Kansrekenen en Statistiek

Sandra Van Aert

4 oktober 2011

Hoor- en werkcolleges

- docenten
  - Sandra Van Aert
    - theorie + practica FYS
  - Stefan Van Dongen
    - practica BIR
  - Annick De Backer
    - practica BIR en FYS

Hoor- en werkcolleges

- theorie
  - BIR+FYS: 1x dinsdag 10.45u–12.45u in U.024
  - BIR+FYS: 11x donderdag 10.45u–12.45u in U.024
  - FYS: 2x vrijdag 10.45u–12.45 in US.103

- practica
  - BIR: 11x dinsdag 13.45u–15.45u in US.103
  - BIR: 2x maandag 10.45u–12.45u in US.103
  - FYS: 11x woensdag 10.45u–12.45u in US.103
  - FYS: 3x vrijdag 10.45u–12.45 in US.103
  - opgaven oefeningen:
    - elke week op Blackboard
    - mee te brengen naar het practicum

Cursusmateriaal

- cursustekst
  - auteurs: Sandra Van Aert, Stefan Van Dongen en Peter Goos
  - verkrijgbaar bij de Cursusdienst

- formularium
  - achteraan cursus
  - beschikbaar op Blackboard
  - mag gebruikt worden tijdens het examen

Software

- BIR: statistisch pakket R
  - voor iedereen beschikbaar
  - bevat heel wat statistische functies
  - zal gebruikt worden tijdens practicum en in de vervolgcursus
- FYS: Matlab
  - vaak als programmeertaal gebruikt door fysici
  - sluit aan bij het vak Computerpracticum BA1
  - zal bebruikt worden tijdens het practicum
- JAVA applets
  - illustratie van concepten tijdens hoorcolleges

Evaluatie

- schriftelijk **examen** in januari
  - telt mee voor 90%
  - gesloten boek, maar MET formularium
  - theorievragen en oefeningen krijgen gelijk gewicht (45% theorie, 45% oefeningen)
  - leerstof: cursusnota's, lessen, oefeningensessies

- **permanente evaluatie**
  - telt mee voor 10%
  - BIR + FYS: meerkeuzevragen test lesweek 8
  - BIR: verplichte oefeningen lesweek 5 en 9 (in te leveren lesweek 7 en 11)
  - FYS: verplichte oefeningen lesweek 5, 9 en 11 (in te leveren lesweek 7, 11 en 13)

Vragen

- vragen stellen mag tijdens en na de les
- voor serieuze problemen
  - afspraak maken met één van de docenten per e-mail
  - sandra.vanaert@ua.ac.be
    stefan.vandongen@ua.ac.be
    annick.debacker@ua.ac.be

Waarom statistiek?

Iedereen wordt vroeg of laat met de analyse van gegevens geconfronteerd:

- wetenschappelijk onderzoek
- industrie (waaraan zijn defecte producten te wijten?)
- verzekeringsmaatschappijen (aan wie kennen ze een verzekering toe? hoe hoog moet de premie zijn?)
- overheid (hoe goed werken beleidsmaatregelen?)
- bachelorproef of masterproef

Wat is statistiek?

- een statistiek verwijst altijd naar numerieke informatie
- geheel van methodologieën voor het verzamelen, voorstellen, analyseren en interpreteren van data of gegevens
- belangrijke hulpwetenschap
- geneeskunde, economie, chemie, biologie, fysica, politieke wetenschappen, criminologie, ...

Voorbeelden

- **luchtvaartmaatschappijen**: no-shows, bagagegewicht, . . . (vb. overboekingen)
- **grootwarenhuizen**: gekochte producten, gespendeerde bedragen, betalingswijzen, . . . (vb. op maat gemaakte reclamefolders)
- **geneeskunde**: lichaamstemperatuur, hartslag, . . . (vb. verband tussen lichaamstemperatuur en hartslag?)
- **scheikunde**: concentratie chemische stoffen en smaakscore van cheddar kazen, . . . (vb. verschillen tussen goede en matige kazen wat melkzuur en azijnzuur in de kaas betreft?)

Studie-object van de statistiek

- populatie van objecten: Belgische bevolking, klanten van een grootwarenhuis, patiënten, ...
- processen die objecten genereren: industriële en chemische productieprocessen
- gegevens: geregistreerde eigenschappen of karakteristieken → variabelen
- steekproef: slechts een deel van de objecten wordt bestudeerd

Takken van de statistiek

- beschrijvende statistiek
  - beschrijven van steekproefgegevens
  - overzichtelijk voorstellen
  - berekenen van een aantal kenmerkende waarden (gemiddelde, variantie, ...)
- verklarende of inferentiële statistiek
  - analyseren en interpreteren van steekproefgegevens
  - antwoorden vinden op vragen of hypothesen
  - nagaan wat de waarde is van een model
  - veralgemenen naar de ganse populatie of het ganse proces: inferentie

Data en hun voorstelling

Sandra Van Aert

4 oktober 2011

Meetschalen

- gegevens worden verzameld over meerdere eigenschappen of variabelen
- voorbeelden
  - kleur van wijn
  - hematocrietgehalte van wielrenner
- kwalitatieve of categorische variabelen:
  - nominale meetschaal
  - ordinale meetschaal
- kwantitatieve variabelen:
  - intervalschaal
  - ratio meetschaal

Nominale variabelen

- elementen van steekproef/populatie worden in een klasse of categorie geplaatst
- voorbeelden
  - geslacht (man/vrouw)
  - nationaliteit (Belg/Nederlander/...)
  - godsdienst (katholiek/protestants/...)
  - type autoverzekering (omnium/burgerlijke aansprakelijkheid/geen)
  - gemeente
- cijfercodes
  - man = 0, vrouw = 1
  - postnummers van gemeenten
  - cijfercodes impliceren geen volgorde: rekenkundige bewerkingen zinloos (behalve percentages)

Ordinale variabelen

- nominale variabelen waarbij er een ordening is tussen de klassen of categorieën
- voorbeelden
  - aantal Michelinsterren van een restaurant
  - antwoorden op enquêtes: “1: helemaal eens”, “2: eerder eens”, “3: noch eens, noch oneens”, “4: eerder oneens” of “5: helemaal oneens”
- rekenkundige bewerkingen zinloos (behalve percentages)

Kwantitatieve variabelen

- worden uitgedrukt in een aantal vaste meeteenheden
- voorbeelden
  - lengte
  - gewicht
  - aantal verkochte auto's
  - temperatuur
  - duurtijd
  - aantal Kb per tijdseenheid
  - ...
- bijna alle rekenkundige bewerkingen zinvol

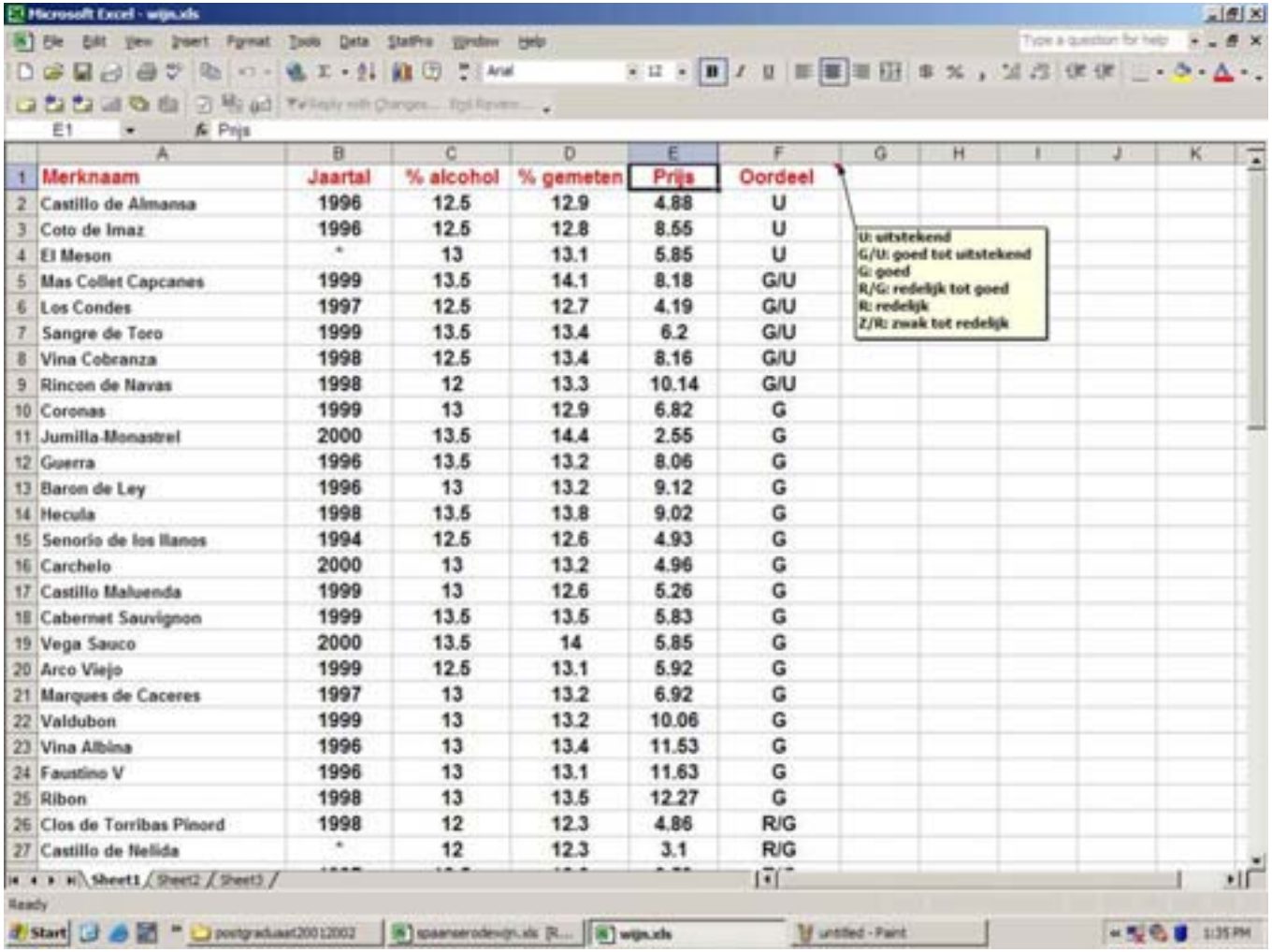
Kwantitatieve variabelen

- intervalschaal:
  - geen natuurlijk nulpunt
  - voorbeeld: temperatuur (Celsius, Fahrenheit), tijd afgelezen op een klok
  - verschil tussen 2 en 4 uur = verschil tussen 21 en 23 uur
  - verhoudingen houden geen steek
  - 4 uur is niet dubbel zo laat als 2 uur
- ratioschaal:
  - wel absoluut nulpunt
  - voorbeeld: lengte, gewicht, ...
  - verhoudingen zijn wel zinvol
  - 2 meter is dubbel zo lang als 1m

