**Samenvatting beschrijvende statistiek**: formules, definities,werkwijzen en eigenschappen

VERZAMELINGEN

**Kardinaalgetal** = het aantal elementen van een verzameling

**Machtsverzameling** = de verzameling van alle mogelijke deelverzamelingen

$∩$: doorsnede
$∪$: unie
$∖$: verschil
$ ^{c}$: complement
$\#A$: kardinaalgetal
$D⊂A$: deelverzameling
$2^{A}$: machtsverzameling
$A\_{1} × A\_{2}$: productverzameling

**Eigenschappen van verzamelingen:**
A⋂B = B⋂A
A⋂(B⋂C) = (A⋂B) ⋂C
A⋂∅ = ∅
A⋂(B⋃C) = (A⋂B) ⋃(A⋂C)
(A⋂B)c = Ac⋃BcA⋃B = B⋃A
A⋃(B⋃C) = (A⋃B) ⋃C
A⋃∅ = A
A⋃(B⋂C) = (A⋃B) ⋂(A⋃C)
(Ac)c = A
(A⋃B)c = Ac⋂Bc

**Partitie** = opsplitsing van een verzameling in een stel niet-lege en niet-overlappende deelverzamelingen

**Productverzameling** = verzameling van alle geordende koppels

**Eigenschappen van het sommatieteken:**$$\sum\_{i}^{} \left(x\_{i}+y\_{i}\right)= \sum\_{i}^{} x\_{i}+\sum\_{i}^{} y\_{i}$$$$\sum\_{i}^{} \left(cx\_{i}\right)=c\sum\_{ }^{} x\_{i}$$$$\sum\_{i=1}^{n} c=n c$$$$\left(\sum\_{i=1}^{n}x\_{i} \right)\left(\sum\_{j=1}^{m}y\_{i} \right)=\sum\_{i=1}^{n} \sum\_{j=1}^{m} x\_{i}y\_{j}$$

ℕ: de natuurlijke getallen (0,1,2,3, …)
ℤ: de gehele getallen (…, -2, -1, 0, 1, 2, …)
ℚ: breuken
ℝ: de reële getallen
∅: lege verzameling
𝒰: universele verzameling

BESCHRIJVENDE STATISTIEK

**Kwalitatieve variabele** = het bereik bestaat uit een aantal waarden (categorieën) waarover geen verdere claims worden gemaakt

**Kwantitatieve variabele** = het bereik bestaat uit numerieke waarden waarvoor ordening, optellen en aftrekken zinvol is

$$\sum\_{j=1}^{m}freq(x\_{j})=freq\left(x\_{1}\right)+ freq\left(x\_{2}\right)+ …+freq(x\_{m})$$

p(xj) = $\frac{freq(x\_{j})}{n}$

**Lijndiagram, frequentiefunctie:**
Y-as = frequentie
X-as = X
Lijntjes tekenen

**Staafdiagram, frequentiefunctie:**
Y-as = frequentie
X-as = X
Staven tekenen (met spaties, WANT geen histogram)

**Histogram, proportiefunctie:**
Y-as = proportie
X-as = X
staven tekenen (geen spatie, WANT geen staafdiagram)

**Polygoonvoorstelling:**
🡪 middelpunten van de bovenste lijnstukken van de histogramstaven met elkaar verbinden

**Cumulatieve frequentie** = $cfreq\_{ }\left(x\_{j}\right)$ = het aantal observaties met waarde kleiner of gelijk aan Xi

|  |  |  |
| --- | --- | --- |
| naam | notatie | definitie |
| percentielen | Pc1 … Pc99 | x.01 x.02 x.03 … x.99 |
| decielen | D1 … D9 | x.1 x.2 x.3 … x.9 |
| kwartielen | Q1 … Q3 | x.25 x.50 x.75 |

**Cumulatieve proportie** = $F\_{ }(x\_{j})$

**Cumulatieve proportiefunctie:**
Y-as = F
X-as = X
Lijnen tekenen (gekleurde bol, lege bol)
Laten doorlopen tot het oneindige

**r-de kwantiel zoeken:**
GEVAL 1: geen xi geobserveerd met F(xi) = r
🡪 kiezen voor het eerstvolgende getal 🡪 dat is dan je r-de kwantiel
GEVAL 2: er is een xi geobserveerd met F(xi) = r
🡪 het gemiddelde van dat getal en het eerstvolgende getal

