INLEIDING TOT DE BIOMEDISCHE STATISTIEK

CD-ROM

Wiskundige beschouwingen

Georges De Moor
Georges Van Maele

Inhoudsopgave

Wiskundige beschouwingen

Inleiding

In het boek wordt op verschillende plaatsen verwezen naar dit appendix voor de verduidelijking van mathematische wetmatigheden (formules) en voor de verklaring van de afleiding van wiskundige begrippen.

Deze “Wiskundige Beschouwingen” zijn dan ook een nuttige aanvulling voor diegenen die zich iets verder willen verdiepen in de materie of graag de achtergrond van een afleiding of de uitwerking van een bepaalde formule willen kennen.

W 0. Odds Ratio en Relatief Risico

~ 2.2.2

Odds Ratio (er bestaat geen Nederlandstalig equivalent) (OR) wordt veelvuldig aangewend in klinische studies en meer specifiek bij zogenaamde ‘case-control’ studies die een uitgesproken **retrospectief** karakter hebben.
De resultaten worden voorgesteld in een 2×2-tabel:

	gevallen (vertonen de ziekte)	**controles (zonder ziektebeeld)**
blootgesteld	a	b
niet blootgesteld	c	d
totaal	a + c	b + d

Voorbeeld: case-control studie voor het verband tussen het gebruik van contraceptiva en borstkanker (Vessey et al. 1983).

Orale contraceptiva	**gevallen (borstkanker)**	**controles (geen borstkanker)**
	537	554
	639	622
totaal	1176	1176

De ‘odds ratio’ is de verhouding van beide ‘odds’. De ‘odds’ van een gebeurtenis is de verhouding van de probabiliteit van het optreden van een gebeurtenis tegenover de probabiliteit van het niet optreden van de gebeurtenis:

$$\mathbf{odds} = \frac{\mathbf{kans}}{\mathbf{1 - kans}}\ ;\ \ \mathbf{kans} = \frac{\mathbf{odds}}{\mathbf{odds + 1}}$$

In patiëntcontroleonderzoek geeft de odds-ratio de verhouding van de odds weer van de odds voor blootstelling van patiënten en die van controles (blootstellingsodds-ratio). In cohortonderzoek geeft de odds-ratio de verhouding weer van de odds voor de bestudeerde uitkomst (ziekte) van blootgestelden en die van niet-blootgestelden (ziekteodds-ratio).

Ziekteodds-ratio en blootstellingsodds-ratio verschillen weliswaar conceptueel, maar zijn mathematisch identiek. Bij zeldzame uitkomsten (waarvan doorgaans sprake is in patiëntcontroleonderzoek) is de odds-ratio een goede schatter van het *relatief risico*.

In bovenstaande tabel leest men af dat voor de patiënten (borstkanker, cases) de odds voor contraceptiva gelijk is aan $\frac{\mathbf{a}}{\mathbf{c}}$, voor de controles is dit $\frac{\mathbf{b}}{\mathbf{d}}$.

De odds ratio is gelijk aan: $$\mathbf{OR} = \frac{\frac{\mathbf{a}}{\mathbf{c}}}{\frac{\mathbf{b}}{\mathbf{d}}} = \frac{\mathbf{ad}}{\mathbf{bc}} \quad (\mathbf{OR} \in [0,+\infty[)$$

Let trouwens op de symmetrie van de formule: men bekomt hetzelfde resultaat indien men de odds beschouwt binnen de 'contraceptiva' en binnen de 'niet-gebruikers'.

De odds ratio in bovenstaand voorbeeld is OR = 0.94. De odds ratio verandert niet als men een rij of kolom met een getal vermenigvuldigt. De odds ratio is dus onafhankelijk van de marginale totalen maar enkel van de aantallen binnen de cellen. De odds ratio hangt met andere woorden niet af van de opzet van het studieonderzoek; de odds ratio drukt een maat uit voor de samenhang tussen twee variabelen (in het voorbeeld tussen gebruik contraceptiva en borstkanker).

