Statistiek

Powerpoint op toledo

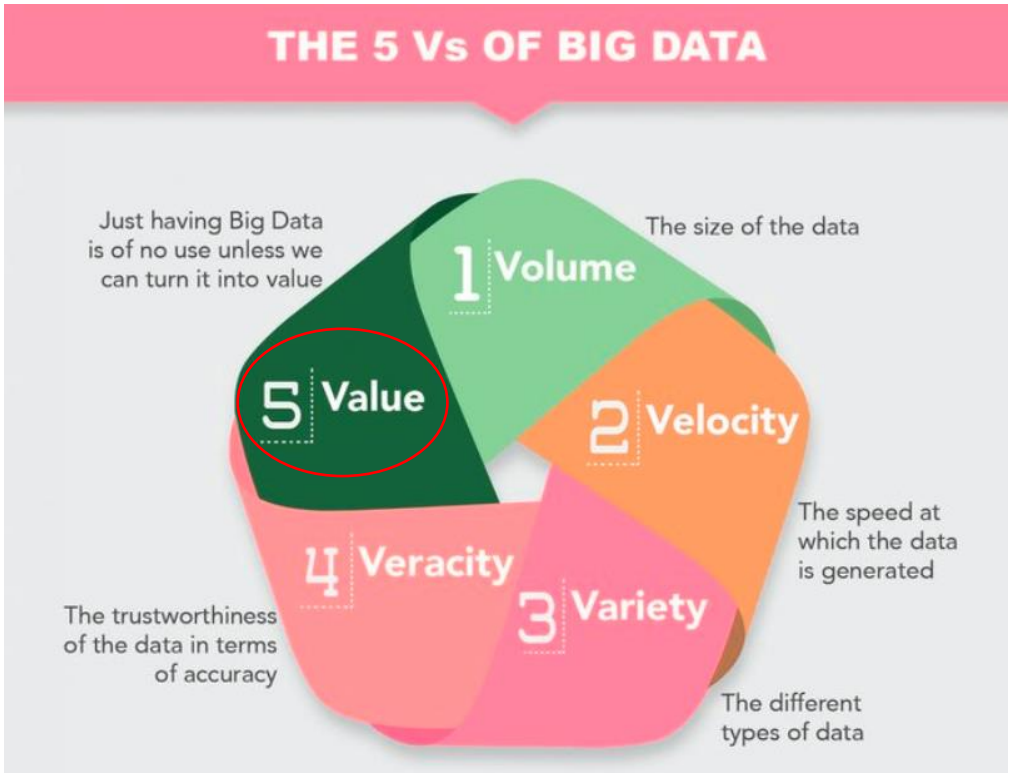
Legende:

Definities: geel gekleurd

Oefeningen: blauw gekleurd

Introductie

Big Data: 1 of meer datasets die te groot zijn voor reguliere databasemanagementsystemen



Populatie: verzameling van eenheden die we bestuderen

Steekproef: deelverzameling van eenheden van de populatie

Kwantitatief: gegevens waarmee we kunnen rekenen en het zinvol is om met te rekenen

- Continu: alle waarden in een bepaald interval mogelijk
- Discreet: bepaald aantal waarden mogelijk

Kwalitatief: gegevens waarmee we niet kunnen rekenen of het niet zinvol is te rekenen

Oefening Kwalitatief - Kwantitatief:

a) 2
b) 1C
c) 2
d) 2
e) 1D

f) 2
g) 2
h) 1D
i) 2
j) In score: 1D
Zonder score: 2
k) 2
l) 1C

Opdracht:

a) Mensen in een republiek.
b) Kwalitatief
c) 2000 Inwoners
d) Neen, niet iedereen heeft een telefoon

Gemiddelde: alle waarden delen door de hoeveelheid waarden
Wanneer uiterste gegevens uitzonderlijk verschillen zal het gemiddelde beïnvloed worden!

Mediaan: middelste waarde van een reeks

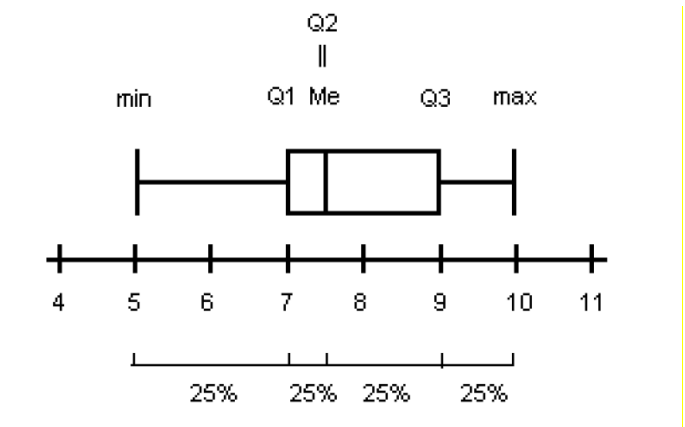
Modus: Meest voorkomende waarde

Opdrachten:

a) 40.85
b) 40.85
c) Geen
A. 8.375, 8.5
B. 8.25, 8
C. 6.25, 5.5
a. Gemiddelde: 7.8125 (alle gemiddelden maal de grootte van hun steekproef delen door de totale steekproef)
b. Mediaan: Geen snelle manier, alle gegevens samen zetten en zien

