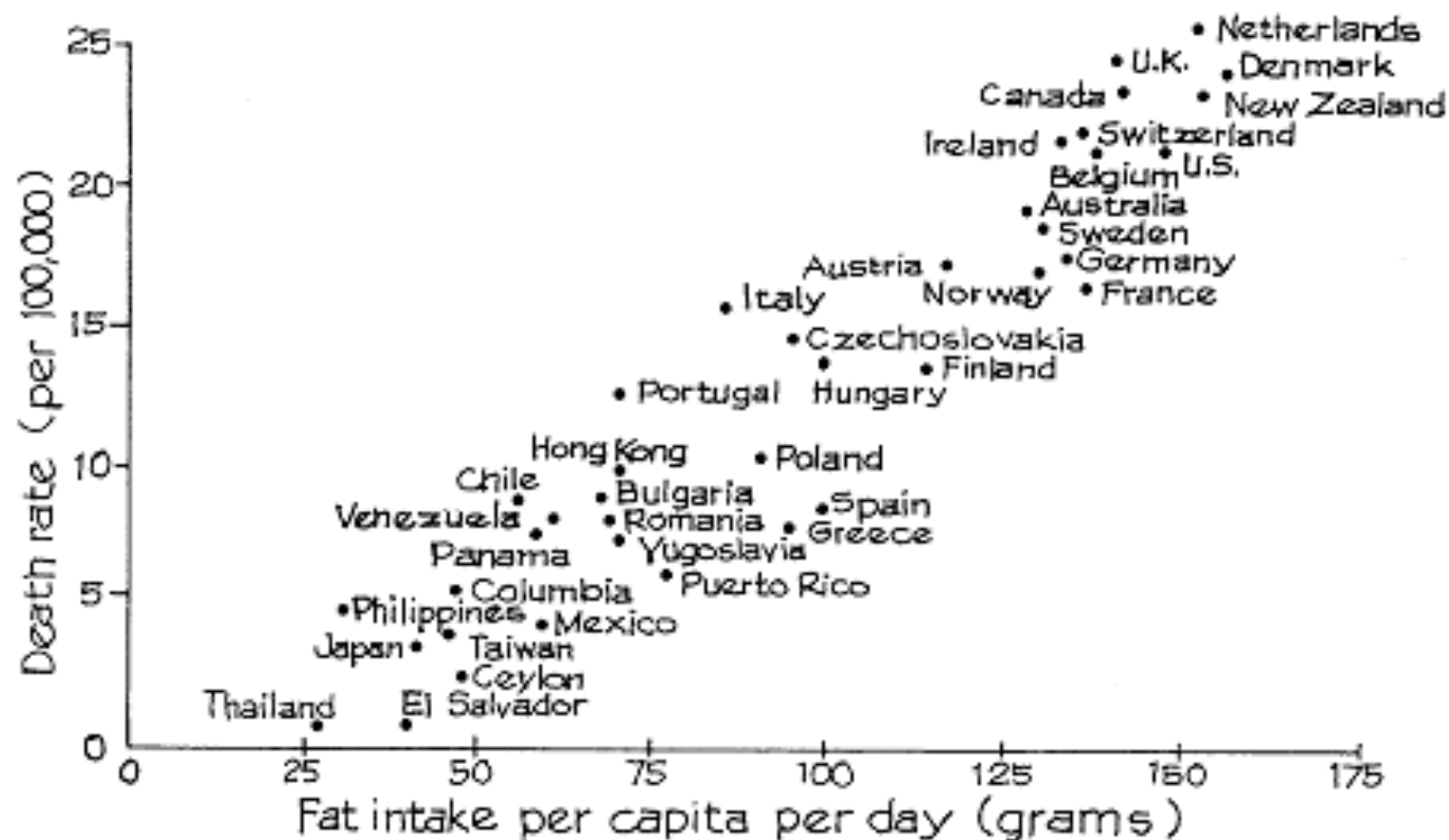
De “semi-attached figure”

- Als je iets niet echt statistisch kan aantonen, toon dan iets aan wat er op lijkt.
 - “In een streng laboratoriumonderzoek is aangetoond dat dit middel tegen verkoudheden gemiddeld $56\% \pm 7\%$ van de ziektekiemen doodde in 2.2 ± 0.5 dagen tijd.”
 - “Vier maal meer verkeersdoden om 7h dan om 19h”
- Het niet vergelijken van je data met een referentie.
 - Bv. “26% meer sap!!!”

Correlatie

- Een krachtig gereedschap in statistiek het berekenen van correlatie (met correlatiematrices, regressiecoëfficiënten, ...)
- Echter correlatie wordt al te vaak misbruikt om causale verbanden te leggen tussen twee factoren.
 - Sommige klinken aannemelijk: “kinderen van ouders met laag inkomen hebben slechtere studieresultaten”.
 - Maar je kan echt alles correleren als je wil: “leesvaardigheid hangt af van de schoenmaat”
- Als er een causaal verband is, dan is een correlatie.
- Een causaal verband werkt in twee richtingen, soms moeilijk te onderscheiden in welke richting.
- Soms is er geen rechtstreeks verband tussen de twee factoren.

Figure 8. Cancer rates plotted against fat in the diet, for a sample of countries.



Source: K. Carroll, "Experimental evidence of dietary factors and hormone-dependent cancers," *Cancer Research* vol. 35 (1975) p. 3379. Copyright by *Cancer Research*. Reproduced by permission.

Reïficatie

- Reïficatie = “verdingelijking”
- Een complex fenomeen wordt in een numerieke voorstelling gegoten, waarbij alle nuancering verloren gaat.
- Bv.
 - IQ
 - Temperatuur

Getallen na de komma

- “Als er veel getallen na de komma staan, dan moet het wel degelijk onderzoek zijn.”
- Als je een getal geeft, rond dan steeds netjes af.
 - Vuistregel: twee significante cijfers van de standaard deviatie
gemiddelde = 98.2346321 st.dev. = 0.00342879
wordt 98.2346 ± 0.0034

Aan de slag met statistiek

- Als je statistiek wenst te gebruiken voor het rapporteren van onderzoek.
 - Gebruik dan de statistische methodes van gelijkaardig onderzoek, dit laat vergelijking toe.
 - Zoek een goed recht-toe-recht-aan boek.

Kritische vragen bij het interpreteren van statistiek

- Who says so?
- How does he know?
- What's missing?
- Did somebody change the subject?
- Does it make sense?

- Benjamin Disraeli: “There are three kinds of lies: lies, damn lies, and statistics.”