In een situatie waarbij de studie-opzet deze is van een **prospectieve studie** beschouwt men twee groepen (cohorten $\rightarrow$ cohort study) naargelang men al dan niet blootgesteld werd aan een risicofactor en men gaat na hoeveel gevallen er binnen een bepaalde periode al dan niet het ziektebeeld hebben ontwikkeld:

risicofactor	**ziek**	**niet ziek**	**tijdens periode van follow-up**
blootgesteld	a	b	a + b
niet blootgesteld	c	d	c + d

Het **relatief risico** (RR) wordt gedefinieerd als de verhouding van het optreden van de ziekte in de blootgestelde groep tegenover het optreden van de ziekte in de niet-blootgestelde groep:

$$\mathbf{RR} = \frac{\frac{\mathbf{a}}{\mathbf{a}+\mathbf{b}}}{\frac{\mathbf{c}}{\mathbf{c}+\mathbf{d}}} = \frac{\mathbf{a(c+d)}}{\mathbf{(a+b)c}}$$

$$= \frac{\textbf{incidentie in de blootgestelde groep}}{\textbf{incidentie in de niet blootgestelde groep}} \qquad \mathbf{RR} \in [0,+\infty[$$

Betrouwbaarheidsintervallen:

Odds Ratio: 95%-betrouwbaarheidsinterval van ln(OR) is gelijk aan

$$\mathbf{\ln(OR)} - 1.96\sqrt{\frac{1}{\mathbf{a}}+\frac{1}{\mathbf{b}}+\frac{1}{\mathbf{c}}+\frac{1}{\mathbf{d}}},\ \mathbf{\ln(OR)} + 1.96\sqrt{\frac{1}{\mathbf{a}}+\frac{1}{\mathbf{b}}+\frac{1}{\mathbf{c}}+\frac{1}{\mathbf{d}}}$$

Relatief risico (RR): 95%-betrouwbaarheidsinterval van ln(RR) is gelijk aan

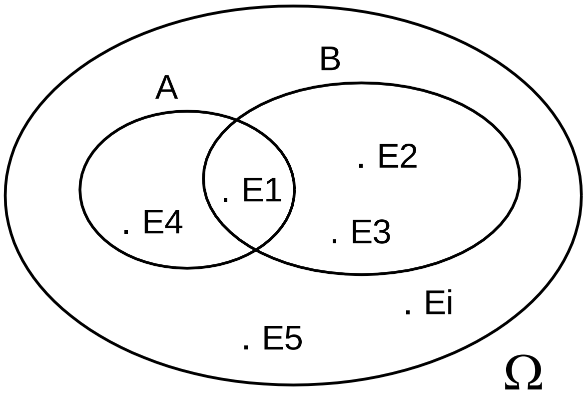
$$\ln(RR) - 1.96\sqrt{\frac{\frac{b}{a}}{a+b} + \frac{\frac{d}{c}}{c+d}},\ \ln(RR) + 1.96\sqrt{\frac{\frac{b}{a}}{a+b} + \frac{\frac{d}{c}}{c+d}}$$

Men dient derhalve de antilog van beide betrouwbaarheidsgrenzen te berekenen en bekomt aldus een asymmetrisch betrouwbaarheidsinterval.

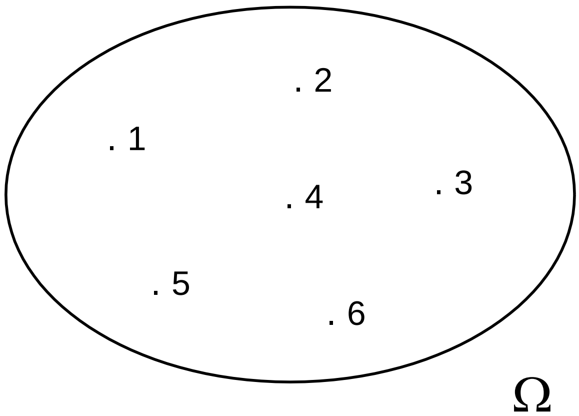
W 1. Probabiliteit: voorstelling met verzamelingen

~ 2.3.5

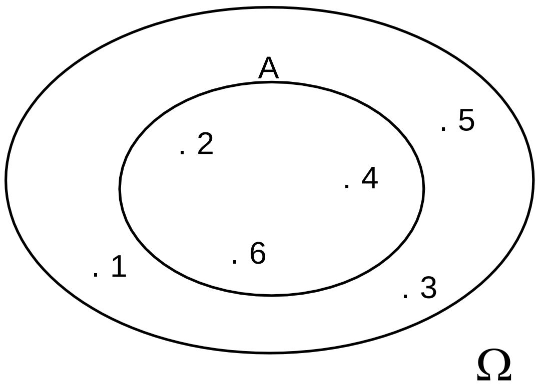
- universum of uitkomstenruimte Ω
- gebeurtenissen A en B
- uitkomsten E_i



