

Statistiek voor A.I.

College 1

Dinsdag 11 September 2012

- Website: <http://phil.uu.nl/statistiek/>
- Applied Statistics for the Behavioral Sciences - 5th edition, Dennis E. Hinkle, William Wiersma, and Stephen G. Jurs.
- Introduction to Probability, Charles M. Grinstead and J. Laurie Snell: http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html.
- Syllabi *Bayesiaans leren* en *Schatten* (op website).
- Optioneel: software *R*.

Kansrekening en Statistiek?

- Wat is de kans dat ik de loterij win? [Kansspelen](#)
- Bevordert luieren de fantasie? [Psychologie](#)
- Hoe leert mijn spamfilter wat spam is? [Informatica](#)
- Does God play dice with the universe? [Natuurkunde](#)
- Is Lucia de B. schuldig? Nee. [Rechtspraak](#)
- Werkt paracetamol? [Geneeskunde](#)
- Wat is een kans? [Filosofie](#)

Deductief redeneren (logica): Uit dit volgt dat.

Inductief redeneren (kansrekening):

Uit dit volgt met deze waarschijnlijkheid dat.

Als dit gebeurt is de kans dat dat gebeurt zo en zoveel.

- 1 Beschrijvende statistiek.
- 2 Deductieve statistiek:
 - Kansrekening/Bayesiaanse statistiek
 - Klassieke/orthodoxe statistiek.

FOKKE & SUKKE
VOELEN DAT AAN HUN WATER

DE KANS DAT VRIJWEL ALLE
HOOGLERAREN STATISTIEK HET
EENS ZIJN

...IS NATUURLIJK
HEEL ERG KLEIN.



RGvT

www.foksuk.nl

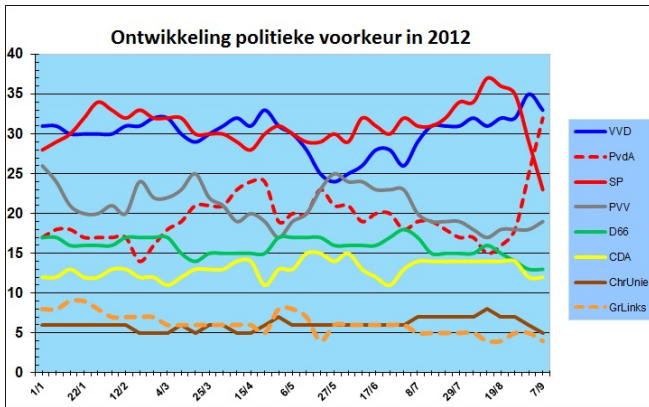
1 Beschrijvende statistiek

1 Beschrijvende statistiek

Vandaag:

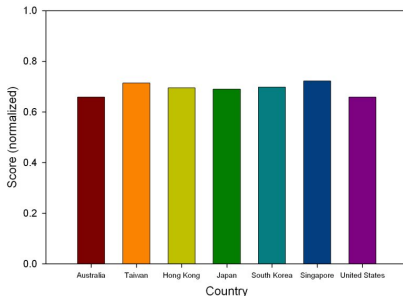
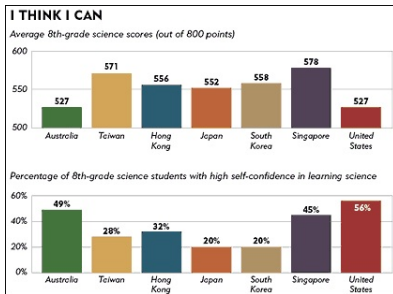
- Populatie en steekproef
- Maten
- Weergave
- Standaardscores
- Normale verdeling

Verkiezingen 2012



Suggestie

Dezelfde resultaten gepresenteerd via scores en standaardscores:



Populatie en steekproef

Definitie Een *populatie* bevat alle elementen van een bepaalde groep.

Een *parameter* is een eigenschap van de elementen in de populatie.

Een *steekproef* is een deelverzameling van de groep.

Een *statistiek* is een eigenschap van de elementen in de steekproef.

