H1: Gegevens

1.1 soorten gegevens

1.1.1 kwalitatieve gegevens

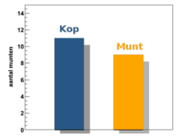
Kwalitatieve gegevens	= kenmerken, classes of categorieën - nominale variabele = waarde van de variabele plaatst meetgegeven in bepaalde categorie vb: geslacht, nationaliteit, kleur - ordinale variabele = er is logische ordening tussen elementen in een steekproef vb: sterren voor hotels/restaurants > heeft geen vaste meeteenheid

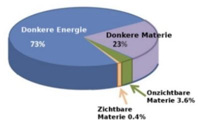
1.1.2 kwantitatieve gegevens

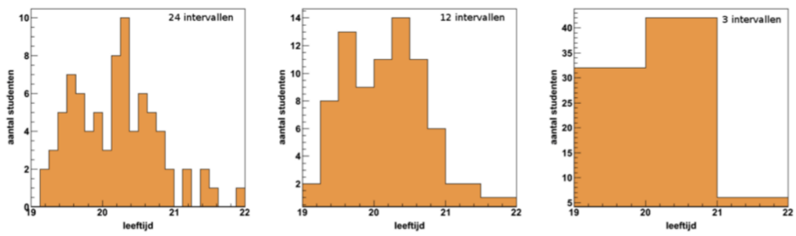
Kwantitatieve gegevens	= kunnen in vaste meeteenheden uitgedrukt worden - discrete gegevens = worden in gehele getallen uitgedrukt - continue gegevens = worden in reële getallen uitgedrukt, mist afronding
schaal van kwantitatieve gegevens	- intervalschaal = geen natuurlijk nulpunt > verhoudingen niet zinvol berekenbaar - ratioschaal = wel een nulpunt > verhoudingen zijn zinvol

1.1.4 gegevens voorstellen

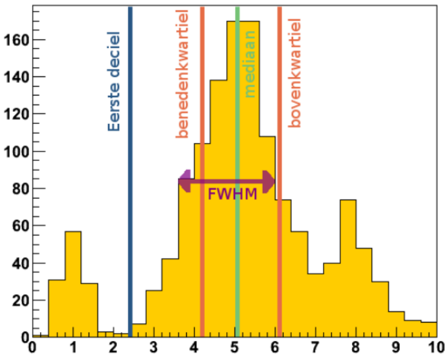
def: absolute frequentie van een klasse	= het aantal elementen van de steekproef die tot de klasse behoren
def: relatieve frequentie ve klasse	= de verhouding van de absolute frequentie tot het aantal waarnemingen vd proef
kwalitatieve variabelen voorstellen	1: staafdiagram: 2: taartdiagram:
Kwantitatieve variabelen voorstellen	histogrammen: maak de afweging tss nauwkeurigheid en leebaarheid nl: hoe breder de intervallen, hoe minder nauwkeurig <> meer leesbaar smaller meer minder

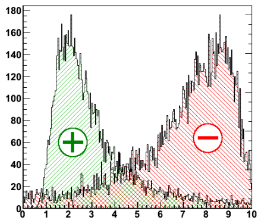
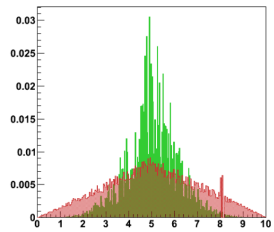