Voorbeeld: gooien van een dobbelsteen: $\Omega = \{1,2,3,4,5,6\}$

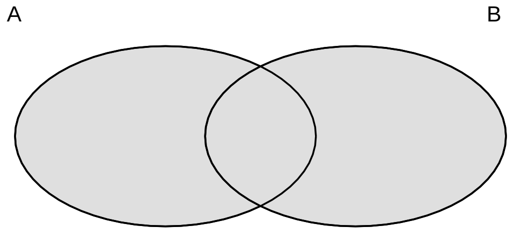


gebeurtenis A: gooien van een even getal



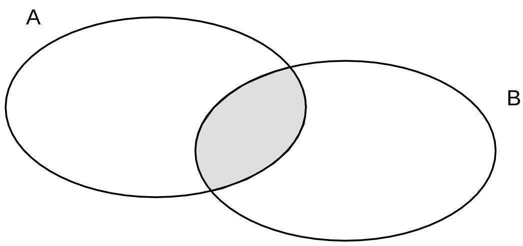
$$P(A) = P\,(\text{worp} = 2) + P\,(\text{worp} = 4) + P\,(\text{worp} = 6)$$
$$= \mathbf{\frac{1}{6}} + \mathbf{\frac{1}{6}} + \mathbf{\frac{1}{6}} = \mathbf{\frac{1}{2}}$$

Samengestelde gebeurtenissen



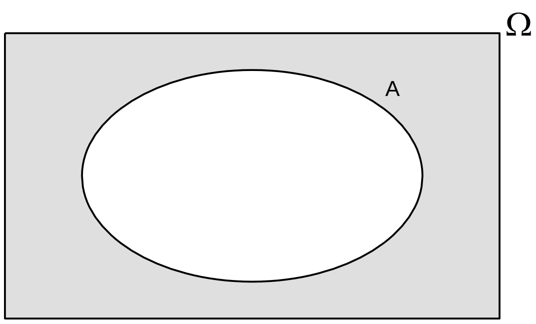
$A \cup B$ unie

$P(A+B) = P(A \text{ of } B) = P(A \cup B)$



$A \cap B$ doorsnede

$P(A.B) = P(A \text{ en } B) = P(A \cap B)$



Complement $\overline{A}$

$P(\overline{A}) = 1 - P(A)$

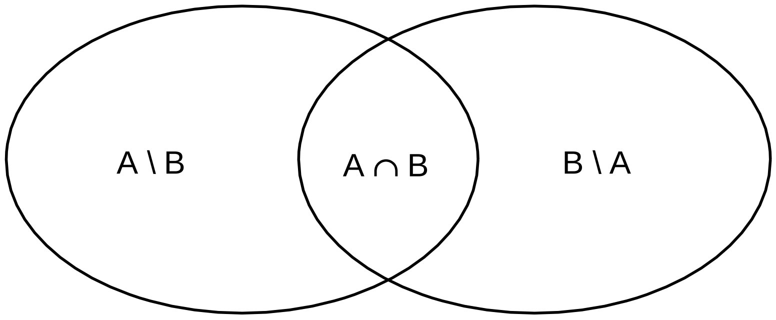
Afleiding van de algemene optellingswet (gelijkheid van BOOLE)

$$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$$

<u>Opmerking</u>: voor $A \cap B = \varnothing$ (disjuncte verzamelingen)
geldt $P(A \cap B) = 0$

en wordt dit de gewone somregel:

$$P(A \cup B) = P(A) + P(B)$$



$$A \cup B = (A \backslash B) \cup (A \cap B) \cup (B \backslash A)$$

$$P(A \cup B) = P((A \backslash B) \cup (A \cap B) \cup (B \backslash A))$$

er geldt dat: $\begin{cases} \mathbf{(A \backslash B)} \cap \mathbf{(A \cap B)} = \varnothing & \\ \mathbf{(A \backslash B)} \cap \mathbf{(B \backslash A)} = \varnothing & \textbf{mutueel} \\ \mathbf{(A \cap B)} \cap \mathbf{(B \backslash A)} = \varnothing & \textbf{exclusief} \end{cases}$

