Zomercursus Wiskunde A

Week 4, les 1

Gerrit Oomens
G.Oomens@uva.nl

Korteweg-de Vries Instituut voor Wiskunde
Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam

1 augustus 2011

Soorten data

serienr.	kleur	gewicht (g)	productiejaar	drijfvermogen
120	geel	21	2003	gemiddeld
345	groen	42	2005	laag
81	geel	24	2009	hoog

Kwalitatief: eigenschappen

- Nominaal: ongeordende labels.
  Voorbeeld: kleur, serienummer.
- Ordinaal: geordende labels, verschillen hebben geen betekenis.
  Voorbeeld: drijfvermogen, "eens" / "neutraal" / "oneens".

Kwantitatief: getallen (discreet/continu)

- Interval: geordend, verschillen hebben betekenis. Nulpunt is willekeurig en heeft geen betekenis.
  Voorbeeld: productiejaar, temperatuur in Celsius.
- Ratio: geordend met vast nulpunt. Verschillen en quotiënten hebben betekenis.
  Voorbeeld: gewicht, lengte, temperatuur in Kelvin.

Kwantitatieve data

Beschouw de volgende dataset (leeftijden van studenten)

18.2	21.1	18.0	19.5	20.1	19.5	19.3	19.8
19.2	19.8	18.1	18.4	18.3	19.7	0.1	20.4
19.4	18.7	19.3	18.9	19.4	20.6	18.5	19.3
18.7	20.1	21.3	19.4	21.6	18.1	20.1	18.1

We sorteren de data:

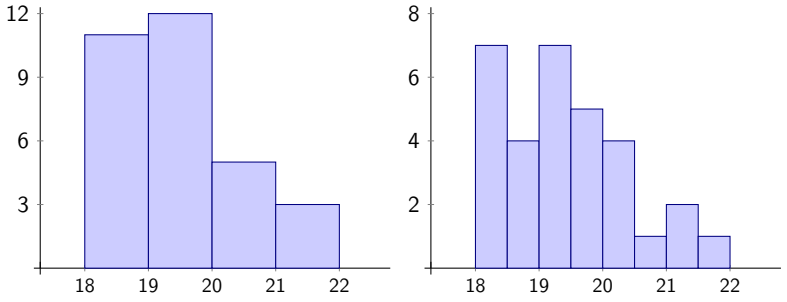
0.1	18.0	18.1	18.1	18.1	18.2	18.3	18.4
18.5	18.7	18.7	18.9	19.2	19.3	19.3	19.3
19.4	19.4	19.4	19.5	19.5	19.7	19.8	19.8
20.1	20.1	20.1	20.4	20.6	21.1	21.3	21.6

Het getal 0.1 is een *uitschieter*: een getal dat niet binnen het patroon van de data valt.

Histogram

Hoe geven we de data grafisch weer?

0.1	18.0	18.1	18.1	18.1	18.2	18.3	18.4
18.5	18.7	18.7	18.9	19.2	19.3	19.3	19.3
19.4	19.4	19.4	19.5	19.5	19.7	19.8	19.8
20.1	20.1	20.1	20.4	20.6	21.1	21.3	21.6



Numerieke samenvattingen

Sommige getallen kunnen ons informatie over de data geven.

0.1	18.0	18.1	18.1	18.1	18.2	18.3	18.4
18.5	18.7	18.7	18.9	19.2	19.3	19.3	19.3
19.4	19.4	19.4	19.5	19.5	19.7	19.8	19.8
20.1	20.1	20.1	20.4	20.6	21.1	21.3	21.6

Het *gemiddelde*:

$$\bar{x} = \frac{0.1 + 18.0 + 18.1 + \cdots + 21.1 + 21.3 + 21.6}{32} = 18.819.3.$$

De *mediaan* of middelste getal. In het geval dat er geen middelste getal is (even aantal waarnemingen), nemen we het gemiddelde van de twee middelste getallen:

$$\text{mediaan} = \frac{19.3 + 19.4}{2} = 19.35.$$

Het gemiddelde is gevoelig voor uitschieters, de mediaan niet.

Numerieke samenvattingen

Sommige getallen kunnen ons informatie over de data geven.

18.0	18.0	18.1	18.1	18.1	18.2	18.3	18.4
18.5	18.7	18.7	18.9	19.2	19.3	19.3	19.3
19.4	19.4	19.4	19.5	19.5	19.7	19.8	19.8
20.1	20.1	20.1	20.4	20.6	21.1	21.3	21.6

Het *gemiddelde*:

$$\bar{x} = \frac{18.0 + 18.0 + 18.1 + \cdots + 21.1 + 21.3 + 21.6}{32} = 18.819.3.$$

De *mediaan* of middelste getal. In het geval dat er geen middelste getal is (even aantal waarnemingen), nemen we het gemiddelde van de twee middelste getallen:

$$\text{mediaan} = \frac{19.3 + 19.4}{2} = 19.35.$$

Het gemiddelde is gevoelig voor uitschieters, de mediaan niet.