Een *constante* is een eigenschap die hetzelfde is voor alle elementen van de populatie.

Een *variabele* is een eigenschap die verschillende waarden/scores kan aannemen voor verschillende elementen van de populatie.

Voorbeeld

- De verzameling van alle Nederlanders is een populatie.
- Elke groep Nederlanders is een steekproef van deze populatie.
- De gemiddelde lengte van alle Nederlanders is een parameter.
- De gemiddelde lengte van een groep Nederlanders is een statistiek.
- De eigenschap *Nederlander* is een constante voor de populatie.
- De eigenschap *vrouw* is een variabele voor de populatie.

Variabelen worden onderverdeeld in vier groepen:

- nominale schaal
- ordinale schaal
- intervalschaal
- ratioschaal.

Weergave

- Frequentiedistributie.
- Frequentiedistributie bij intervallen.
- Cumulatieve frequentiedistributie.
- Boxplot.

- Functies.
- Scatterplots.
- Histogrammen.
- Frequentiepolygonen.
- Cumulatieve frequentiedistributie.

Maten

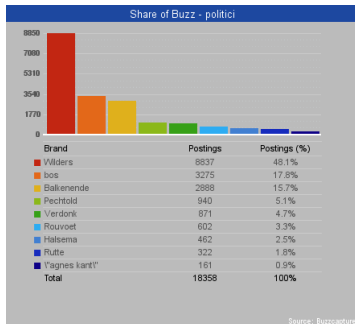
Definitie De *modus* is de score met de hoogste frequentie.

Het x *percentiel* is dat punt in de verdeling waarop of beneden x procent van de waarden valt.

De *mediaan* is het 50e percentiel (de score waar beneden de helft van de waarden valt). Bij een oneven aantal waarden is het de middelste score, bij een even aantal waarden het midden tussen de twee middelste waarden.

Maten: percentiel, modus en mediaan

Voorbeeld Van <http://www.politiekedialoog.nl>: Nederlanders discussiëren op het internet over allerlei zaken die hen bezighouden. Dus ook over politiek.



De modus is Wilders.

De modus is een van de weinige maten die toepasbaar is op nominale distributies.

Er zijn geen percentielen, en dus ook geen mediaan, omdat de schaal geen getalswaarden zijn.

Voorbeeld Het aantal verkochte ijsjes in een ijskraam per smaak per dag:

vanille	pistache	straciatelle	framboos
100	110	80	110

De modus is/zijn pistache en framboos. De modus hoeft niet uniek te zijn.

Er zijn geen percentielen, en dus ook geen mediaan, omdat de schaal nominaal is.

Maten: percentiel, modus en mediaan

Voorbeeld Het aantal computers per gezin in een bepaald dorp:

aantal computers	0	1	2	3	4	5	6	7
frequentie	1	1	3	2	0	1	0	1

Er zijn 9 scores. De modus is 2. De mediaan is de middelste waarde, de waarde van de 5e score: 2. Wanneer de waarden op een rij gezet worden is dat makkelijk te zien:

012223357.

7 waarden vallen op en beneden score 3. 7 is 78% van 9, dus het 78e percentiel is 3.

Voorbeeld Het aantal computers per gezin in een bepaald dorp:

aantal computers	0	1	2	3	4	5	6	7
frequentie	1	1	3	2	0	1	0	2

Er zijn 10 waarden. De mediaan is het midden tussen de twee middelste waarden (de waarden van de 5e en de 6e score), die 2 en 3 zijn. De mediaan is dus 2.5.

Definitie Het *gemiddelde* is het rekenkundig gemiddelde van de waarden: de som van de waarden gedeeld door het aantal:

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

waarbij X_i de waarden zijn en n het aantal waarnemingen is. Het gemiddelde van de populatie wordt vaak aangegeven met μ .

Merk op: Als de data gegeven zijn via een frequentie distributie, dan kan het gemiddelde zo berekend worden:

$$\bar{X} = \frac{\sum_{i=1}^n f_i X_i}{n},$$

waarbij X_i de waarden (klassen) zijn, n het aantal waarnemingen is en f_i de frequentie (het aantal) van de waarnemingen in klasse X_i .