(m.a.w. deze drie deelverzamelingen vormen een **partitie** van $A \cup B$)

Daaruit volgt:

$P(A \cup B) = P(A \backslash B) + P(A \cap B) + P(B \backslash A)$

gewone somregel

uitwerking voor $P(A \backslash B)$ en $P(B \backslash A)$

men heeft: $A = (A \backslash B) \cup (A \cap B)$
en $(A \backslash B) \cap (A \cap B) = \varnothing$

$\rightarrow P(A) = P((A \backslash B) \cup (A \cap B))$
$= P(A \backslash B) + P(A \cap B)$

$\rightarrow P(A \backslash B) = P(A) - P(A \cap B)$

Analoog voor $P(B \backslash A) = P(B) - P(A \cap B)$

Na substitutie:

$$P(A \cup B) = P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B)$$
$$= P(A) + P(B) - P(A \cap B)$$

<u>Voorbeeld</u>:

$P(\overline{A \cup B}) = P(\bar{A} \cap \bar{B})$

linkerlid:

$$
\begin{aligned}
P(\overline{A \cup B}) &= 1 - P(A \cup B) \\
&= 1 - [\mathrm{P(A) + P(B) - P(A\ en\ B)}] \\
&= 1 - \mathrm{P(A) - P(B) + P(A) \times P(B)}
\end{aligned}
$$

rechterlid:

$$
\begin{aligned}
P(\bar{A} \cap \bar{B}) &= P(\bar{A}) \times P(\bar{B}) \\
&= [1 - \mathrm{P(A)}] \times [1 - \mathrm{P(B)}] \\
&= 1 - \mathrm{P(A) - P(B) + P(A) \times P(B)}
\end{aligned}
$$

W 2. Veralgemening vermenigvuldigingswet

~ 2.3.6.1

P(A en B en C) = P(A en B) × P(C | A en B) = P(A) × P(B | A) × P(C | A en B)

P(A en B en C en D) = P(A en B en C) × P(D | A en B en C)
= P(A) × P(B | A) × P(C | A en B) × P(D | A en B en C)

Algemeen:

P(A en B en C en … en Z)
= P(A) × P(B | A) × P(C | A en B) × … × P(Z | A en B en … en Y)

In het geval van onafhankelijke gebeurtenissen:

P(A en B en C en … en Z) = P(A) × P(B) × P(C) × … × P(Y) × P(Z)

W 3. Veralgemening optellingswet

~ 2.3.6.2

P(A of B of C) = P(A of B) + P(C) – P[(A of B) en C]
= P(A) + P(B) – P(A en B) + P(C) – P[(A en C) of (B en C)]
= P(A) + P(B) + P(C) – P(A en B) – [P(A en C) + P(B en C) – P(A en C en B en C)]
= P(A) + P(B) + P(C) – P(A en B) – P(A en C) – P(B en C) + P(A en C en B en C)

In geval van twee-à-twee mutueel exclusieve gebeurtenissen:

P(A of B of C) = P(A) + P(B) + P(C)

W 4. Afleiding regel van Bayes

~ 3.3

$$P(D \times S) = P(D) \times P(S \mid D) = P(S) \times P(D \mid S)$$

$$P(D \mid S) = \frac{P(S \mid D) \times P(D)}{P(S)}$$

$$P(D \mid S) = \frac{P(S \mid D) \times P(D)}{P(S \mid D) \times P(D) + P(S \mid \overline{D}) \times P(\overline{D})}$$

$$= \frac{se \times prev}{se \times prev + aspec \times (1 - prev)}$$

Afleiding van de uitdrukking in de noemer:

$$P(D) + P(\overline{D}) = 1$$

$$P(D \mid S) + P(\overline{D} \mid S) = 1$$

$$\frac{P(D \text{ en } S)}{P(S)} + \frac{P(\overline{D} \text{ en } S)}{P(S)} = 1$$

$$P(D) \times P(S \mid D) + P(\overline{D}) \times P(S \mid \overline{D}) = P(S)$$

