Hoofdstuk 3 :

Numerieke beschrijving van data

Marnix Van Daele

Marnix.VanDaele@UGent.be

Vakgroep Toegepaste Wiskunde en Informatica
Universiteit Gent

Beschrijvende maten

- We beschrijven populaties en steekproeven d.m.v. karakteristieken
  - populaties worden gekenmerkt door parameters
    $\mu,\ \sigma,\ \rho,\ \ldots$
  - steekproeven worden gekenmerkt door statistieken
    $\overline{x},\ s,\ r,\ \ldots$
- 3 soorten karakteristieken
  - centraliteitsmaten beschrijven de ligging (location)
  - spreidingsmaten beschrijven de spreiding (dispersion)
  - vormmaten beschrijven de vorm

Centraliteitsmaten

- het rekenkundig gemiddelde
- de mediaan
- de modus
- en het meetkundig gemiddelde

Het rekenkundig gemiddelde

Het (rekenkundig) gemiddelde (mean), van $x_1, x_2, \ldots, x_n$ is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

Het gemiddelde van de waarden 1, 2, 3, 4 en 5 bedraagt

$$\overline{x} = \frac{1+2+3+4+5}{5} = 3$$

Het gemiddelde van 1, 2, 3, 4 en 50 bedraagt

$$\overline{x} = \frac{1+2+3+4+50}{5} = 12$$

$\overline{x}$ is gemakkelijk te berekenen maar is gevoelig voor uitschieters.

Middel tegen die gevoeligheid : trimmed mean

Het rekenkundig gemiddelde

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad\qquad \overline{f(x)} = \frac{1}{n}\sum_{j=1}^{n} f(x_j)$$

$$\overline{a\,x + b} = \frac{1}{n}\sum_{j=1}^{n}(a\,x_j + b) = a\,\overline{x} + b$$

Bijzonder geval : $\overline{x - \overline{x}} = 0$

$$\begin{aligned}\overline{f(x) + g(x)} &= \frac{1}{n}\sum_{j=1}^{n}(f(x_j) + g(x_j)) \\ &= \frac{1}{n}\sum_{j=1}^{n} f(x_j) + \frac{1}{n}\sum_{j=1}^{n} g(x_j) = \overline{f(x)} + \overline{g(x)}\end{aligned}$$

Het rekenkundig gemiddelde

Gegeven : frequentietabel

Gevraagd : bepaal $\overline{x}$.

- discrete data : heeft x_i absolute frequentie n_i, dan

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}\sum_{i} n_i\, x_i$$

Voorbeeld : gemiddelde van 1, 2, 1, 3 en 2

$$\begin{aligned}\overline{x} &= \frac{1}{5}\sum_{j=1}^{5} x_j = \frac{1+2+1+3+2}{5} = 1.8\\ &= \frac{1}{5}\sum_{i=1}^{3} n_i\, x_i = \frac{2\times 1+2\times 2+1\times 3}{5} = 1.8\end{aligned}$$

Het rekenkundig gemiddelde

Gegeven : frequentietabel

Gevraagd : bepaal $\overline{x}$.

- continue data : benader elke x_i door het klassemidden t_j waarvoor $t_i - \dfrac{\Delta_i}{2} \leq x_j < t_i + \dfrac{\Delta_i}{2}$

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \approx \frac{1}{n}\sum_{i} n_i\, t_i$$

$$\begin{aligned} \overline{x} &= \frac{1}{117}\sum_{j=1}^{117} x_j = 180.538 \\ &\approx \frac{1}{117}\sum_{i} n_i\, t_i = 180.5385 \end{aligned}$$

Mediaan

De mediaan (median) van $x_1, x_2, \ldots, x_j, \ldots, x_n$ is de middelste waarde als de metingen gerangschikt worden van klein naar groot.

De mediaan van de waarden 1, 2, 3, 4 en 5 bedraagt 3.
De mediaan van 1, 2, 3, 4 en 50 bedraagt 3.

De mediaan is minder gevoelig dan het gemiddelde en kan ook gebruikt worden bij ordinale data.

De modus

De modus (mode) van een verzameling meetwaarden wordt gedefinieerd als de waarde waarvoor de frequentie het hoogst is. In geval gewerkt wordt met klassen, spreekt men van de modale klasse.