1.2 numerieke beschrijving van gegevens	
1.2.1 beschrijving van de centrale waarde van gegevens	
def: aritmetisch gemiddelde /rekenkundig	Het aritmetisch gemiddelde van N elementen van x is: $$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$ voor N elementen in B intervallen: $$\bar{x} = \frac{1}{N}\sum_{j=1}^{B} n_j x_j$$ waarbij de interval j slaat op de waarde x_j en bevat n_j tellen. Voor een functie f(x): $$\bar{f} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$ en met de verdeling in B intervallen $$\bar{f} = \frac{1}{N}\sum_{j=1}^{B} n_j f(x_j)$$
def: geometrisch gemiddelde /meetkundig	Voor N elementen van x: $$g = \left(\prod_{i=1}^{N} x_i\right)^{1/N}$$
stelling: logaritme en gemiddelden	De logaritme van het meetkundig gemiddelde is het rekenkundig gemiddelde van de afzonderlijke logaritmen nl: $\ln g = \ln \sqrt[N]{x_1 \cdot x_2 \cdot \ldots \cdot x_N}$ $= \frac{1}{N}\ln(x_1 \cdot x_2 \cdot \ldots \cdot x_N)$ $= \frac{1}{N}(\ln x_1 + \ln x_2 + \ldots + \ln x_N)$ $= \overline{\ln x}$
def: harmonisch gemiddelde	= reciproke van rekenkundig gemiddelde van de reciproken $$h = \left(\frac{1}{N}\sum_{i=0}^{N} x_i^{-1}\right)^{-1}$$
def: kwadratisch gemiddelde	$$k = \sqrt{\frac{\sum_i x_i^2}{N}}$$
def: modus	= meest populaire waarde in een groep gegevens
stelling: modus van een verdeling	Voor continue gegevens vind je de modus door de verdeling te differentiëren en gelijk te stellen aan 0
def: mediaan	= het halfweg-punt > de helft van de punten bevinden zich beneden het mediaan boven dus: • als het aantal elementen N oneven is, is de mediaan het $\frac{(N+1)}{2}$-de element; • als het aantal elementen N even is, is de mediaan het gemiddelde van het $\frac{N}{2}$-de en $(\frac{N}{2}+1)$-de element.
1.2.2 beschrijving van de spreiding van gegevens	
def: spreidingsbreedte/range R	= het verschil tss de hoogste en laagste waarde van gegevens: $$R = x_{\max} - x_{\min}$$
def: gemiddelde absolute afwijking	= rekenkundig gemiddelde van de absolute waarden van de afwijkingen tov het rekenkundig gemiddelde $$GAA = \frac{1}{N}\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert$$

def: variantie	= de gemiddelde waarde van het kwadraat van de afwijkingen vd variabele vh gemiddelde $V(x) \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$ > hetzelfde kan met een functie: $V(f) \equiv \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - \bar{f})^2$
stelling: variantie en gemiddelde	variantie = het gemiddelde van de kwadraten min kwadraat van gemiddelde: $V(x) \equiv \frac{1}{N}\sum_i (x_i - \bar{x})^2 \quad = \quad \frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x\right)^2 \quad = \quad \overline{x^2} - \bar{x}^2$
def: standaardafwijking	= positieve vierkantswortel van de variantie: $\sigma = \sqrt{V(x)} = \sqrt{\overline{x^2} - \bar{x}^2} = \sqrt{\frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x\right)^2} = \sqrt{\frac{1}{N}\sum_i (x_i - \bar{x})^2}$
def: schatting voor standaardafwijking	$s = \sqrt{\frac{1}{N-1}\sum_i (x_i - \bar{x})^2}$
def: p-de percentiel c_p	= een getal dat groter is dan p% van de waarnemingen en kleiner dan (1-p)% van de waarnemingen > vb: mediaan is het 50ste percentiel
def: bedenkwartiel bovenkwartiel	= 25e percentiel = 75e percentiel
def: volle breedte op halve hoogte	= afstand tss de twee snijpunten van het histogram op de helft van de hoogte van de centrale piek = FWHM
stelling: FWHM en Gaussische verdeling	De FWHM van een Gaussische verdeling is 2.35σ
verdelingen visueel:	
1.2.3 momenten van de gegevens	
def: r-de steekproefmoment	= de gemiddelde waarde van de r-de macht van de waarnemingen: $m'_r = \frac{1}{N}\sum_{i=1}^{N} x_i^r$
def: r-de centrale moment	= het gemiddelde vd r-de macht vd afwijkingen ten opzichte vh steekproefgemiddelde $m'_r = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^r$
def: scheefheid	$\gamma = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(x_i - \bar{x})^3 = \frac{1}{\sigma^3}\overline{(x - \bar{x})^3}$ > pos. scheefheid impliceert staart naar rechts neg. scheefheid links
stelling: scheefheid en steekproefmoment	$m_3 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3 = \gamma\sigma^3$