Regel van Bayes in zijn odds-likelihood gedaante:

$$P(D \mid S) = \frac{P(S \mid D) \times P(D)}{P(S \mid D) \times P(D) + P(S \mid \overline{D}) \times P(\overline{D})}$$

$$P(\overline{D} \mid S) = \frac{P(S \mid \overline{D}) \times P(\overline{D})}{P(S \mid \overline{D}) \times P(\overline{D}) + P(S \mid D) \times P(D)}$$

$$\frac{P(D \mid S)}{P(\overline{D} \mid S)} = \frac{P(S \mid D) \times P(D)}{P(S \mid \overline{D}) \times P(\overline{D})}$$

$$\underbrace{\frac{P(D \mid S)}{P(\overline{D} \mid S)}}_{\text{posterior odds}} = \underbrace{\frac{P(D)}{P(\overline{D})}}_{\text{prior odds}} \times \underbrace{\frac{P(S \mid D)}{P(S \mid \overline{D})}}_{\text{likelihood ratio}}$$

W 5. Berekening van kwartielen

~ 4.3.2.2

mediaan: $\text{rang}\{\frac{n}{2} + 0.5\}$

eerste kwartiel (Q1): $\text{rang}\{\frac{n}{4} + 0.5\}$

derde kwartiel (Q3): $\text{rang}\{\frac{3n}{4} + 0.5\}$

Afrondingen:

1. rang eindigt op .0 $\Rightarrow$ OK
2. rang eindigt op .25 $\Rightarrow$ afronden naar beneden
3. rang eindigt op .75 $\Rightarrow$ afronden naar boven
4. rang eindigt op .50 $\Rightarrow$ gemiddelde van i^{de} en $(i+1)^{de}$ waarde

Voorbeelden:

rang = 6.25 $\Rightarrow$ 6^e waarde

rang = 6.75 $\Rightarrow$ 7^e waarde

rang = 6.50 $\Rightarrow$ $(6^e + 7^e)/2$

W 6. Begrip vrijheidsgraden

~ 4.3.2.5

Afleiding van $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}x_i - n\bar{x} = \sum_{i=1}^{n}x_i - n\frac{\sum_{i=1}^{n}x_i}{n} = 0$$

De gelijkheid $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ laat toe een willekeurige afwijking $x_i - \bar{x}$ te berekenen, gegeven de n - 1 andere verschillen.

Immers, uit $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

volgt dat $x_j - \bar{x} = -\sum_{i=1(i \neq j)}^{n}(x_i - \bar{x}) \quad (\forall j \in \{1,\ldots,n\})$

W 7. Rekenformule voor de standaarddeviatie

Vertrekkend van formule voor de standaarddeviatie heeft men:

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

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

$$\begin{aligned}
s^2 &= \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right)\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - 2\overline{x}\sum_{i=1}^{n}x_i + n\overline{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - 2n\overline{x}^2 + n\overline{x}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - n\overline{x}^2\right]
\end{aligned}$$

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

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

W 8. Scheefheid - kurtosis

~4.3.3

Scheefheid

Wordt gedefinieerd als: $\mathbf{skewness} = \frac{\sum(x_i - \overline{x})^3}{(n-1)s^3}$

Zar JH in Biostatistical Analysis 3rd ed. (Prentice Hall, 1974) definieert skewness als g_1 in een iets andere gedaante:

$$\mathbf{skewness} = \frac{n\sum(x_i - \overline{x})^3}{(n-1)(n-2)s^3}$$

Kurtosis

Wordt gedefinieerd als: $\textbf{kurtosis} = \dfrac{\sum(x_i - \bar{x})^4}{(n-1)s^4} - 3$

Zar JH definieert kurtosis als g_2 in een aangepaste gedaante:

$$\textbf{kurtosis} = \frac{\frac{n(n+1)}{n-1}\sum(x_i - \bar{x})^4 - 3\left[\sum(x_i - \bar{x})^2\right]^2}{(n-2)(n-3)s^4}$$

W 9. Covariantie - correlatie

~4.5.2

Covariantie drukt de lineaire samenhang uit tussen twee continue toevalsveranderlijken X en Y.