**Klassegrenzen** = xi en xh
**Klassemiddelpunt** = $\frac{x\_{i}+x\_{h}}{2}$
**Histogram met klassen:**
Y-as = F
X-as = X
X is telkens het klassemiddelpunt van die klasse

**Gegroepeerde frequentietabel:**
🡪 stam- en loofdiagram

**Modus** = elke waarde waarvoor freq(x) maximaal is

1 modus = unimodaal
2 modi = bimodaal

**Mediaan** = (Mex): Pc50 = D5 = Q2

**Mediaan zoeken:**
🡪orden de observaties naar grootte-orde en hernummer ze
🡪als n oneven is: $Me\_{x}=x\_{\frac{n+1}{2}}$
🡪als n even is: $Me\_{x}$= $\frac{x\_{\frac{n}{2}}+x\_{\frac{n}{2}+1}}{2}$

 $\overbar{x}=\frac{1}{n}\sum\_{i=1}^{n}x\_{i}$

$$\overbar{x}=\frac{1}{n}\sum\_{j=1}^{m}freq\left(x\_{j}\right)x\_{j}=\sum\_{j=1}^{m}p(x\_{j})x\_{j}$$

**Eigenschappen van het gemiddelde:**
 $\sum\_{i=1}^{n}\left(x\_{i}-\overbar{x}\right)=0$
Als c $\ne \overbar{x}$ dan $\sum\_{i=1}^{n}(x\_{i}-\overbar{x})^{2}<\sum\_{i=1}^{n}(x\_{i}-c)^{2}$
Regel van Steiner: $\sum\_{i}^{}(x\_{i}-c)^{2}= \sum\_{i}^{}(x\_{i}-\overbar{x})^{2}+ n (\overbar{x}-c)^{2}$

**Eigenschap van mediaan:**
Als c $\ne Me\_{x}$ dan $\sum\_{i=1}^{n}|x\_{i}-Me\_{x}|\leq \sum\_{i=1}^{n}|x\_{i}-c|$

**Bereik**: verzamelingen van alle functiewaarden

**Bereik**
max(x) – min(x)

**Interkwartielbereik**
Q3 – Q1

**Variantie**: de gemiddelde kwadratische afstand van de observaties t.o.v. hun gemiddelde

**Variantie**
 $\left(s\_{x}^{2}\right)= \frac{1}{n}\sum\_{i=1}^{n}(x\_{i}-\overbar{x})^{2}$
$$\left(s\_{x}^{'2}\right)= \frac{1}{n-1}\sum\_{i=1}^{n}(x\_{i}-\overbar{x})^{2}$$

**Standaarddeviatie**: een maat voor de spreiding van een variabele of van een verdeling

**Standaarddeviatie**
 $\left(s\_{x}\right)= \sqrt{\frac{1}{n}\sum\_{i=1}^{n}(x\_{i}-\overbar{x})^{2}}$
$$(s\_{x}^{'})= \sqrt{\frac{1}{n-1}\sum\_{i=1}^{n}(x\_{i}-\overbar{x})^{2}}$$

**Berekening van** $s\_{x}^{2} en s\_{x}$**:**
METHODE 1: berekenen per observatie $x\_{i}-\overbar{x}$ en maat direct gebruik van de bovenstaande definities
METHODE 2: maak gebruik van de volgende gelijkheid:
 $s\_{x}^{2}=\frac{1}{n}\sum\_{i}^{}x\_{i}^{2}-\overbar{x^{2}}$
 Bereken daartoe per observatie $x\_{i}^{2}$ enz…
METHODE 3: Vertrek van de frequentie- of proportietabel en maak gebruik van de formules
 $s\_{x}^{2}=\frac{1}{n}\sum\_{j=1}^{m}freq\left(x\_{j}\right)(x\_{j}-\overbar{x})^{2}$ OF $s\_{x}^{2}=\frac{1}{n}\sum\_{j=1}^{m}freq\left(x\_{j}\right)x\_{j}^{2}-\overbar{x^{2}}$
 $s\_{x}^{2}=\sum\_{j=1}^{m}p\left(x\_{j}\right)(x\_{j}-\overbar{x})^{2}$ OF $s\_{x}^{2}=\sum\_{j=1}^{m}p\left(x\_{j}\right)x\_{j}^{2}-\overbar{x^{2}}$