Locatie maten: gemiddelde en mediaan

Het gemiddelde van een verzameling data $x_1, x_2, \ldots, x_n$ wordt gegeven door

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Dus het gemiddelde van de dataset [3 4 2 2 5] is

$$\bar{x} = \frac{3 + 4 + 2 + 2 + 5}{5} = \frac{16}{5}.$$

Wat is de mediaan? Eerst ordenen we de data: [2 2 3 4 5], dan nemen we de middelste waarde als het aantal waarnemingen oneven is en anders het gemiddelde van de middelste twee. In dit geval is de mediaan gelijk aan 3.

Als we een getal veranderen, kan het gemiddelde dramatisch veranderen: het gemiddelde van [3 4 2 200 5] is $\frac{214}{5}$. De mediaan wordt hier echter niet sterk door beïnvloed: hij is nu 4.

Spreidingsmaten: quartielen

Het gemiddelde en de mediaan geven beiden een idee van de locatie van de data. We willen ook graag weten hoe de data uitgespreid zijn rond deze waarden. Bijvoorbeeld, de datasets

$$x = [1 \quad 2 \quad 2 \quad 7 \quad 8 \quad 8 \quad 10]$$
$$y = [5 \quad 5 \quad 6 \quad 7 \quad 12 \quad 12 \quad 13]$$

hebben beiden mediaan 7, maar hun spreiding is verschillend. We kunnen *quartielen* gebruiken: we delen de (geordende) data op in de observaties links van de mediaan en die rechts ervan, dan nemen we de mediaan van elk van de twee kleinere datasets.

Voor x krijgen we [1 2 2] en [8 8 10], met medianen 2 and 8. We zeggen dat het *eerste quartiel* van x gelijk is aan 2 en het *derde quartiel* aan 8. Voor y krijgen we 5 en 12.

Samenvattingen samengevat

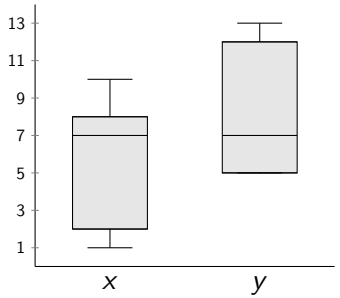
We hebben twee datasets

$$x = [1 \quad 2 \quad 2 \quad 7 \quad 8 \quad 8 \quad 10]$$
$$y = [5 \quad 5 \quad 6 \quad 7 \quad 12 \quad 12 \quad 13]$$

met

	minimum	eerste quartiel	mediaan	derde quartiel	maximum
x	1	2	7	8	10
y	5	5	7	12	13

We kunnen deze informatie samenvatten in een *boxplot*.



Spreiding om het gemiddelde meten

Beschouw

$$x = [4 \quad 8 \quad 9 \quad 9 \quad 10]$$
$$y = [1 \quad 8 \quad 9 \quad 9 \quad 18]$$

We hebben $\bar{x} = 8$ en $\bar{y} = 9$. Het is duidelijk dat y meer uitgespreid is dan x. Hoe meten we dit?
Beschouw de afstand van elke waarneming tot het gemiddelde:

$$(10-8)^2 = 2^2 = 4$$
$$(9-8)^2 = 1^2 = 1$$
$$(9-8)^2 = 1^2 = 1$$
$$(8-8)^2 = 0^2 = 0$$
$$(4-8)^2 = (-4)^2 = 16$$

Neem de gemiddelde afstand:

$$\frac{4+1+1+0+16}{5} = \frac{22}{5}.$$

Voor y:

$$\frac{81+0+0+1+64}{5} = \frac{146}{5}.$$

Variantie en standaarddeviatie

Voor een dataset $x_1, x_2, \ldots, x_n$ is de *variantie* gelijk aan

$$\text{var} = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1}.$$

Merk op dat we delen door $n-1$ en niet door n! De *standaarddeviatie* wordt dan gegeven door

$$\sigma = \sqrt{\text{var}} = \sqrt{\frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2}{n-1}}.$$

Beschouw $x = [1 \quad 1 \quad 3 \quad 4 \quad 8 \quad 13]$. We hebben $\bar{x} = 5$ en

$$\sigma = \sqrt{\frac{4^2+4^2+2^2+1^2+3^2+8^2}{5}} = \sqrt{\frac{110}{5}} = \sqrt{22}.$$