Covariantie wordt gedefinieerd als:

$$\begin{aligned}\sigma_{XY} &= \mathrm{Cov}(X,Y) = \sum_x \sum_y (x - \mu_x)(y - \mu_y)P\left[X = x_i, Y = y_i\right] \\ &= E(X - \mu_x)(Y - \mu_y) \\ &= E(XY) - \mu_x \mu_y\end{aligned}$$

Immers:

$$\begin{aligned}E(X - \mu_x)(Y - \mu_y) &= E(XY - \mu_x Y - \mu_y X + \mu_X \mu_y) \\ &= E(XY) - \mu_x E(Y) - \mu_y E(X) + \mu_X \mu_y \\ &= E(XY) - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y \\ &= E(XY) - \mu_x \mu_y\end{aligned}$$

Correlatie: Correlatie wordt gedefinieerd: $\rho_{XY} = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$

Indien X en Y onafhankelijk zijn dan is de correlatie gelijk nul:

$$\sigma_{XY} = \rho_{XY} = 0$$

Indien X en Y onafhankelijk zijn van elkaar kan de gezamenlijke kansfunctie
P[X = x, Y = y] geschreven worden als het product van de afzonderlijke kansfuncties: P[X = x, Y = y] = P[X = x].P[Y = y]

(vergelijk met P(A×B) = P(A) × P(B)
indien A en B onafhankelijke gebeurtenissen zijn; vanuit
P(A | B) = P(A) en P(B | A) = P(B)

Men heeft dan $\sigma_{XY} = \sum\sum(X - \mu_x)(Y - \mu_y)P[X = x].P[Y = y]$

$$= \sum \underbrace{(X-\mu_x)}_{=0} P[X=x] \sum \underbrace{(Y-\mu_y)}_{=0} P[Y=y]$$

$$= 0$$

W 10. Pearson-correlatie: rekenformule

~ 4.5.3

Overgang van $$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum y_i - \bar{y})^2}}$$

naar de gedaante die handiger is voor praktisch rekenwerk

$$r = \frac{\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{\sqrt{\sum x^2 - \frac{\left(\sum x\right)^2}{n}} \sqrt{\sum y^2 - \frac{\left(\sum y\right)^2}{n}}}$$

naar analogie met W 7

W 11. Spearman rangcorrelatiecoëfficiënt r_s

~ 4.5.4

W 11. 1. Rangsommen

$\sum_{i=1}^{n} i$ (Som van de kwadraten van de rangen)

$$\sum_{i=1}^{n} i = 1 + 2 + 3 + 4 + \ldots + (n-1) + n$$

voor n oneven

$$\sum i = 1 + n + (2 + (n-1)) + (3 + (n-2)) + \ldots + \left(\frac{n-1}{2} + (n - (\frac{n-1}{2} - 1)) + \frac{n+1}{2}\right)$$

$$= \underbrace{(n+1) + (n+1) + (n+1) + \ \ldots \ + (n+1)}_{\frac{n-1}{2} \text{ termen}} + \left(\frac{n+1}{2}\right)$$

$$= (n+1)\left(\frac{n-1}{2}\right) + \frac{n+1}{2} = \frac{1}{2}(n^2 - 1 + n + 1) = \frac{1}{2}n(n+1)$$

voor n even

$$\sum_{i=1}^{n} i = (1+n) + (2+(n-1)) + (3+(n-2)) + \ldots + \left(\frac{n}{2} + \left(n - \left(\frac{n}{2} - 1\right)\right)\right)$$

$$= \underbrace{(n+1) + (n+1) + (n+1) + \ldots\ldots + (n+1)}_{\frac{n}{2}\ \text{termen}} = \frac{1}{2} n(n+1)$$

Daaruit volgt: $\sum_{i=1}^{n} i = \frac{1}{2} n(n+1), \qquad \forall n \in N$

W 11. 2. Som van de kwadraten van de rangen

$$\sum_{i=1}^{n} i^2 = 1^2 + 2^2 + 3^2 + \ldots + (n-1)^2 + n^2$$

$$= 1 + 2.2 + 3.3 + \ldots + (n-1)(n-1) + n.n$$

$$= 1 + (2+2) + (3+3+3) + \ldots + \underbrace{((n-1) + (n-1) + \ldots + (n-1))}_{n-1\ \text{termen}} + \underbrace{n + n + \ldots + n}_{n\ \text{termen}}$$