**Eigenschappen van variantie:**
Eigenschap 1:
 $2s\_{x}^{2}$ = het gemiddelde kwadratische verschil tussen alle paren observaties
 $2s\_{x}^{2}= \frac{\sum\_{i=1}^{n}\sum\_{i^{'}=1}^{n}(x\_{i}-x\_{i'})^{2}}{n^{2}}$
Eigenschap 2: ongelijkheid van Tchebychev
 als $k>1$ dan $p(\left|X-\overbar{x}\right|\geq k s\_{x})\leq \frac{1}{k^{2}}$
 equivalent met als $k'>1$ dan $p[\left(X-\overbar{x}\right)^{2}\geq k' s\_{x}^{2}]\leq \frac{1}{k^{'}}$

**Boxplot:**Zie boek statistiek p.44-45

**Z-transformatie**
$$Z\_{X}\left(x\_{i}\right)=Z\left(x\_{i}\right)= \frac{x\_{i}-\overbar{x}}{s\_{x}}$$

**Invloed van transformaties op frequentiefuncties:**
als y = f(x)
dan
$$freq\_{y}(y)=freq\_{x}(f^{-1}(y))=\sum\_{x}^{}freq\_{x}(x)$$

 f(X)=Y

**Invloed van transformaties op centrale tendensmaten:**
Als y = f(x) met f(x) = ax + b
dan
$$\overbar{f(x)}=f\left(\overbar{x}\right)= a\overbar{x}+b$$

Bijzonder geval:
$$\overbar{z\_{x}}=0$$

**Invloed van transformaties op spreidingsmaten:**
Stel y = f(x) met f(x) = ax + b
dan
$$s\_{f(x)}^{2}=a^{2}s\_{x}^{2} en dus ook s\_{f(x)}=|a|s\_{x}$$

Bijzonder geval:
$$s\_{z\_{x}}^{2}=1$$

Gevolg:
 Zx als a $>$ 0
ZaX + b=
 -Zx als a $<$ 0

BESCHRIJVENDE STATISTIEK MET 2 VARIABELEN

**Bivariate frequentiefuncties 🡪 bivariate frequentietabel**

|  |  |
| --- | --- |
| Aantal anderen |  |
| Plaats | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | FreqX |
| Thuis | 4 | 5 | 2 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 |
| Werk | 1 | 3 | 2 | 1 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| Buitenshuis | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4 |
| FreqY | 7 | 8 | 4 | 5 | 1 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 30 |

**Bivariate proportiefuncties 🡪 bivariate proportietabel**
Hetzelfde als bovenstaand voorbeeld maar dan met proporties

**Lijndiagrammen**
Per plaats een andere lijndiagram

**Rug-aan-rug stam-en-loofdiagram**
Hetzelfde als een stam-en-loofdiagram maar dan rug aan rug, 2 variabelen

**Rijconditionele proportie**
$p\_{y|x=x\_{j}}(y\_{j^{'}})=\frac{freq\_{x,y}(x\_{j},y\_{j^{'}})}{freq\_{x}(x\_{j})}$=$\frac{p\_{x,y}(x\_{j}y\_{j^{'}})}{p\_{x}(x\_{j})}$

**Kolomconditionele proportie**
$p\_{x|y=y\_{j'}}(x\_{j})=\frac{freq\_{x,y}(x\_{j},y\_{j^{'}})}{freq\_{y}(y\_{j'})}$=$\frac{p\_{x,y}(x\_{j}y\_{j^{'}})}{p\_{y}(y\_{j'})}$

**Scatterdiagram**
Zie boek p.61-62

**Conditionele gemiddelde**
$$\overbar{y|X= x\_{j}}=\frac{1}{freq\_{x}(x\_{j})}\sum\_{j^{'}=1}^{m'}freq\_{x,y}(x\_{j},y\_{j^{'}})y\_{j'}=\sum\_{j^{'}=1}^{m'}p\_{y|x=x\_{j}}(y\_{j^{'}})y\_{j'}$$$$\overbar{x|Y= y\_{j'}}=\frac{1}{freq\_{y}(y\_{j'})}\sum\_{j=1}^{m}freq\_{x,y}(x\_{j},y\_{j^{'}})x\_{j}=\sum\_{j=1}^{m}p\_{x|y=y\_{j'}}(x\_{j})x\_{j}$$