Meetschalen

- kwantitatieve variabelen
  - discreet: vb. aantal passagiers op lijnvlucht
  - continu: vb. lengte, duurtijd, ...
- hiërarchie
  - variabelen gemeten op ratioschaal zijn meest informatief
  - gegevens gemeten op een hogere schaal kunnen omgezet worden in gegevens op een lagere schaal, maar niet omgekeerd!
  - statistische methoden voor lagere meetschalen kunnen gebruikt worden voor hogere meetschalen, maar niet omgekeerd!

Datamatrix of gegevensmatrix



	Merknaam	Jaartal	% alcohol	% gemeten	Prijs	Oordeel
2	Castillo de Almansa	1996	12.5	12.9	4.88	U
3	Coto de Imaz	1996	12.5	12.8	8.55	U
4	El Meson	*	13	13.1	5.85	U
5	Mas Collet Capcanes	1999	13.5	14.1	8.18	G/U
6	Los Condes	1997	12.5	12.7	4.19	G/U
7	Sangre de Toro	1999	13.5	13.4	6.2	G/U
8	Vina Cobranza	1998	12.5	13.4	8.16	G/U
9	Rincon de Navas	1998	12	13.3	10.14	G/U
10	Coronas	1999	13	12.9	6.82	G
11	Jumilla-Monastrel	2000	13.5	14.4	2.55	G
12	Guerra	1996	13.5	13.2	8.06	G
13	Baron de Ley	1996	13	13.2	9.12	G
14	Hecula	1998	13.5	13.8	9.02	G
15	Senorio de los llanos	1994	12.5	12.6	4.93	G
16	Carchelo	2000	13	13.2	4.96	G
17	Castillo Maluenda	1999	13	12.6	5.26	G
18	Cabernet Sauvignon	1999	13.5	13.5	5.83	G
19	Vega Sauco	2000	13.5	14	5.85	G
20	Arco Viejo	1999	12.5	13.1	5.92	G
21	Marques de Caceres	1997	13	13.2	6.92	G
22	Valdubon	1999	13	13.2	10.06	G
23	Vina Albina	1996	13	13.4	11.53	G
24	Faustino V	1996	13	13.1	11.63	G
25	Ribon	1998	13	13.5	12.27	G
26	Clos de Torribas Pinord	1998	12	12.3	4.86	R/G
27	Castillo de Nelida	*	12	12.3	3.1	R/G

U: uitstekend
G/U: goed tot uitstekend
G: goed
R/G: redelijk tot goed
R: redelijk
Z/R: zwak tot redelijk

Voorstellen van univariate kwalitatieve variabelen

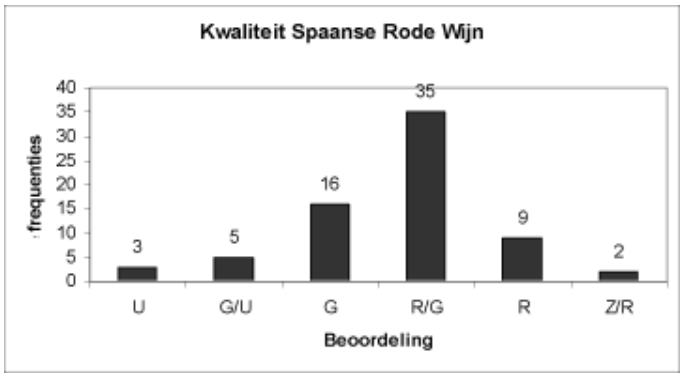
- frequenties en relatieve frequenties
- staafdiagram
- cirkel-, sector- of taartdiagram

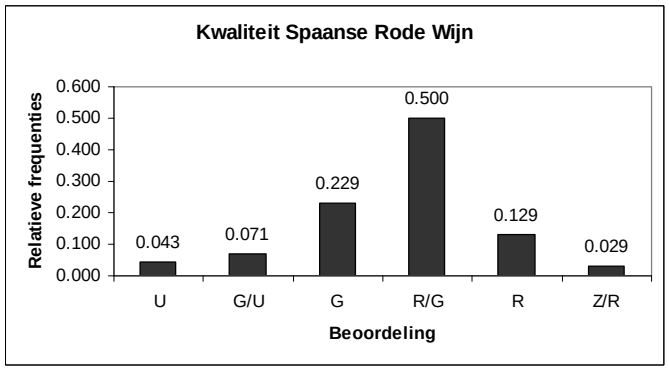
(Relatieve) frequenties en staafdiagram

- frequenties en relatieve frequenties

oordeel	U	G/U	G	R/G	R	Z/R	totaal
frequentie	3	5	16	35	9	2	70
rel. frequentie	.043	.071	.229	.500	.129	.029	1

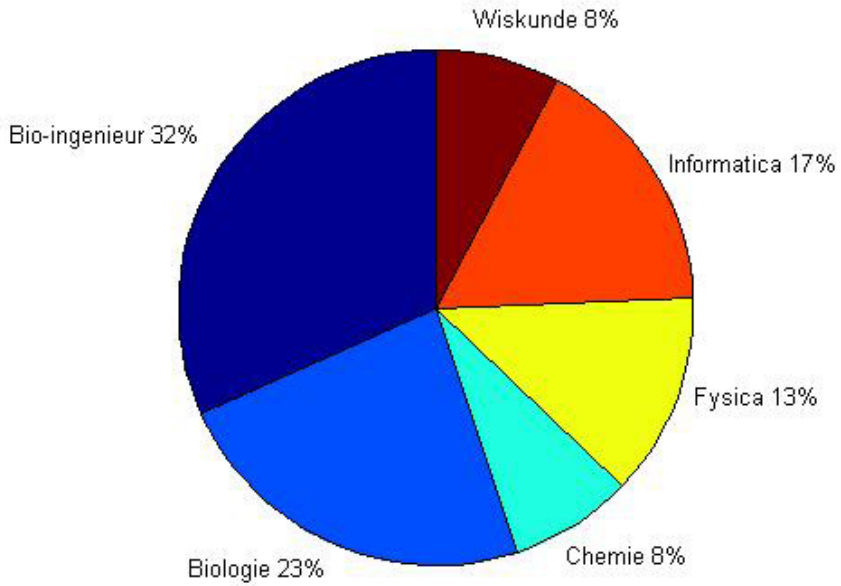
- staafdiagram





Cirkel-, sector- of taartdiagram

Aantal inschrijvingen bacheloropleidingen UA op 27 september 2011



Voorstellen van univariate kwantitatieve variabelen

- stam- en bladdiagram
- staafdiagram
- histogram
- (frequentie)polygoon
- empirische cumulatieve verdelingsfunctie

Stam- en bladdiagram

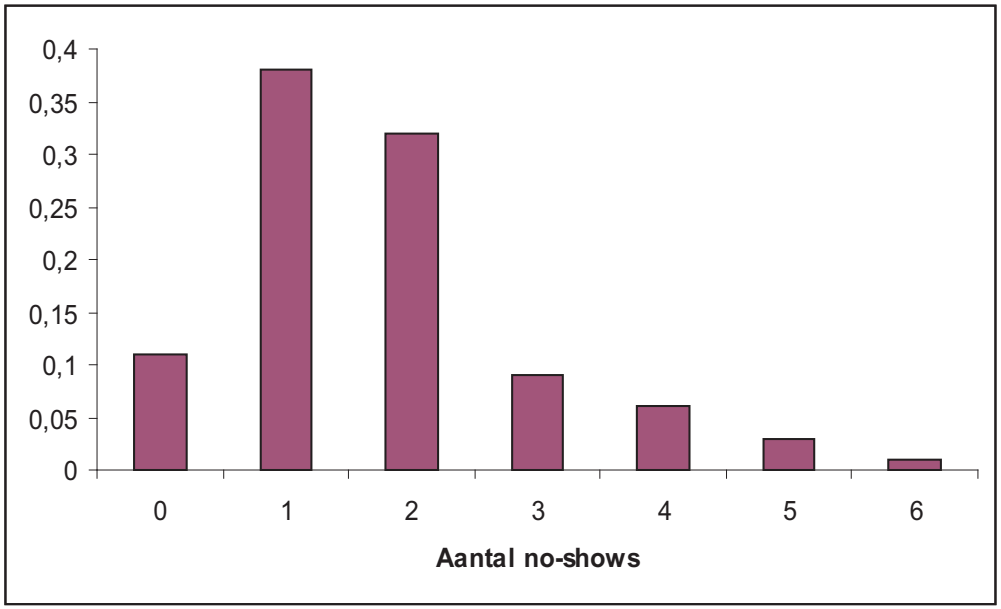
Stam- en bladdiagram van de variabele prijs