Maten: gemiddelde, modus en mediaan

Voorbeeld Gegevens van *TNS NIPO*. Het aantal grove woorden per uur in 2004:

NOS	0.3
IKON	1.5
KRO	1.7
Veronica	2.5
VARA	4.1
BNN	5.1

Modus: elke score.

Mediaan:

$$\frac{1.7 + 2.5}{2} = 2.1.$$

Gemiddelde:

$$\frac{0.3 + 1.5 + 1.7 + 2.5 + 4.1 + 5.1}{6} = 2.533333.$$

Maten: gemiddelde

Voorbeeld Het aantal computers per huis in een dorp met 9 huizen, waarbij de categorieën de waarden bevatten:

huizen	H1	H2	H3	H4	H5	H6	H7	H8	H9
aantal computers	2	3	0	2	5	2	3	1	7

Het gemiddelde aantal computers:

$$\frac{2 + 3 + 0 + 2 + 5 + 2 + 3 + 1 + 7}{9} = \frac{25}{9} = 2.8.$$

Het aantal computers per huis in hetzelfde dorp, waarbij de categorieën de waarden zijn en hun inhoud de frequentie van het voorkomen van de score:

aantal computers	0	1	2	3	4	5	6	7
aantal huizen	1	1	3	2	0	1	0	1

Het gemiddelde aantal computers:

$$\frac{0 \cdot 1 + 1 \cdot 1 + 2 \cdot 3 + 3 \cdot 2 + 0 \cdot 4 + 1 \cdot 5 + 0 \cdot 6 + 1 \cdot 7}{9} = \frac{25}{9} = 2.8.$$

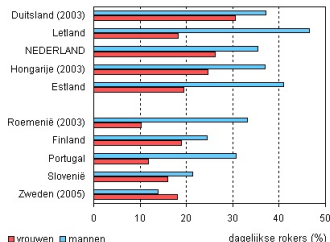
Maten: afwijking

Definitie Voor elke score (waarde) X_i is de *afwijking (van het gemiddelde)*, x_i , het verschil met het gemiddelde:

$$x_i = (X_i - \bar{X}).$$

Maten: afwijking

Voorbeeld Dagelijkse rokers in 10 landen (2006):



Het gemiddelde percentage van de dagelijkse rokers onder de mannen in de 10 landen:

$$\bar{X} = \frac{37 + 47 + 35 + 37 + 41 + 32 + 25 + 31 + 22 + 13}{10} = 32\%.$$

De afwijking van de score van Duitsland is $37\% - 32\% = 5\%$. Die van Finland is $25 - 32 = -7\%$. Evenzo voor de andere landen:

Nederland	Hongarije	Estland	Roemenië	Portugal	...
3	5	9	0	-1	...

Stelling De som van de afwijkingen van het gemiddelde is 0.

Bewijs

$$\sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n X_i - n\bar{X} = \sum_{i=1}^n X_i - n \frac{\sum_{i=1}^n X_i}{n} = \sum_{i=1}^n X_i - \sum_{i=1}^n X_i = 0.$$

Definitie De *gemiddelde afwijking* is het gemiddelde van de absolute waarden van de afwijkingen van het gemiddelde:

$$\frac{\sum_{i=1}^n |X_i - \bar{X}|}{n}.$$

Voorbeeld Het aantal kinderen per persoon van 4 personen A, B, C, D :

A	B	C	D
0	2	1	2

Het gemiddelde is $\frac{5}{4}$. De afwijkingen van het gemiddelde zijn

$$x_A = -\frac{5}{4} \quad x_B = \frac{3}{4} \quad x_C = -\frac{1}{4} \quad x_D = \frac{3}{4}.$$

De gemiddelde afwijking is

$$\frac{\frac{5}{4} + \frac{3}{4} + \frac{1}{4} + \frac{3}{4}}{4} = \frac{\frac{12}{4}}{4} = \frac{3}{4}.$$

Definitie De *variantie* van een populatie ter grootte N is het gemiddelde van de gekwadrateerde afwijkingen van het gemiddelde μ :