**Conditionele variantie**
$$s\_{y|X=x\_{j}}^{2}=\frac{1}{freq\_{x}(x\_{j})}\sum\_{j^{'}=1}^{m'}freq\_{x,y}(x\_{j},y\_{j^{'}})y\_{j'}^{2}-(\overbar{y|X=x\_{j}})^{2}$$

**Covariantie**
$$\left(s\_{xy}\right):\frac{1}{n}\sum\_{i=1}^{n}\left(x\_{i}-\overbar{x}\right)(y\_{i}-\overbar{y})$$$$s\_{xy}^{'}=\frac{1}{n-1}\sum\_{i=1}^{}\left(x\_{i}-\overbar{x}\right)(y\_{i}-\overbar{y})$$

$$s\_{xx}=s\_{x}^{2}$$

Zie verder boek p.67

**Eigenschappen covariantie**$$s\_{xy}=s\_{yx}$$$$s\_{ax+b y}=a s\_{xy}$$

**Correlatie**$$\left(r\_{xy}\right): s\_{z\_{x}z\_{y}}$$

**Eigenschappen correlatie**$$r\_{xy}=\frac{1}{n}\sum\_{i}^{}Z\_{x}(x\_{i})Z\_{y}(Y\_{i})=\frac{s\_{xy}}{s\_{x}s\_{y}}$$$$r\_{xx}=1$$$$r\_{xy}=r\_{yx}$$$$r\_{ax+b y}=r\_{xy}als a>0;r\_{ax+b y}=-r\_{xy} als a<0$$$$-1\leq r\_{xy}\leq +1$$

**Algemene optimale voorspelling**
$$y\_{j}^{est}=f(x\_{j})$$

**Gekwadrateerde standaardfout van estimatie**$$s\_{y.x}^{2}=\frac{1}{n}\sum\_{i=1}^{n}(y\_{i}-y\_{i}^{est})^{2}$$

**Optimale lineaire voorspelling**$$y\_{i}^{est}=b\_{0}+b\_{1}x\_{i}$$

b0 is de regressieconstante
b1 is het regressiegewicht van X

$$b\_{0}=\overbar{y}-b\_{1}\overbar{x}$$$$b\_{1}=r\_{xy}\frac{s\_{y}}{s\_{x}}$$

$\frac{y\_{i}^{est}-\overbar{y}}{s\_{y}}=r\_{xy}\left(\frac{x\_{i}-\overbar{x}}{s\_{x}}\right)$

$$Z\_{y}\left(y^{est}\right)=r\_{xy}Z\_{x}(x\_{i})$$

$\frac{y\_{i}^{est}-y\_{i'}^{est}}{s\_{y}}=r\_{xy}\left(\frac{x\_{i}-x\_{i'}}{s\_{x}}\right)$

$$Z\_{y}\left(y\_{i}^{est}\right)-Z\_{y}\left(y\_{i'}^{est}\right)=r\_{xy}\left[Z\_{x}\left(x\_{i}\right)-Z\_{x}(x\_{i'})\right]$$

$$\frac{1}{n}\sum\_{i=1}^{n}(y\_{i}-\overbar{y})^{2}=\frac{1}{n}\sum\_{i=1}^{n}(y\_{i}-y\_{i}^{est})^{2}+\frac{1}{n}\sum\_{i=1}^{n}(y\_{i}^{est}-\overbar{y})^{2}$$$$s\_{y}^{2}=s\_{y.x}^{2}+s\_{verkl}^{2}$$

$$\frac{s\_{verkl}^{2}}{s\_{y}^{2}}=r\_{xy}^{2}$$

**Somvariabelen**

$$\overbar{x+y}=\overbar{x}+\overbar{y}$$

$$s\_{x+y}^{2}=s\_{x}^{2}+s\_{y}^{2}+2s\_{xy}$$

$$s\_{x+y z}=s\_{xz}+s\_{yz}$$

**Oefeningen werkwijze:**
-Eerst somvariabelen eigenschappen
-Daarna lineaire transformaties

BESCHRIJVENDE STATISTIEK MET MEER DAN 2 VARIABELEN

$$y\_{j}^{est}=b\_{0}+b\_{1}x\_{1i}+b\_{2}x\_{2i}$$