Frequentie	Stam &	Blad
4,00	2 .	2567
4,00	3 .	0126
11,00	4 .	02236788999
12,00	5 .	023345789999
4,00	6 .	2389
4,00	7 .	2229
7,00	8 .	1224667
2,00	9 .	01
4,00	10 .	0015
2,00	11 .	56
3,00	12 .	134
2,00	13 .	56

Staafdiagram voor discrete variabelen

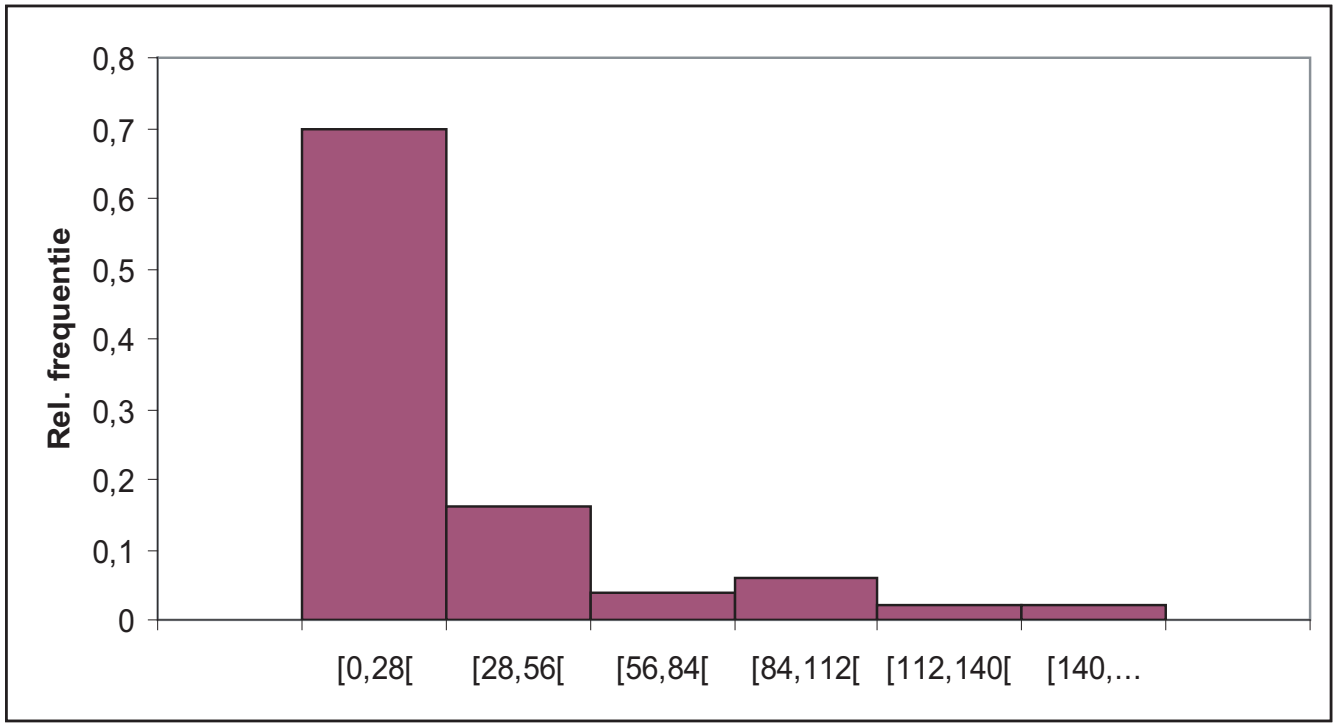
Aantal no-shows	0	1	2	3	4	5	6
Frequentie	11	38	32	9	6	3	1
Rel. frequentie	11%	38%	32%	9%	6%	3%	1%

Staafdiagram aantal no-shows op 100 vluchten



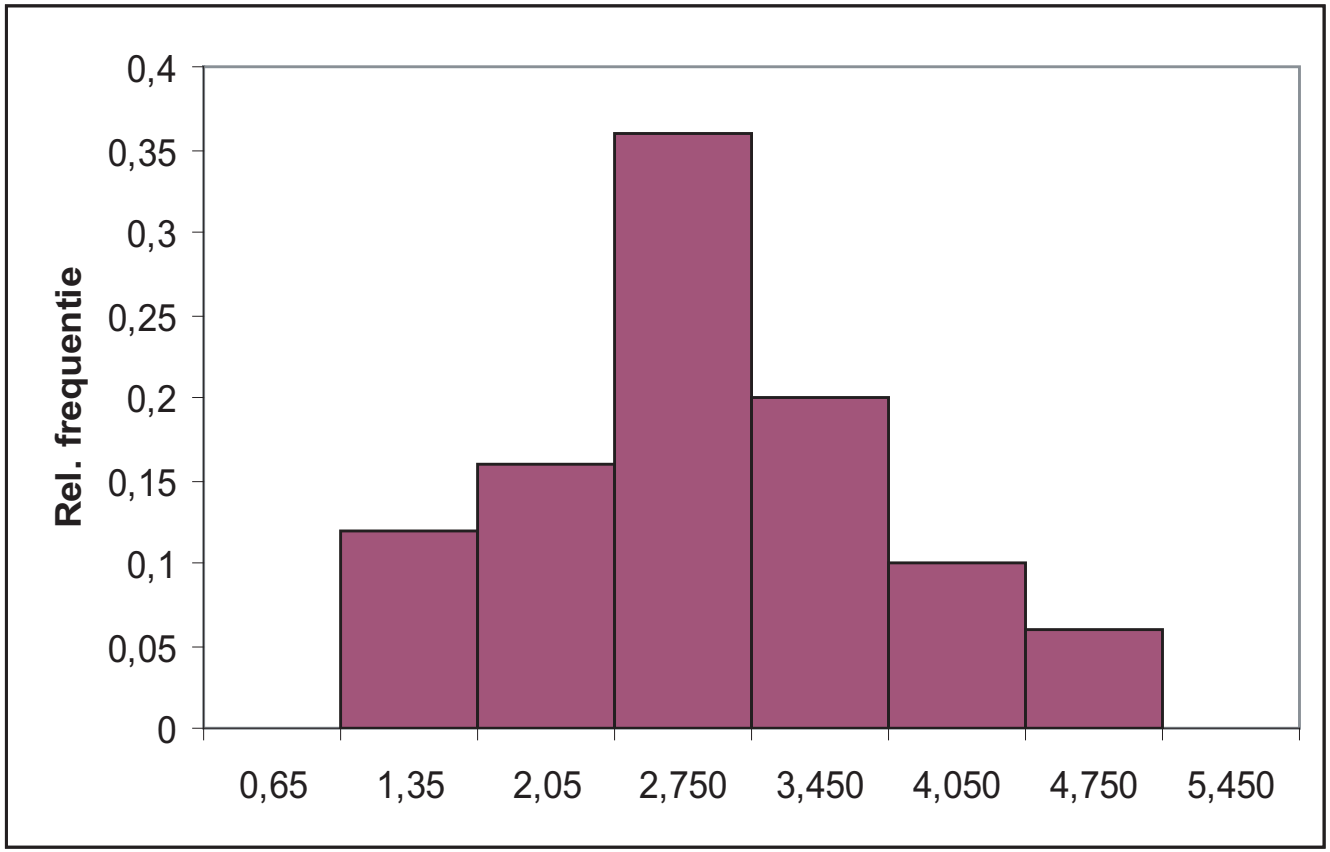
Histogram voor continue variabelen

Histogram voor 50 breeksterktes (uitgedrukt in kg)