Gebruik :

- bij grote steekproeven de meest populaire waarde aanduiden
- bij bimodale verdelingen

Het meetkundig gemiddelde

Het meetkundig gemiddelde (geometric mean) van $x_1, x_2, \ldots, x_j, \ldots, x_n$ wordt gedefinieerd als

$$\text{GM} = \sqrt[n]{x_1\ x_2\ \cdots\ x_j\ \cdots\ x_n}.$$

$$\log \text{GM} = \frac{1}{n}\sum_{i=1}^{n} \log x_i = \overline{\log x}$$

De logaritme van GM $=$ het (rekenkundig) gemiddelde van de logaritme van de waarnemingen.

Het GM van 10, 100 en 1000 bedraagt 100 vermits

$$\text{GM} = \sqrt[3]{10 \times 100 \times 1000} = 100.$$

$$\log_{10} \text{GM} = \frac{1}{3}\sum_{i=1}^{3} \log_{10}\ x_i = \frac{1}{3}(1+2+3) = 2 \Longrightarrow \text{GM} = 10^2 = 100$$

Centraliteitsmaten : richtlijnen

Twee factoren spelen een rol :

- de schaal (kwantitatief of niet-kwantitatief)
- symmetrisch- of scheef-zijn van de verdeling van de waarnemingen

Richtlijnen :

- $\overline{x}$: bij kwantitatieve data en voor (min of meer) symmetrische distributies
- mediaan : bij ordinale data en voor kwantitatieve data waarvan de distributie scheef is
- modus : bij bimodale verdelingen
- meetkundig gemiddelde : bij observaties gemeten op een logaritmische schaal

Spreidingsmaten

- minimum en maximum
- range
- standaarddeviatie en variantie
- variatiecoëfficiënt
- percentielen

De range

De range van een verzameling meetwaarden $x_1, x_2, \ldots, x_j, \ldots, x_n$ wordt gedefinieerd als het verschil tussen de grootste en de kleinste meetwaarde.

Minimum en maximum

Kleinste en grootste meetwaarde

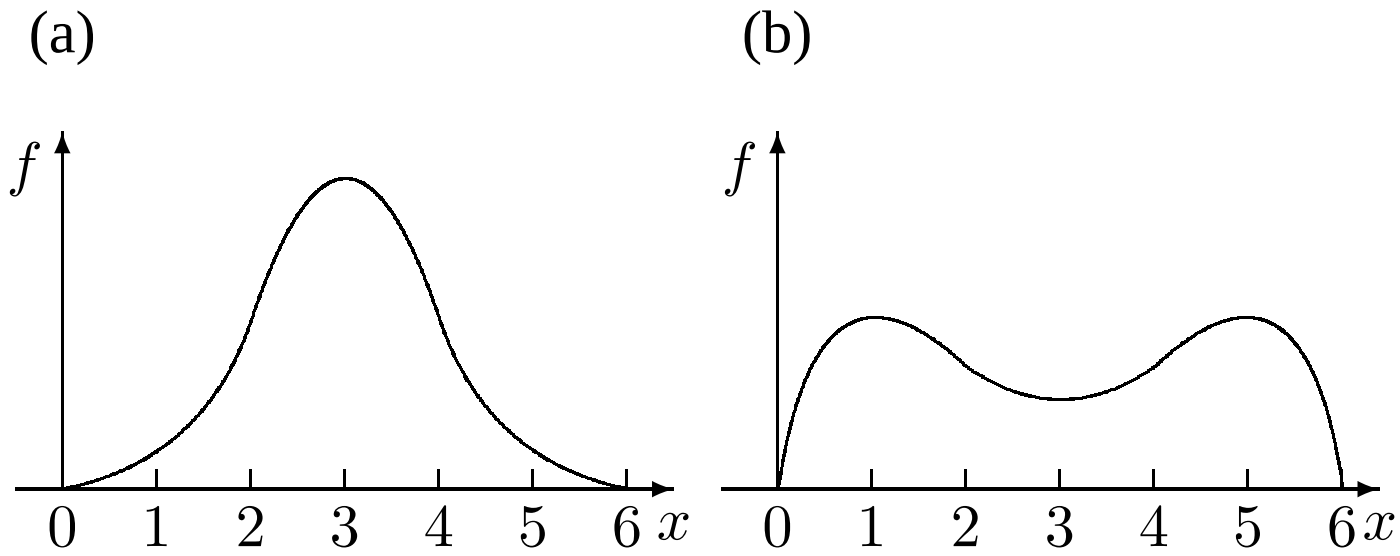
Dit geeft iets meer informatie dan de range.