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

De *standaard afwijking* van een populatie is de wortel uit de variantie:

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

De *variantie* van een steekproef ter grootte n is

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

De *standaard afwijking* van een steekproef is de wortel uit de variantie:

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

Maten: variantie

Voorbeeld De populatie is een vijver met 5 vissen. De lengtes in cm. zijn:

v_1	v_2	v_3	v_4	v_5
6	5	9	4	6

Gemiddelde van de populatie: $\mu = 6$ cm. Variantie van de populatie:

$$\sigma^2 = \frac{(6-6)^2 + (5-6)^2 + (9-6)^2 + (4-6)^2 + (6-6)^2}{5} = 2.8.$$

Standaardafwijking van de populatie: $\sigma = \sqrt{2.8} = 1.67332$.

Er wordt een steekproef uit de vijver gehengeld bestaande uit v_2 , v_4 en v_5 .
Gemiddelde van de steekproef:

$$\overline{X} = \frac{5 + 4 + 6}{3} = 5 \text{ cm.}$$

Variantie van de steekproef:

$$s^2 = \frac{(5-5)^2 + (4-5)^2 + (6-5)^2}{3-1} = \frac{0+1+1}{2} = 1.$$

Standaardafwijking van de steekproef : $s = \sqrt{1} = 1$.

Stelling De som van de gekwadrateerde afwijkingen van het gemiddelde is minimaal.

Bewijs We bewijzen dat voor iedere steekproef, en het geldt dus in het bijzonder voor de gehele populatie. Laat

$$f(y) = \sum_{i=1}^n (X_i - y)^2 = \sum_{i=1}^n (X_i^2 - 2X_i y + y^2) = \sum_{i=1}^n (X_i^2 - 2X_i y) + ny^2.$$

$$f'(y) = \sum_{i=1}^n -2X_i + 2ny = 2\left(-\sum_{i=1}^n X_i + ny\right) \quad f'(\bar{X}) = 0.$$

Dus f is minimaal in $\bar{X}$.

Notatie:

	populatie	steekproef
aantal	N	n
gemiddelde	μ	$\bar{X}$
variantie	σ^2	s^2
standaard afwijking	σ	s

Standardscores

Definitie Gegeven een steekproef met gemiddelde $\bar{X}$ en standaardafwijking s , is de *standaardscore* (*standard score*) van waarde/score X :

$$\frac{X - \bar{X}}{s}.$$

De standaarscore van waarde X_i wordt vaak aangegeven met z_i of x_i .

Standaardcores

Voorbeeld Waarden: $X_1 = 2$ $X_2 = 4$ $X_3 = 6$. Gemiddelde: $\bar{X} = 4$.

$$s^2 = \frac{(2-4)^2 + (4-4)^2 + (6-4)^2}{2} = 4 \quad s = \sqrt{4} = 2.$$

Standaardcores:

$$\text{van } X_1: x_1 = \frac{2-4}{2} = -1, \quad \text{van } X_2: x_2 = \frac{4-4}{2} = 0, \quad \text{van } X_3: x_3 = \frac{6-4}{2} = 1.$$

Voorbeeld Waarden: $Y_1 = 1$ $Y_2 = 2$ $Y_3 = 3$ (de vorige gedeeld door 2).
Gemiddelde: $\bar{X} = 2$.

$$s^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{2} = 1 \quad s = \sqrt{1} = 1.$$

Standaardcores:

$$\text{van } Y_1: y_1 = \frac{1-2}{1} = -1, \quad \text{van } Y_2: y_2 = \frac{2-2}{1} = 0, \quad \text{van } Y_3: y_3 = \frac{3-2}{1} = 1.$$

Merk op: De verdelingen van X en Y lijken zodanig op elkaar dat hun standaardcores gelijk zijn.

Stelling De verdeling van de standaardscores van een of andere verdeling X heeft altijd gemiddelde 0. X_i ligt op n standaardafwijkingen van $\bar{X}$ precies dan als geldt $x_i \in \{n, -n\}$.

Finis