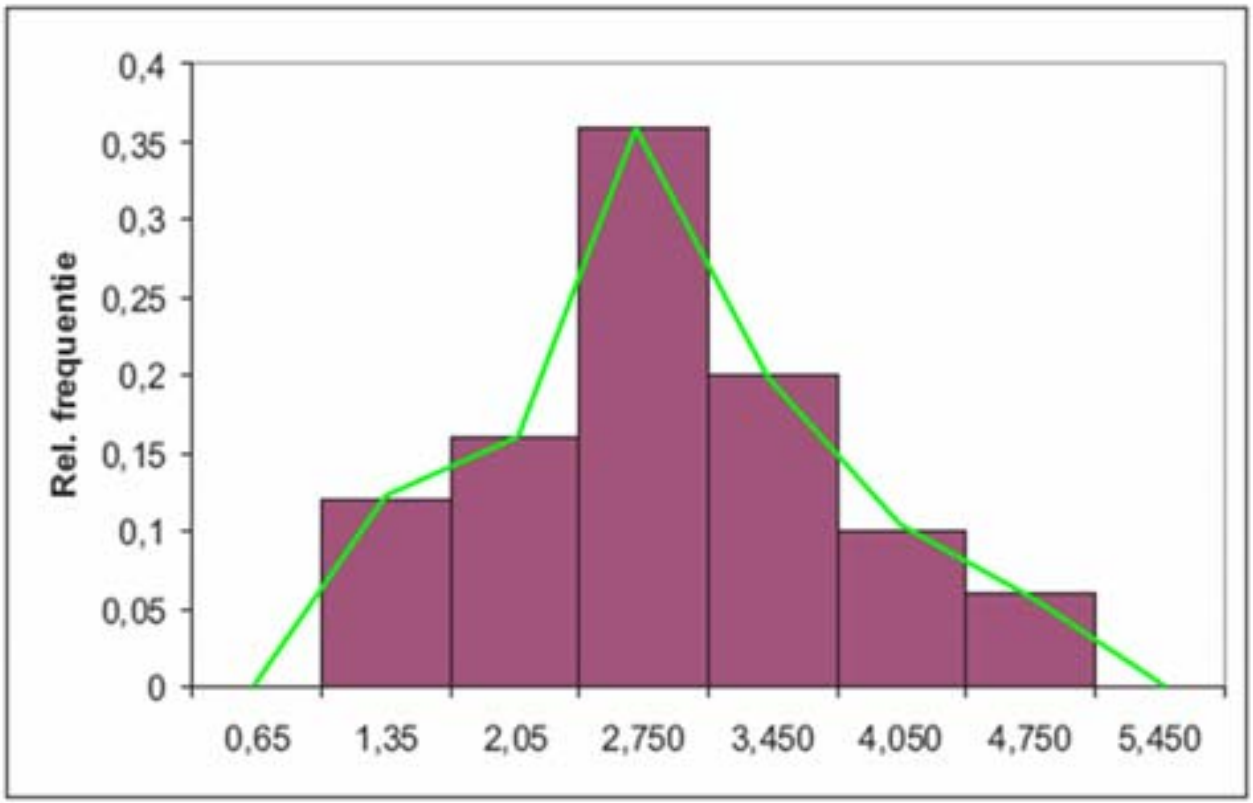
Histogram voor continue variabelen

Histogram voor natuurlijk logaritme van 50 breeksterktes



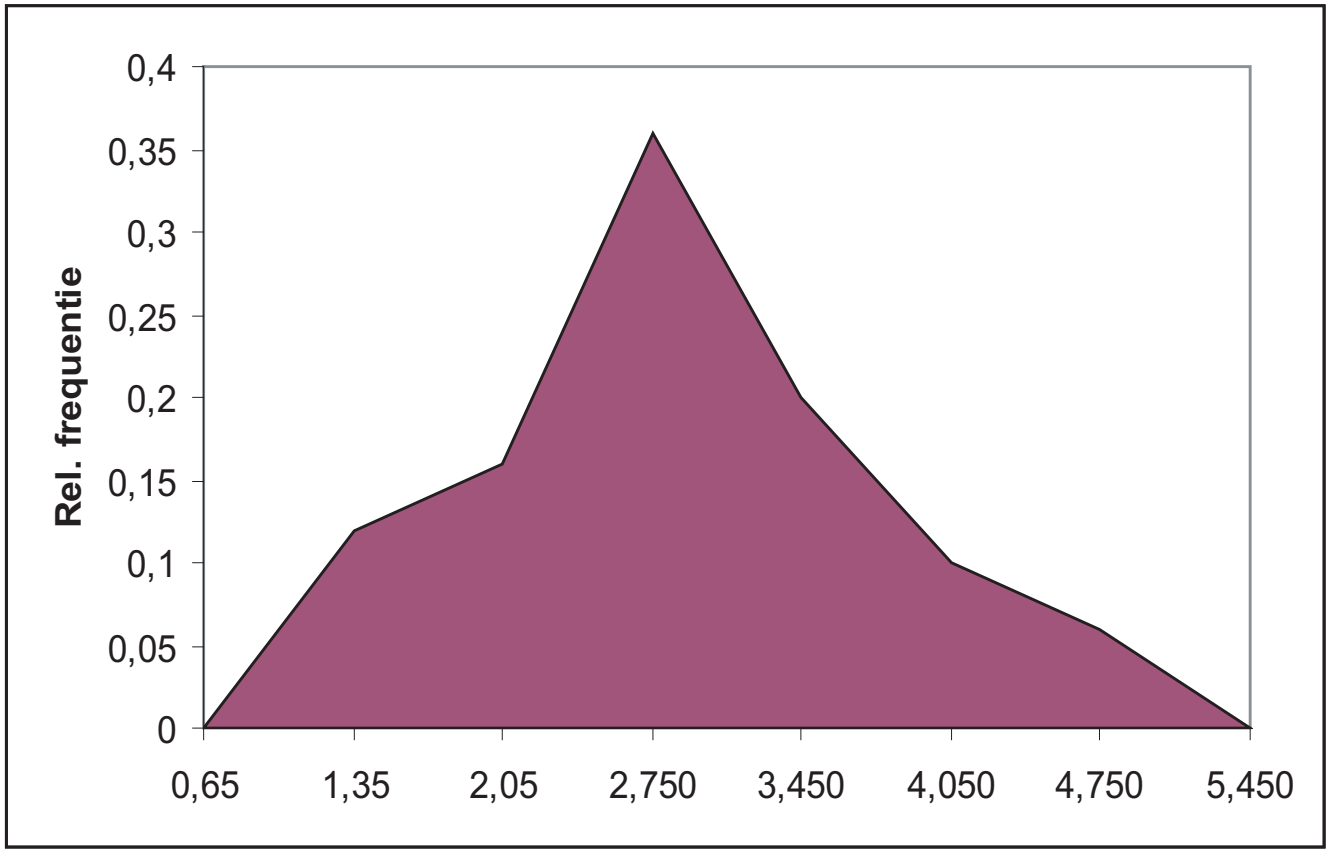
Log(breeksterkte)

Histogram voor continue variabelen



Log(breeksterkte)

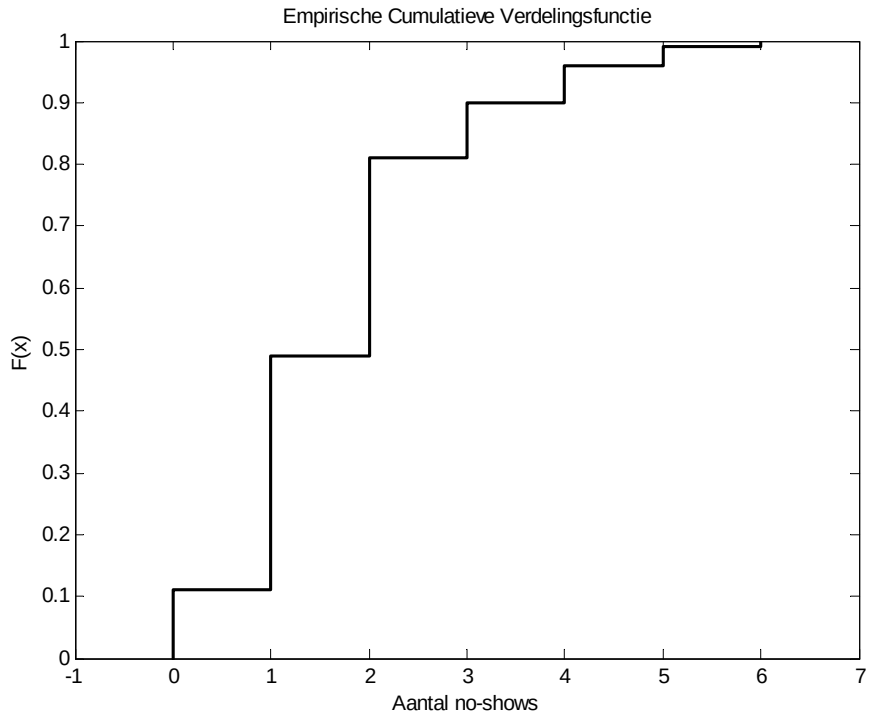
Frequentiepolygoon voor continue variabelen



Log(breeksterkte)

Empirische cumulatieve verdelingsfunctie

Aantal no-shows	0	1	2	3	4	5	6
Frequentie	11	38	32	9	6	3	1
Rel. frequentie	11%	38%	32%	9%	6%	3%	1%
Cum. rel. frequentie	11%	49%	81%	90%	96%	99%	100%



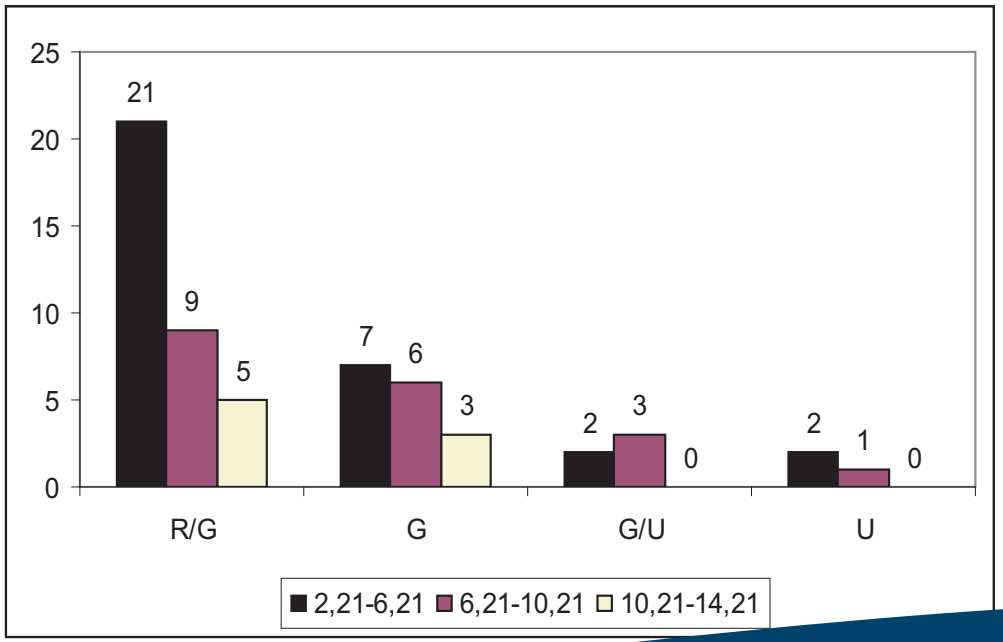
Voorstellen van bivariate variabelen

- kruistabel
- meervoudig staafdiagram
- puntenwolk

Kruistabel voor bivariate variabelen

	R/G	G	G/U	U	Totaal
[2.21 – 6.21[	21	7	2	2	32
[6.21 – 10.21[	9	6	3	1	19
[10.21 – 14.21[	5	3	0	0	8
Totaal	35	16	5	3	59