$$\begin{aligned}
= 1 + 2 + 3 + 4 + \ldots &+ (n-1) + n \\
+ 2 + 3 + 4 + \ldots &+ (n-1) + n \\
+ 3 + 4 + \ldots &+ (n-1) + n \\
+ 4 + \ldots &+ (n-1) + n \\
&\quad . \\
&\quad . \\
&\quad . \\
&+ (n-1) + n \\
&+ n
\end{aligned}$$

$$\begin{aligned}
= \sum_{i=1}^{n} i & \\
&+ \sum i - 1 \\
&+ \sum i - (1+2) \\
&+ \sum i - (1+2+3) \\
&\quad . \\
&\quad . \\
&\quad . \\
&+ \sum i - (1+2+\ldots+(n-2)) \\
&+ \sum i - (1+2+\ldots+(n-1))
\end{aligned}$$

$$= n\sum_{i=1}^{n} i - \sum_{j=1}^{n-1} \sum_{i-1}^{j} I \qquad \left[\sum i = \frac{1}{2} n(n+1)\right]$$

$$\sum_{i=1}^{n} i^2 = \frac{n}{2} n(n+1) - \sum_{j=1}^{n-1} \frac{1}{2} j(j+1)$$

$$= \frac{n^2}{2}(n+1) - \frac{1}{2}\sum_{j=1}^{n-1} j^2 - \frac{1}{2}\sum_{j=1}^{n-1} j$$

$$= \frac{n^2}{2}(n+1) - \frac{1}{2}\left(\sum_{j=1}^{n-1} i^2 - n^2\right) - \frac{1}{2}.\frac{1}{2}.(n-1)n$$

$$= \frac{1}{2}(n^3 + n^2) - \frac{1}{4}(n^2 - n) - \frac{1}{2}\sum_{i=1}^{n} i^2 + \frac{n^2}{2}$$

$$\frac{3}{2}\sum_{i=1}^{n} i^2 = \frac{1}{4}(2n^3 + 2n^2 - n^2 + n + 2n^2)$$

$$\sum_{i=1}^{n} i^2 = \frac{1}{6}(2n^3 + 3n^2 + n)$$

$$= \frac{n}{6}(2n^2 + 3n + 1)$$

$$= \frac{1}{6} n(n+1)(2n+1)$$

$$\boxed{\sum i^2 = \frac{1}{6} n(n+1)(2n+1)}$$

W 11. 3. Gemiddelde en variantie van n rangen

Men heeft:

$$\overline{x} = \frac{\sum x}{n} \qquad\qquad s_x^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

gemiddelde rang: $\overline{x}_{(i)} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{1}{2}n(n+1) = \frac{n+1}{2}$

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

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

$$= \frac{1}{n-1}\left(\frac{1}{6}n(n+1)(2n+1) - \frac{1}{n}\frac{1}{4}n^2(n+1)^2\right)$$

$$= \frac{1}{n-1}\left(\frac{1}{12}(2n(2n^2 + 3n + 1) - 3n(n^2 + 2n + 1)\right)$$

$$= \frac{1}{12(n-1)}\left(4n^3 + 6n^2 + 2n - 3n^3 - 6n^2 - 3n\right)$$

$$= \frac{1}{12(n-1)}(n^3 - n) = \frac{n(n^2-1)}{12(n-1)} = \frac{n(n+1)}{12}$$

gemiddelde van n rangen:	$\frac{n+1}{2}$
variantie van n rangen:	$\frac{n(n+1)}{12}$

W 11. 4. Spearman rangcorrelatiecoëfficiënt r_s

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

(waarbij d_i de absolute waarde van het verschil is tussen de overeenkomstige rang van het (x, y) paar)