def: kurtosis	$c = \frac{1}{\sigma^4}\overline{(x-\bar{x})^4} - 3$ > pos. kurtosis impliceert relatief hoge piek en lange staarten neg. kurtosis plompere verdeling met brede piek en korte staarten  *Figuur 1.6: Algemene distributies met positieve en negatieve scheefheid*  *Figuur 1.7: Algemene distributies met positieve en negatieve kurtosis*
1.3 multivariabele variabelen	
1.3.1 bivariate variabelen: covariantie en correlatie	
bivariate variabele	= steekproef op twee variabelen X en Y > elk evenement bestaat uit een paar van getallen: $\{(x_1,y_1),(x_2,y_2),(x_3,y_3),\ldots,(x_N,y_N)\}$ Hierbij heeft elke variabele X en Y onafhankelijk van elkaar: gemiddelden $\bar{x}$, $\bar{y}$, de varianties $V(x)$, $V(y)$ en de standaardafwijkingen σ_x en σ_y
def: covariantie	covariantie tss de variabelen X en Y: $\mathrm{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) = \overline{(x-\bar{x})(y-\bar{y})} = \overline{xy} - \bar{x}\bar{y}$
stelling: verband variantie en covariantie	De covariantie is een triviale uitbereiding van de univariabele variantie > variantie is de covariantie van de variabele met zichzelf $V(x) = \mathrm{cov}(x,x)$
def: correlatiecoëfficiënt	= de verhouding tss de covariantie en de individuele standaardafwijkingen $\rho = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y}$ > is een getal zonder eenheden, begrensd door -1 en +1 > geeft aan in welke mate er een lineair verband is tss de twee variabelen
1.3.2 multivariabele uitbreiding: covariantie- en correlatie matrix	
multivariabele variabelen	Een steekproef met N elementen van m variabelen wordt geschreven als: $\{(x_{(1)},x_{(2)},\ldots,x_{(m)})_1 \quad , \quad (x_{(1)},x_{(2)},\ldots,x_{(m)})_2 \quad , \quad \ldots \quad , \quad (x_{(1)},x_{(2)},\ldots,x_{(m)})_N\}$ - subscript tss haakjes duiden element aan van individueel evenement - subscript zonder haakjes duid op evenement binnen steekproef
multivariabele covariantie	tussen elk paar variabelen geldt: $\mathrm{cov}(x_{(i)},x_{(j)}) = \overline{x_{(i)}x_{(j)}} - \overline{x_{(i)}}\;\overline{x_{(j)}}$
covariantiematrix	We kunnen de multivariabele covariantie in matrixvorm schrijven: met elk element V_{ij}: $V_{ij} = \mathrm{cov}(x_{(i)},x_{(j)})$ dus: $\begin{pmatrix} \sigma_1^2 & \mathrm{cov}(x_{(1)},x_{(2)}) & \cdots & \mathrm{cov}(x_{(1)},x_{(m)}) \\ \mathrm{cov}(x_{(2)},x_{(1)}) & \sigma_2^2 & \cdots & \mathrm{cov}(x_{(2)},x_{(m)}) \\ \vdots & & \ddots & \\ \mathrm{cov}(x_{(m)},x_{(1)}) & \mathrm{cov}(x_{(m)},x_{(2)}) & \cdots & \sigma_m^2 \end{pmatrix}$
correlatiematrix	= matrixvorm van de correlatiecoëfficiënten: $\rho_{ij} = \frac{\mathrm{cov}(x_{(i)},x_{(j)})}{\sigma_i\sigma_j}$