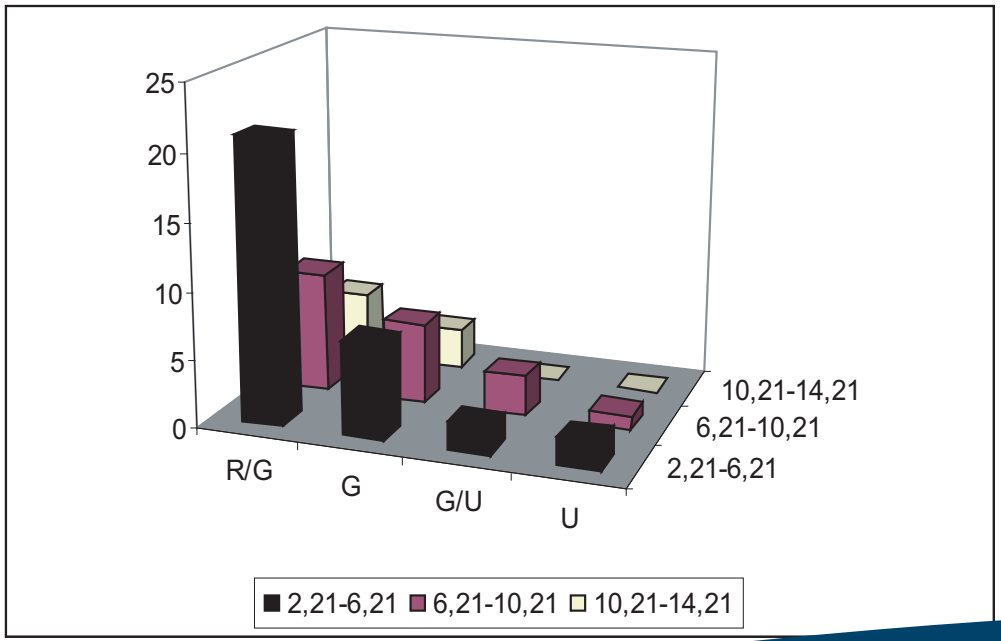
Meervoudig staafdiagram



Kruistabel voor bivariate variabelen

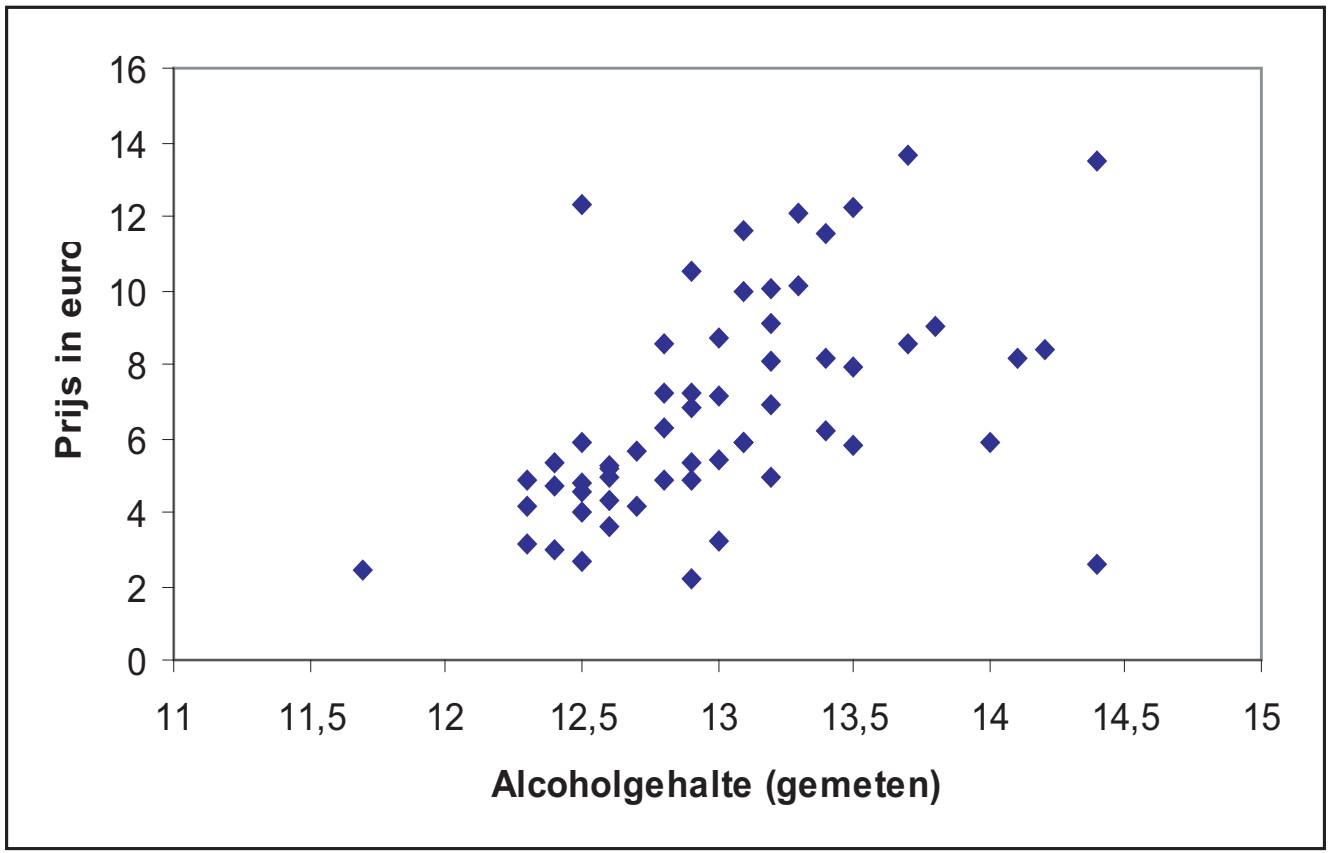
	R/G	G	G/U	U	Totaal
[2.21 – 6.21[	21	7	2	2	32
[6.21 – 10.21[	9	6	3	1	19
[10.21 – 14.21[	5	3	0	0	8
Totaal	35	16	5	3	59

Driedimensionaal staafdiagram



Puntenwolk voor bivariate kwantitatieve variabelen

Elke waarneming wordt door een punt voorgesteld



Beschrijvende statistieken van steekproefgegevens

Sandra Van Aert

4 oktober 2010

Wat?

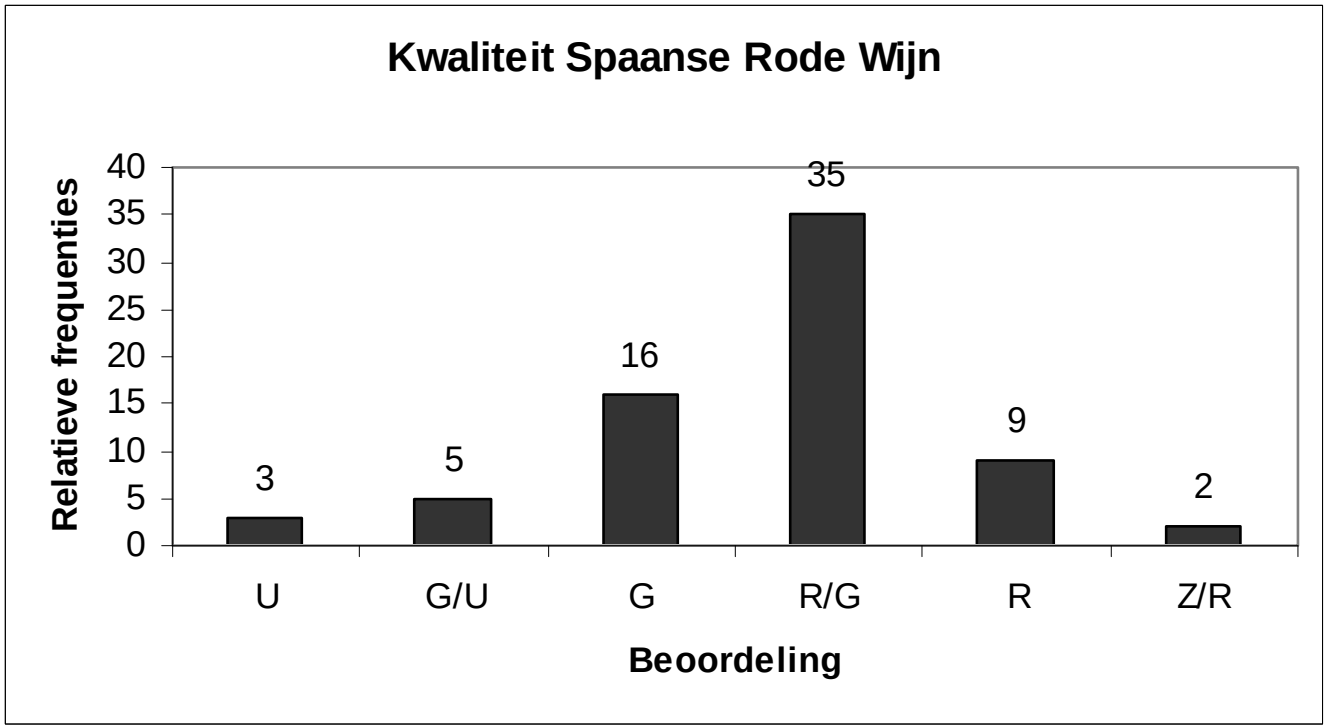
- kengetallen of statistieken → samenvatting van steekproefgegevens
  - ligging, locatie
  - spreiding
  - scheefheid
- aangeduid m.b.v. Romeinse letters
- niet alle kengetallen kunnen voor alle meetschalen gebruikt worden