Voorbeeldsteekproef :

- uit meetwaarden : $\min = 164$ cm en $\max = 196$ cm, d.w.z. range $= 32$ cm
- uit frequentietabel : $\min = 163.5$ cm en $\max = 196$ cm, d.w.z. range $= 33$ cm

Probleem

Noch de range, noch min-max kunnen verschillen detecteren tussen volgende verdelingen :



Spreidingsmaten

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

$$\overline{x - \overline{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\overline{x} = \overline{x} - \overline{x} = 0$$

- gemiddelde afwijking : $\overline{|x - \overline{x}|} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$

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

Steekproefvariantie

De variantie (variance) s_X^2 van een verzameling van n waarden $x_1, x_2, \ldots, x_n$ van de grootheid X wordt gedefinieerd als het gemiddelde van de kwadraten van de afwijkingen van de waarden t.o.v. hun gemiddelde $\overline{x}$:

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

De standaarddeviatie (standard deviation) of standaardafwijking s_X wordt gedefinieerd als de positieve vierkantswortel van de variantie :

$$s_X = \sqrt{s_X^2} \, .$$

Verbeterde steekproefvariantie

De steekproefvariantie

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

is een benadering voor de populatievariantie σ_X^2.
Men kan aantonen dat s_X^2 systematisch een te kleine benadering levert voor σ_X^2 en dat een betere benadering gegeven wordt door de zogenaamde verbeterde steekproefvariantie $s_X'^2$ met

$$s_X'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{n-1}s_X^2 \,.$$

Steekproefvariantie

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

$$\begin{aligned} s_X^2 &= \overline{(x-\overline{x})^2} \\ &= \overline{x^2 - 2\,\overline{x}\,x + \overline{x}^2} \\ &= \overline{x^2} - 2\,\overline{x}\,\overline{x} + \overline{x}^2 \\ &= \overline{x^2} - \overline{x}^2 \end{aligned}$$

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

Steekproefvariantie van functies

$$s^2_{f(X)} = \overline{[f(x)]^2} - \overline{f(x)}^2$$

Toegepast op $f(x) = a\,x + b$

$$\begin{aligned}
s^2_{a\,X+b} &= \overline{(a\,x+b)^2} - \overline{a\,x+b}^2 \\
&= \overline{a^2\,x^2 + 2\,a\,b\,x + b^2} - (a\,\overline{x} + b)^2 \\
&= a^2\,\overline{x^2} + 2\,a\,b\,\overline{x} + b^2 - (a^2\overline{x}^2 + 2\,a\,b\,\overline{x} + b^2) \\
&= a^2\,(\overline{x^2} - \overline{x}^2) \\
&= a^2\,s^2_X
\end{aligned}$$

$$s_{a\,X+b} = |a|\,s_X$$

Ongelijkheid van Chebyshev

Voor om het even welke positieve waarde k geldt : minstens een fractie $1 - 1/k^2$ van alle meetwaarden ligt in het interval

$$]\overline{x} - k\,s, \overline{x} + k\,s[.$$

Bewijs : gegeven n, $\overline{x}$ en s; kies k. Verdeel de meetwaarden in

$$D = \{x_j \mid |x_j - \overline{x}| < k\,s\} \text{ en } V = \{x_j \mid |x_j - \overline{x}| \geq k\,s\}\,,$$

zodat $\#D + \#V = n$

$$n\,s^2 = \sum_{x_j \in D \cup V} (x_j - \overline{x})^2 \geq \sum_{x_j \in V} (x_j - \overline{x})^2 \geq \sum_{x_j \in V} k^2\,s^2 = k^2\,s^2\,(\#V)$$

$$\Longleftrightarrow \frac{\#V}{n} \leq \frac{1}{k^2}\,,$$

d.w.z. de fractie van de n meetwaarden die tot V behoren is hoogstens $1/k^2$ en dus ligt minstens $1 - 1/k^2$ in D.