Bereik: verschil tussen grootste en kleinste gegeven

Interkwartielafstand: gegevens in 4 gelijke groepen verdelen, lengte tussen eerste en derde kwartiel (Q1 en Q3)



Standaarddeviatie:

i	X_i	$X_i - \mu$
1	10	3
2	9	2
3	7	0
4	6	-1
5	6	-1
6	6	-1
7	5	-2
8	8	1
9	8	1
10	7	0
11	2	-5
12	10	3
13	5	-2
14	7	0
15	9	2

Deviaties: verschil tussen elke waarde en het gemiddelde
Variantie: gemiddelde van kwadraten van de deviaties
Standaarddeviatie σ: vierkantswortel van de variatie

Indien steekproef kleiner is als 30 (indien volledige populatie niet toepassen):

$$s^2 = \frac{\sum (X_i - \overline{x})^2}{n-1}$$

$$s = \sqrt{\frac{\sum (X_i - \overline{x})^2}{n-1}}$$

Opdrachten:

a) 49.6
b) 49
c) 50
d) 31
e) 76.93333
f) 8.771165
g) 6
h) 37 - 45 - 49 - 51 – 68
A. B, gegevens liggen verder uit elkaar
B. Juist, want de standaardafwijking is altijd kleiner als het bereik
1. 12,13 11,14 10,15 12.5,12.5 etc.

Correlatie en Regressie

Correlatie: een maat voor de sterkte van een verband tussen twee grootheden

Gemiddelde van alle x-waarden is μ_x /μ_y

Correlatiecoëfficiënt R: Indien deze 1 or -1 is, is er volledige correlatie

$$R(x,y) = \frac{Cov(x,y)}{\sigma_x \cdot \sigma_y}$$

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x)\cdot(y_i - \mu_y)}{n}$$

Berekening zie opdracht 1 voor uitwerking:

1. Berekening gemiddeldes x- en y-waarden
2. Berekening deviaties x- en y-waarden
3. Berekening van de producten van de deviaties $(x_i - \mu_x)$ $(y_i - \mu_y)$
4. Berekening gemiddelde van deze producten
5. Eindproduct delen door product standaardafwijkingen

Opdracht 1 en eigen voorbeeld om als uitleg: 0.935241

10	0	0	11	0.7	0.49		0
12	2	4	14	3.7	13.69		7.4
8	-2	4	9	-1.3	1.69		2.6
13	3	9	13	2.7	7.29		8.1
9	-1	1	9	-1.3	1.69		1.3
10	0	0	9	-1.3	1.69		0
7	-3	9	8	-2.3	5.29		6.9
14	4	16	14	3.7	13.69		14.8
11	1	1	10	-0.3	0.09		-0.3
6	-4	16	6	-4.3	18.49		17.2
10		60	10.3		64.1		5.8
		6	2.44949		6.41	2.531798	
						6.201613	0.935241
Waarden	Gem	Deviaties	kwadraat	Som	Variatie	Standaard Deviatie	
(xi - μx) (yi - μy)		Product σ	R(x,y)				

Opdracht 2: 0.7450994

Opdracht 3: 0.94

Regressielijn berekenen:

Voor de regressierechte $y = ax + b$ door de punten $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ geldt:

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

en $\quad b = \bar{y} - a\bar{x}$

Voorbeeld:

$\mu_x = (1 + 2 + 3 + 4 + 5) / 5 = 3.0$
$\mu_y = (2 + 5 + 6 + 11 + 13) / 5 = 7.4$
$S_{xx} = (1 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10.0$
$S_{xy} = (2 - 7.4) (1 - 3) + (5 - 7.4) (2 - 3) + (6 - 7.4) (3 - 3) + (11 - 7.4) (4 - 3) + (13 - 7.4) (5 - 3) = 28$

De schattingen voor a en b worden dus :
$a = S_{xy} / S_{xx} = 28.0 / 10.0 = 2.8$
$b = \mu_y - a \cdot \mu_x = 7.4 - 2.8 * 3.0 = -1.0$

De regressielijn wordt dus:
$y = 2.8x - 1.0$

Regressiecoëfficiënt β = a

De regressielijn voorspeld ongeveer waar bepaalde punten zullen liggen

Opdracht 1:

- Y = 37.5704x + 5.495252
- 20.52341 mm^2
- 0.678852 g

Opdracht 2: thuis doen