Kengetallen of statistieken

	Nominaal	**Ordinaal**	**Interval/ratio**
Ligging	Modus	Modus Mediaan Kwartielen	Modus Mediaan Kwartielen Rekenkundig gemiddelde
Spreiding		Spreidingsbreedte Interkwartielbreedte	Spreidingsbreedte Interkwartielbreedte Gemiddelde absolute afwijking Variantie en standaardafwijking Variatiecoëfficiënt
Scheefheid			Pearson Fisher
Ligging, spreiding en scheefheid		Box-plot	Box-plot
Correlatie			Correlatiecoëfficiënt

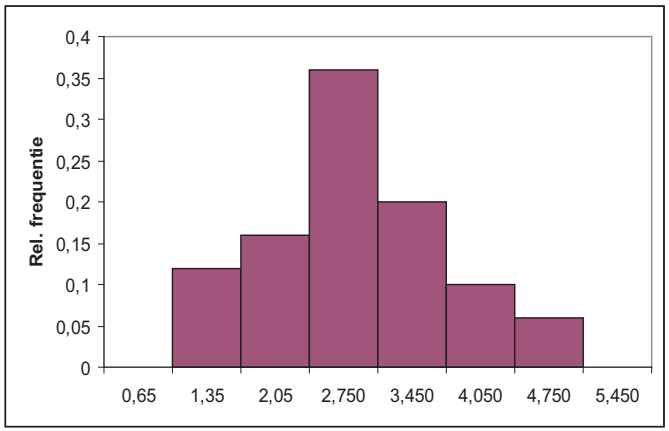
Modus

- kan voor elk type van gegevens gebruikt worden
- M_0: waarneming met de grootste frequentie



Modale klasse

- bij continue, kwantitatieve variabelen heeft modus weinig zin omdat elke waarneming slechts één keer voorkomt
- men stelt dan histogrammen op en bepaalt de klasse met de grootste frequentie
- terminologie: unimodaal, bimodaal, multimodaal



Log(breeksterkte)

Mediaan

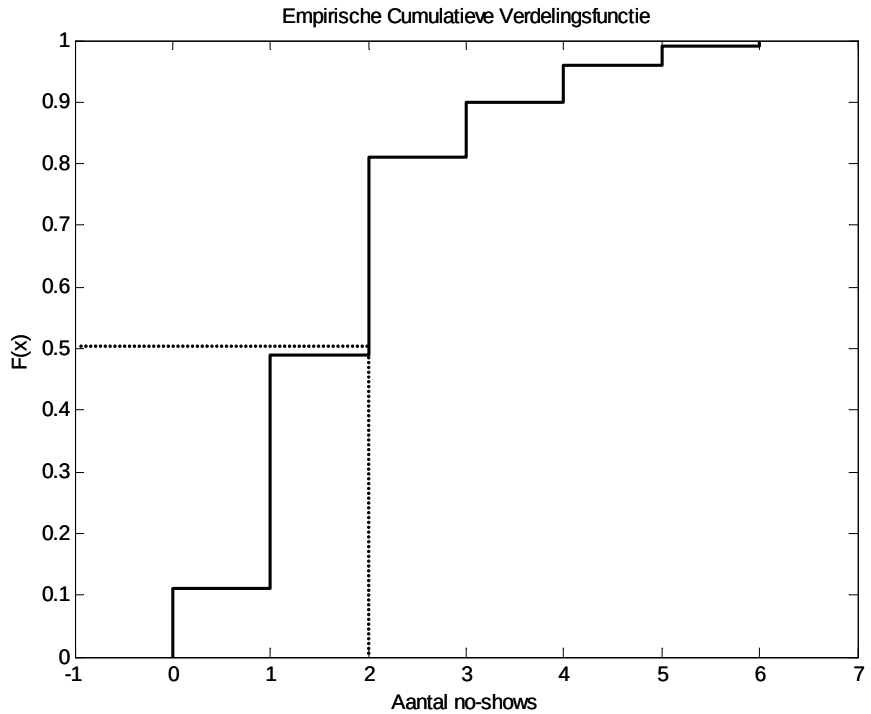
- ordinale gegevens en kwantitatieve gegevens
- M_e: middelste element van geordende data
  - aantal elementen n oneven: $((n+1)/2)^{\text{de}}$ element
  - aantal elementen n even: gemiddelde van het $(n/2)^{\text{de}}$ en het $(n/2+1)^{\text{de}}$ element
- voorbeeld: 16, 13, 14, 17, 14, 16, 17, 16, 15, 13
  - $n = 10 \rightarrow n/2 = 5$ en $n/2+1 = 6$
  - geordend: 13, 13, 14, 14, 15, 16, 16, 16, 17, 17
  - $M_e = (15+16)/2 = 15.5$

Mediaan

- ongeveer 50% van de waarnemingen ligt onder/boven de mediaan
- de mediaan wordt niet beïnvloed door een klein aantal extreme waarnemingen
- bepaling van de mediaan uit de empirische cumulatieve verdelingsfunctie

Mediaan

Aantal no-shows	0	1	2	3	4	5	6
Frequentie	11	38	32	9	6	3	1
Rel. frequentie	11%	38%	32%	9%	6%	3%	1%
Cum. rel. frequentie	11%	49%	81%	90%	96%	99%	100%



Rekenkundig gemiddelde

- het rekenkundig gemiddelde $\overline{x}$ van de waarnemingen $x_1, \ldots, x_n$ is

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

- voorbeeld: 16, 13, 14, 17, 14, 16, 17, 16, 15, 13

$$\begin{aligned}\overline{x} &= \frac{1}{10}(16+13+14+17+14+16+17+16+15+13)\\ &= 15.1\end{aligned}$$

Rekenkundig gemiddelde

- rekenkundig gemiddelde bij gegroepeerde gegevens:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i = \frac{1}{n}(f_1 x_1 + f_2 x_2 + \cdots + f_n x_n)$$

- voorbeeld:

Aantal no-shows	0	1	2	3	4	5	6
Frequentie	11	38	32	9	6	3	1
Rel. frequentie	11%	38%	32%	9%	6%	3%	1 %

$$\overline{x} = \frac{1}{100}(11 \times 0 + 38 \times 1 + 32 \times 2 + 9 \times 3 + 6 \times 4 + 3 \times 5 + 1 \times 6)$$
$$= 1.74$$

Gemiddelde van nominale data

3010	Kessel-Lo	Vlaams-Brabant
1010	Brussel	Brussel
9000	Gent	Oost-Vlaanderen
9270	Laarne	Oost-Vlaanderen
8940	Wervik	West-Vlaanderen
8900	Ieper	West-Vlaanderen
9160	Lokeren	Oost-Vlaanderen
3300	Kumtich	Vlaams-Brabant
9770	Kruishoutem	Oost-Vlaanderen
3670	Meeuwen	Limburg
8000	Brugge	West-Vlaanderen
6370	Bellefontaine	Luxemburg

Voor- en nadelen rekenkundig gemiddelde

- gebruikt alle waarnemingen
- gevoelig voor extreme waarden (in tegenstelling tot mediaan)
- voorbeeld:

$$\overline{x} = \frac{1}{10}(16+13+14+17+14+16+17+16+15+13) = 15.1$$

$$\overline{x} = \frac{1}{10}(16+13+14+17+14+16+17+16+15+130) = 26.8$$

Maatstaven voor relatieve ligging

- ordestatistiek
  - minimum
  - maximum
- percentiel of kwantiel
- deciel
- kwartiel

Ordestatistiek of -kengetal

- i^{de} ordestatistiek of -kengetal $x_{(i)}$:
  - i^{de} waarneming nadat de gegevens gerangschikt zijn van klein naar groot
  - $x_{(i)}$ is het i-de kleinste getal
- voorbeeld: 16, 13, 14, 17, 14, 16, 17, 16, 15, 13
  - $x_{(1)} =?$, $x_{(4)} =?$, $x_{(10)} =?$
  - geordend: 13, 13, 14, 14, 15, 16, 16, 16, 17, 17
  - $x_{(1)} = 13$ (minimum)
  - $x_{(4)} = 14$
  - $x_{(10)} = 17$ (maximum)

Percentielen of kwantielen

- $(100 \times p)^{\text{de}}$ percentiel of kwantiel c_p:
  - reëel getal dat
    - groter is dan $(100 \times p)\%$ van de waarnemingen
    - kleiner is dan $(100 \times (1-p))\%$ van de waarnemingen
  - voorbeeld:
    - 80% van de gegevens is kleiner dan het 80ste percentiel of kwantiel $c_{0.80}$
    - 20% van de gegevens is groter dan het 80ste percentiel $c_{0.80}$
- verschillende berekeningswijzen (levert enkel verschillen in kleine datasets)