Ongelijkheid van Chebyshev

Voor om het even welke positieve waarde k geldt : minstens een fractie $1 - 1/k^2$ van alle meetwaarden ligt in het interval

$$]\overline{x} - k\,s, \overline{x} + k\,s[.$$

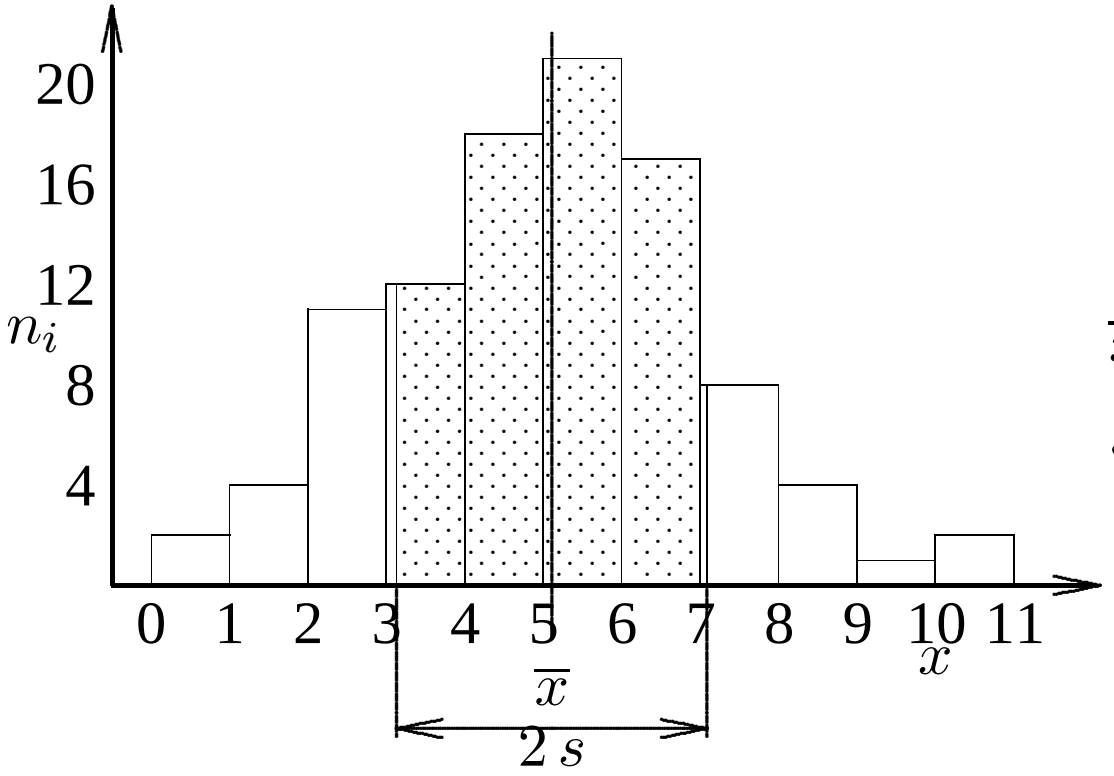
k	$]\overline{x} - k\,s, \overline{x} + k, s[$	$1 - \frac{1}{k^2}$
1	$]\overline{x} - 1\,s, \overline{x} + 1, s[$	$0 = 0\%$
2	$]\overline{x} - 2\,s, \overline{x} + 2\,s[$	$\frac{3}{4} = 75\%$
3	$]\overline{x} - 3\,s, \overline{x} + 3\,s[$	$\frac{8}{9} \approx 90\%$

Deze regel geldt altijd, hoe het histogram er ook uitziet !

In de praktijk zijn de vermelde fracties meestal hoger !

Vuistregel voor belvormige verdelingen

- ongeveer 68 % ligt in $]\overline{x} - s, \overline{x} + s[=]3.128,\ 7.091[$
- ongeveer 95 % ligt in $]\overline{x} - 2\,s, \overline{x} + 2\,s[=]1.146,\ 9.073[$
- bijna alle metingen liggen in
  $]\overline{x} - 3\,s, \overline{x} + 3\,s[=] - 0.836,\ 11.055[$



$\overline{x} = 5.109$

$s = 1.981$

De z-score van een meetwaarde

Als de meetwaarden x_j uitgedrukt zijn in bvb. meter, dan

- is $\overline{x}$ ook in meter
- is s_X^2 in vierkante meter
- is s_x in meter

$$\text{Transformatie : } z_j = \frac{x_j - \overline{x}}{s_X}$$

z_j is dimensieloos met waarden in $[-3, 3]$

Deze transformatie fungeert als

een soort standaardisatie van de meetwaarden.

Variatiecoëfficiënt

De variatiecoëfficiënt (variation coefficient) van een verzameling niet-negatieve meetwaarden $x_1, x_2, \ldots, x_i, \ldots, x_n$ van de grootheid X wordt gedefinieerd als

$$\frac{s}{\overline{x}} .$$

Spreidingsmaten : richtlijnen

- s_X : als $\overline{x}$ wordt gebruikt, d.i. bij min of meer symmetrische kwantitatieve data.
- Percentielen en interquartielen :
  - wanneer de mediaan wordt gebruikt : bij ordinale data of bij scheef-verdeelde kwantitatieve data
  - wanneer $\overline{x}$ wordt gebruikt, maar als het de bedoeling is individuele waarnemingen te vergelijken met een verzameling normen
- interquartiele range : voor de beschrijving van de centrale 50 % van een distributie, onafhankelijk van de vorm
- range : bij kwantitatieve data als het de bedoeling is de nadruk te leggen op extreme waarden
- variatiecoëfficiënt : indien kwantitatieve verdelingen op verschillende schalen worden vergeleken

Vormmaten

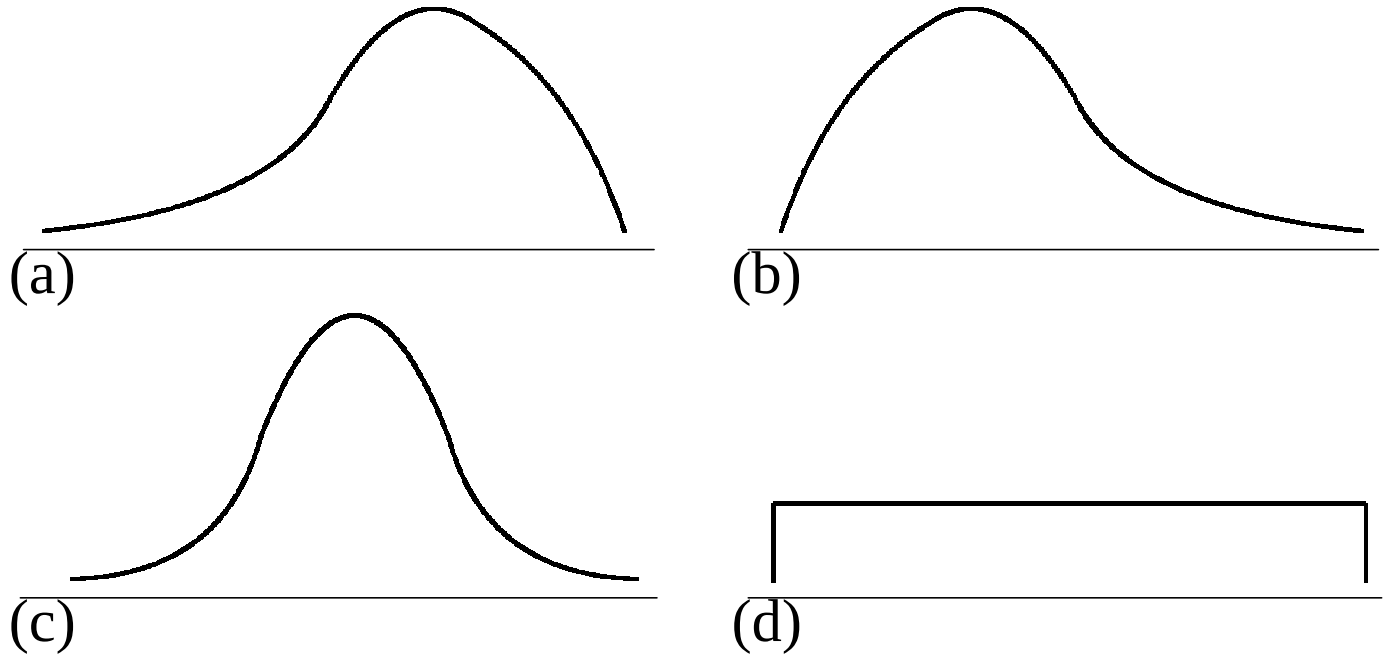
- $\overline{x - \overline{x}} = 0$
- $\overline{(x - \overline{x})^2}$: variantie (spreidingsmaat)
- $\overline{(x - \overline{x})^3}$: scheefheid
- $\overline{(x - \overline{x})^4}$: kurtosis