Percentielen of kwantielen

- $c_p = x_{(q)} + f(x_{(q+1)} - x_{(q)})$ met
  - $a = 1 + p \cdot (n-1)$
  - q = grootste geheel getal kleiner dan a
  - $f = a - q$

- voorbeeld: 16, 13, 14, 17, 14, 16, 17, 16, 15, 13
  - 80^{ste} percentiel = 8^{ste} deciel = ?
  - $a = 1 + 0.8 \times (10-1) = 8.2$
  - $q = 8$
  - $f = 0.2$
  - $c_{0.8} = x_{(8)} + 0.2 \times (x_{(9)} - x_{(8)})$
  - geordend: 13, 13, 14, 14, 15, 16, 16, 16, 17, 17
  - $c_{0.8} = 16 + 0.2 \times (17-16) = 16.2$

Kwartielen

- eerste kwartiel $Q_1 = 25^{\text{ste}}$ percentiel $c_{0.25}$
  - een kwart van de gegevens is kleiner dan of gelijk aan Q_1
  - driekwart van de gegevens is groter dan of gelijk aan Q_1
- tweede kwartiel $Q_2 = 50^{\text{ste}}$ percentiel $c_{0.50}$ = mediaan M_e
- derde kwartiel $Q_3 = 75^{\text{ste}}$ percentiel $c_{0.75}$

Maatstaven voor spreiding

- mediaan, gemiddelde, . . . zeggen niet alles
- voorbeeld:
  - dataset 1: 16, 13, 14, 17, 14, 16, 17, 16, 15, 13
  - dataset 2: 19, 10, 11, 20, 11, 19, 20, 19, 12, 10
  - $M_e = 15.5$ en $\overline{x} = 15.1$ voor beide datasets
- elementaire spreidingsmaten:
  - spreidingsbreedte (range)

$$R = x_{\max} - x_{\min}$$

  - interkwartielbreedte Q

$$Q = Q_3 - Q_1$$

  - kunnen gebruikt worden voor ordinale en kwantitatieve gegevens

Maatstaven voor spreiding

- enkel voor kwantitatieve gegevens
- gemiddelde absolute afwijking
  - mean absolute deviation (MAD)
  - gemiddelde van alle afwijkingen van het rekenkundig gemiddelde in absolute waarde

$$\text{MAD} = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

- steekproefvariantie

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

Maatstaven voor spreiding

- steekproefvariantie

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

- Bewijs dat

$$s^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2) = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right\}$$

Maatstaven voor spreiding

- steekproefvariantie bij gegroepeerde gegevens:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{k} f_i (x_i - \overline{x})^2$$

- voorbeeld:

Aantal no-shows	0	1	2	3	4	5	6
Frequentie	11	38	32	9	6	3	1
Rel. frequentie	11%	38%	32%	9%	6%	3%	1 %

- $s^2 = \frac{1}{99}(11 \times (0 - 1.74)^2 + 38 \times (1 - 1.74)^2 + \ldots$
$+1 \times (6 - 1.74)^2) = 1.53$

Maatstaven voor spreiding

- enkel voor kwantitatieve gegevens
- steekproefstandaarddeviatie

$$s = \sqrt{s^2}$$

- variatiecoëfficiënt
  - dataset 1: 15, 20, 20, 30, 35, 35, 40, 45
  - dataset 2: 1015, 1020, 1020, 1030, 1035, 1035, 1040, 1045
  - variantie is voor beide datasets 114.29
  - gemiddeldes zijn 30 en 1030
  - relatief gezien meer variabiliteit in dataset 1
  - $\text{VC} = \frac{s}{\overline{x}}$
  - dataset 1: 0.3563 dataset 2: 0.0104

Transformaties

- soms heb je een dataset $x_1, x_2, \ldots, x_n$ (vb. temperaturen in °F)
- maar je werkt liever met gegevens $ax_1 + b, ax_2 + b, \ldots, ax_n + b$ (vb. temperaturen in °C)
- stel: $y_1, y_2, \ldots, y_n$
- wat zijn het gemiddelde en de variantie van de nieuwe data?
  - $\overline{y} = a\overline{x} + b$
  - $s_y^2 = a^2 s_x^2$

Transformaties

- Bewijs

$$
\begin{aligned}
s_y^2 &= \frac{1}{n-1}\{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\} \\
&= \frac{1}{n-1}\{\sum_{i=1}^{n} (ax_i + b)^2 - n(\overline{ax+b})^2\} \\
&= \frac{1}{n-1}\{\sum_{i=1}^{n} (a^2x_i^2 + 2abx_i + b^2) - n(a\bar{x} + b)^2\} \\
&= \frac{1}{n-1}\{\sum_{i=1}^{n} (a^2x_i^2 + 2abx_i + b^2) - n(a^2\bar{x}^2 + 2ab\bar{x} + b^2)\}
\end{aligned}
$$

Transformaties

$$
\begin{aligned}
s_y^2 &= \frac{1}{n-1}\{a^2\sum_{i=1}^{n}x_i^2 + 2ab\sum_{i=1}^{n}x_i + nb^2 - na^2\bar{x}^2 - 2nab\bar{x} - nb^2)\} \\
&= \frac{1}{n-1}\{a^2\sum_{i=1}^{n}x_i^2 + 2ab\sum_{i=1}^{n}x_i - na^2\bar{x}^2 - 2nab\bar{x})\} \\
&= \frac{1}{n-1}\{a^2\sum_{i=1}^{n}x_i^2 + 2nab\bar{x} - na^2\bar{x}^2 - 2nab\bar{x})\} \\
&= \frac{1}{n-1}\{a^2\sum_{i=1}^{n}x_i^2 - na^2\bar{x}^2\} \\
&= a^2\frac{1}{n-1}\{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\} \\
&= a^2 s_x^2
\end{aligned}
$$

Populaties i.p.v. steekproeven

- populatiegemiddelde

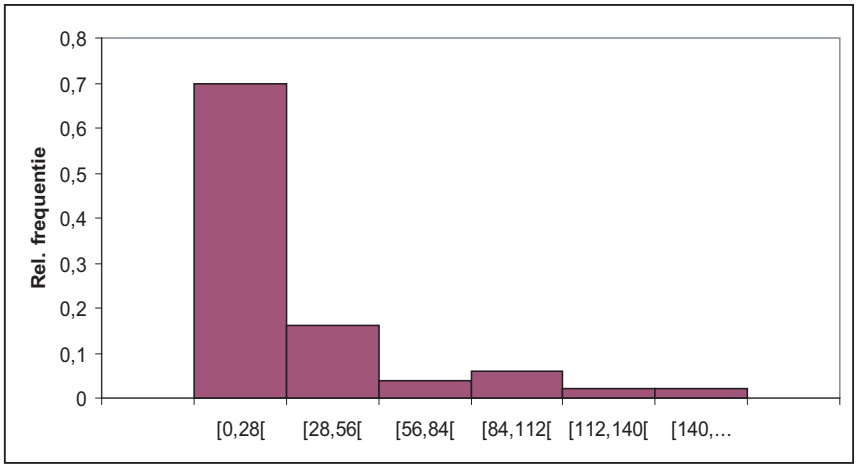
$$\cancel{\bar{x}}^{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- populatievariantie

$$\cancel{s^2}^{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \cancel{\bar{x}}^{\mu})^2}{\cancel{n-1}^{n}}$$

- populatiestandaarddeviatie: σ i.p.v. s

Scheefheid



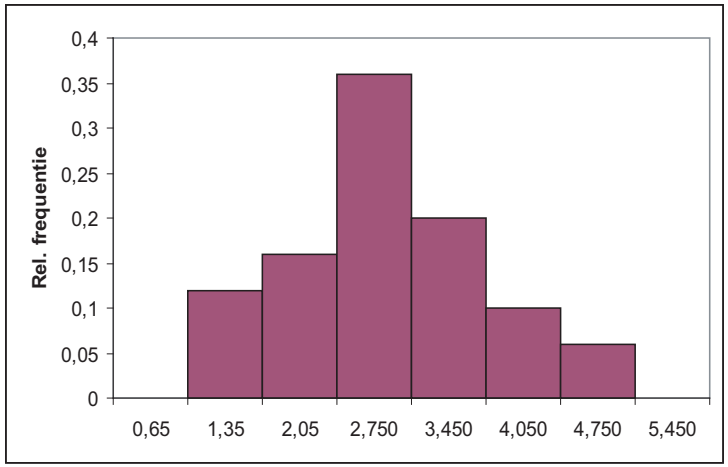
Breeksterkte

rechtsscheef

$M_o = 14.00$

$M_e = 14.80$

$\overline{x} = 28.59$



Log(breeksterkte)

symmetrisch

$M_o = 2.75$

$M_e = 2.58$

$\overline{x} = 2.86$

Scheefheid

- Pearsons scheefheidscoëfficiënt

$$S_P = \frac{3(\overline{x} - M_e)}{s}$$

- $-3 \leq S_P \leq +3$
- symmetrische verdeling : $S_P = 0$
- rechtsscheve verdeling : $S_P > 0$
- linksscheve verdeling : $S_P < 0$

Scheefheid

- Scheefheid van Fisher

$$\frac{m_3}{s^3}$$

- m_3 het derde centrale steekproefmoment
- m_k het k-de centrale steekproefmoment

$$m_k = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^k}{n}$$

- symmetrische verdeling : $m_3 = 0$
- rechtsscheve verdeling : $m_3 > 0$
- linksscheve verdeling : $m_3 < 0$