scheefheid en kurtosis zijn vormmaten

Scheefheid

De scheefheid (skewness) van $x_1, x_2, \ldots, x_j, \ldots, x_n$ wordt gedefinieerd als

$$\frac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^3}{s^3}$$



(a) negatief scheef (b) positief scheef

(c) en (d) symmetrisch

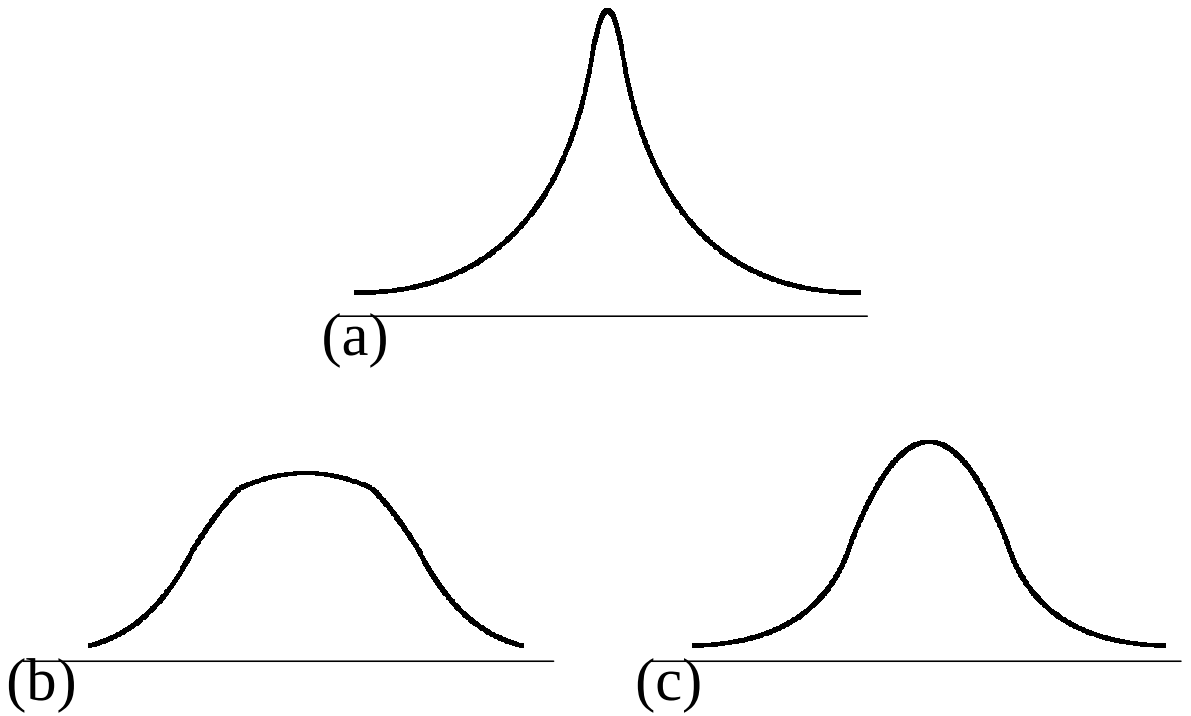
Scheefheid

Verband met ligging van mediaan en gemiddelde

- Zijn de mediaan en het gemiddelde gelijk, dan is de distributie min of meer symmetrisch.
- Is het gemiddelde groter dan de mediaan, dan is de distributie positief scheef.
- Is het gemiddelde kleiner dan de mediaan, dan is de distributie negatief scheef.

Kurtosis

De kurtosis (curtosis) van $x_1, x_2, \ldots, x_j, \ldots, x_n$ wordt gedefinieerd als $\dfrac{\frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^4}{s^4}$.



(a) leptokurtisch **(b)** platykurtisch en **(c)** kurtosis ≈ 3

Een voorbeeld

Descriptives

	GESLACHT			Statistic	Std. Er
GEWICHT	m	Mean		68,87	,79
		95% Confidence Interval for Mean	Lower Bound	67,30	
			Upper Bound	70,44	
		5% Trimmed Mean		68,72	
		Median		68,00	
		Variance		73,320	
		Std. Deviation		8,56	
		Minimum		52	
		Maximum		90	
		Range		38	
		Interquartile Range		12,50	
		Skewness		,206	,224