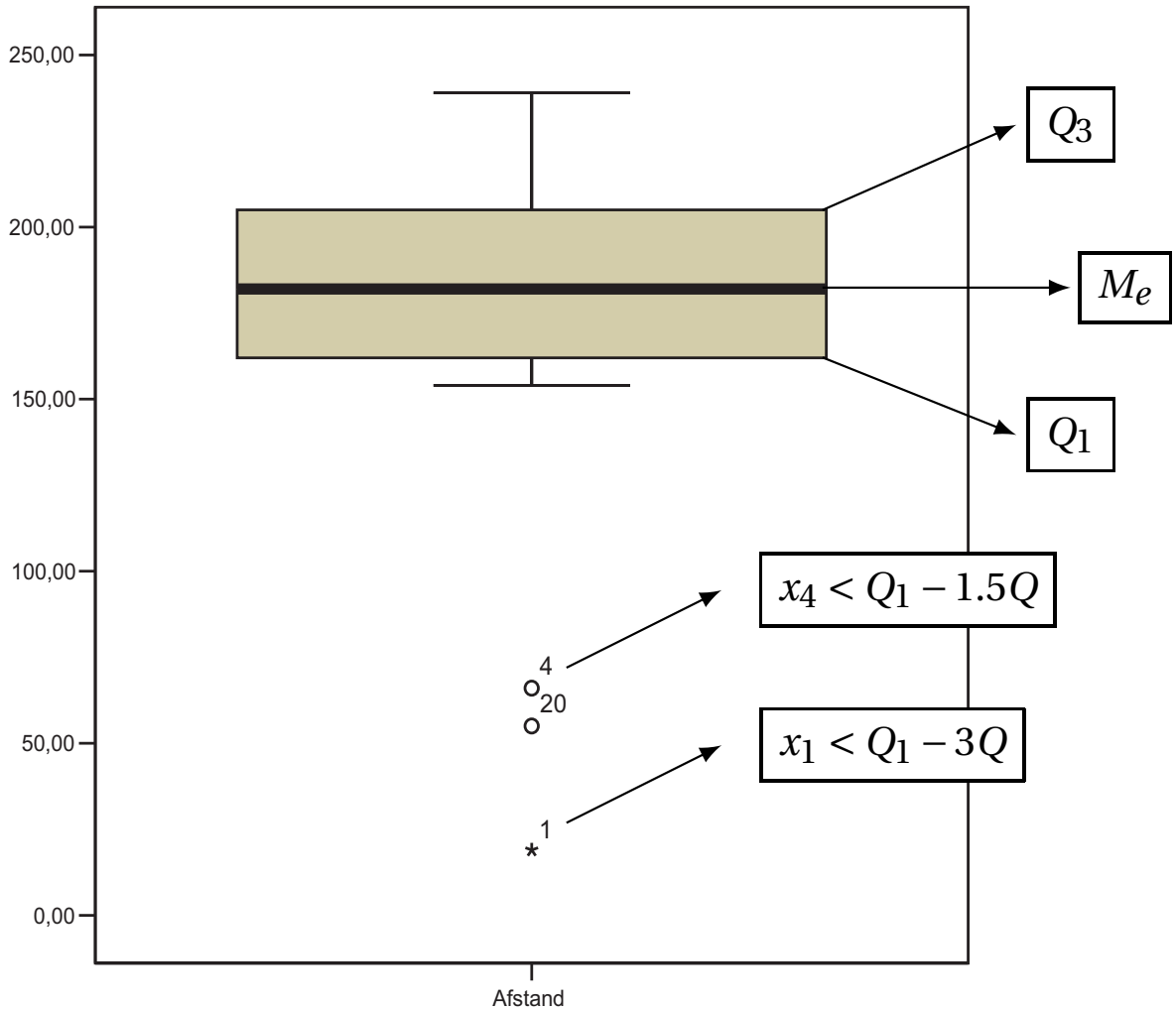
Standaardisatie

- gestandaardiseerde variabele: z

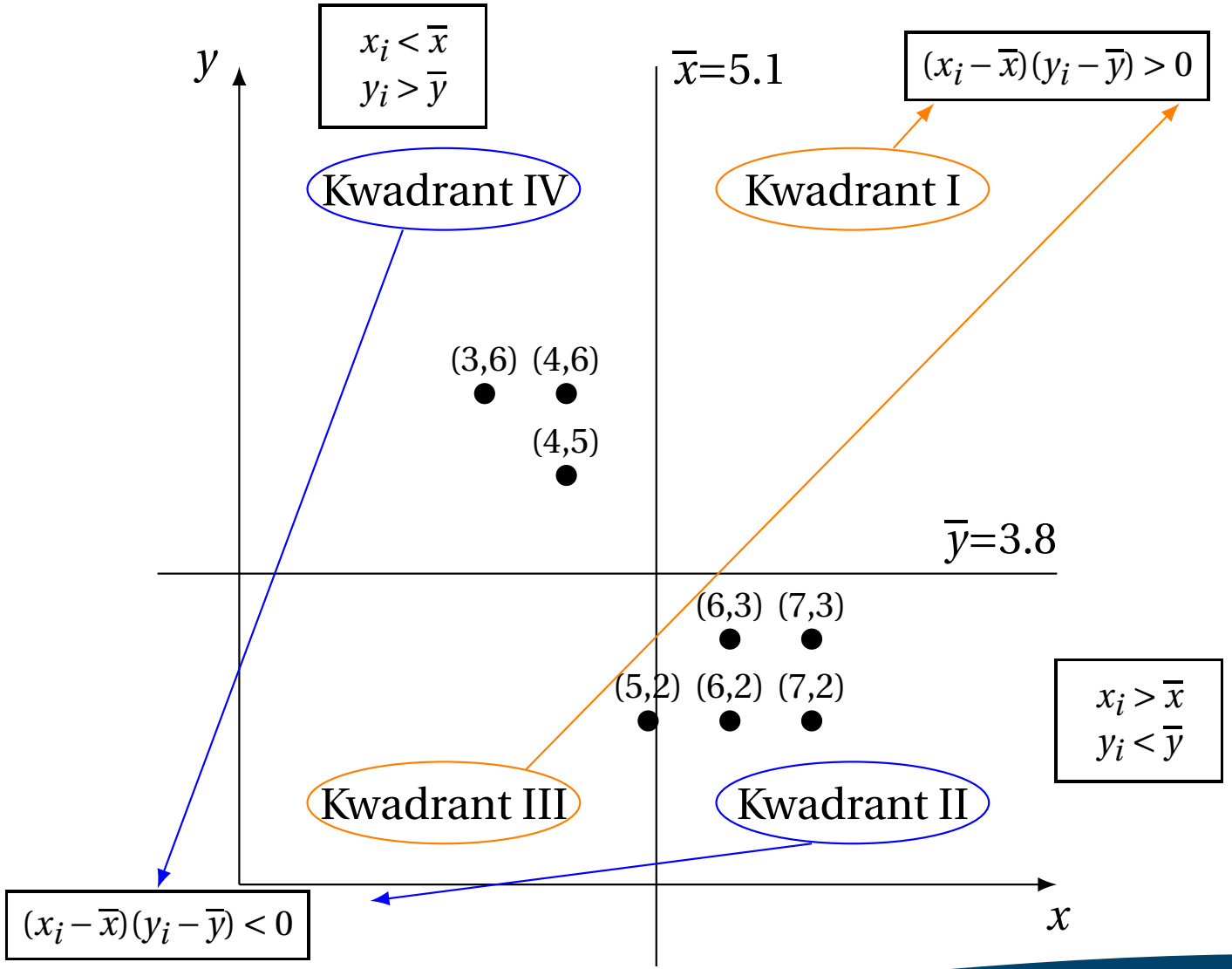
$$z_i = \frac{x_i - \overline{x}}{s}$$

- $\overline{z} =?$ en $s_z^2 =?$
- transformatie met $a = 1/s$ en $b = -\overline{x}/s$
- $\overline{z} = 0$ en $s_z^2 = 1$
- doel standaardisatie:
  een dataset creëren met gemiddelde 0 en variantie 1

Box-plot


250,00
200,00
150,00
100,00
50,00
0,00
Afstand
Q_3
M_e
Q_1
$x_4 < Q_1 - 1.5Q$
$x_1 < Q_1 - 3Q$
4
20
1

Bivariate kwantitatieve variabelen



Bivariate kwantitatieve variabelen

- steekproefcovariantie

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

- populatiecovariantie

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_X)(y_i - \mu_Y)$$

- $s_x^2 = s_{xx}$ en $\sigma_x^2 = \sigma_{xx}$
  variantie = covariantie van een variabele met zichzelf

Bivariate kwantitatieve variabelen

- steekproef- en populatiecovariantie zijn afhankelijk van de meeteenheid
- steekproefcorrelatiecoëfficiënt: (ligt tussen -1 en $+1$)

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- populatiecorrelatiecoëfficiënt: (ligt tussen -1 en $+1$)

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- r_{xy} en ρ_{xy} zijn onafhankelijk van de meeteenheid

Bivariate kwantitatieve variabelen

- correlatiecoëfficiënt geeft aan in welke mate er een lineair verband is tussen twee variabelen
- $y_i = ax_i + b$
- $r_{xy} = ?$
- $s_{xy} = as_{xx}$ en $s_y = |a|s_x$
- $r_{xy} = \frac{a}{|a|} = +1$ of -1

Lineaire combinaties van variabelen

- $u_i = ax_i + by_i + c$
- bewijs dat $\overline{u} = a\overline{x} + b\overline{y} + c$
- bewijs dat $s_u^2 = a^2 s_x^2 + b^2 s_y^2 + 2abs_{xy}$