De berekening van de Spearman correlatiecoëfficiënt r_s wordt opgeleverd door de Pearson correlatie r te berekenen op de rangen:

$$r = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}\sqrt{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}}$$

en daarnet werd aangetoond dat voor rangsommen geldt:

$$\sum x_{(i)} = \sum y_{(i)} = \frac{1}{2}n(n+1)$$

en $$\sum x_{(i)}^2 = \sum y_{(i)}^2 = \frac{1}{6}n(n+1)(2n+1)$$

Werken we vooreerst de term $\sum x_{(i)} y_{(i)}$ uit (product van de corresponderende rangen (zie ook voorbeelden in toepassingen)

$$d_i = \left|x_{(i)} - y_{(i)}\right|$$

$$d_i^2 = \left(x_{(i)} - y_{(i)}\right)^2 = x_{(i)}^2 - 2x_{(i)}y_{(i)} + y_{(i)}^2$$

$$\sum d_i^2 = 2\sum i^2 - 2\sum x_{(i)} y_{(i)}$$

$$\frac{1}{2}\sum d_i^2 = \frac{1}{6} n(n+1)(2n+1) - \sum x_{(i)} y_{(i)}$$

$$\Rightarrow \sum x_{(i)} y_{(i)} = \frac{1}{6} n(n+1)(2n+1) - \frac{1}{2}\sum d_i^2$$

Na vervanging in de basisformule:

$$r_s = \frac{\frac{1}{6} n(n+1)(2n+1) - \frac{1}{2}\sum d_i^2 - \frac{1}{n}\left(\frac{1}{2} n(n+1)\right)^2}{\frac{1}{6} n(n+1)(2n+1) - \frac{n^2 (n+1)^2}{4n}}$$

$$= \frac{\frac{1}{3} n(n+1)(2n+1) - \frac{1}{2} n(n+1)^2 - \sum d_i^2}{\frac{1}{3} n(n+1)(2n+1) - \frac{1}{2} n(n+1)^2}$$

$$= 1 - \frac{\sum d_i^2}{\frac{1}{6}(4n^3 + 6n^2 + 2n - 3n^3 - 6n^2 - 3n)}$$

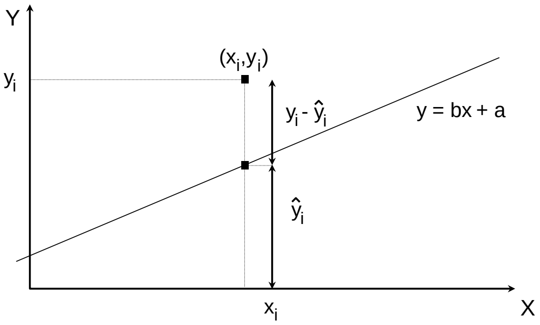
$$= 1 - \frac{6\sum d_i^2}{n^3 - n}$$

$$= 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

W 12. Correlatie - regressie: regressierechte

~4.5.5

De regressierechte als best passende rechte zal beantwoorden aan het wiskundig principe van de ‘best fit’ zodanig dat de totale residuele afwijking (afwijking van de puntenwolk tegenover de rechte) minimaal is.

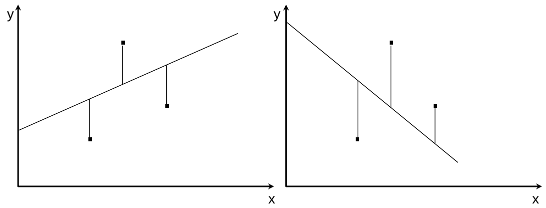


$\hat{y}_i$: beantwoordt aan de rechte
y = bx + a

m.a.w. $(x_i, \hat{y}_i) \in$ regressierechte
$y - \hat{y}_i$ is de residuele afwijking (afstand van een punt tot de regressierechte)

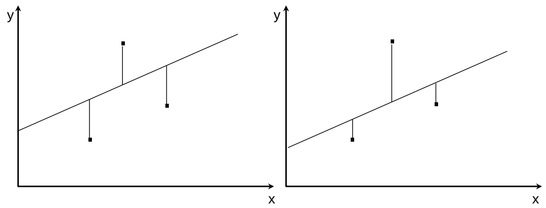
1ste benadering: regressierechte wordt bepaald zodanig dat $\sum_i (y_i - \hat{y}_i)$ minimaal is.

Bij deze benadering worden de positieve verschillen teniet gedaan door de negatieve.
De oplossing is dan alles behalve uniek en zoals uit de afbeelding rechts blijkt zal in dergelijke situatie aan het criterium evengoed voldaan zijn. Visueel is het echter reeds duidelijk dat de oplossing rechts veel minder goed is dan in de figuur links.



2de benadering: de regressierechte zodanig bepaald dat de uitdrukking $\sum_i (y_i - \hat{y}_i)$ minimaal wordt.

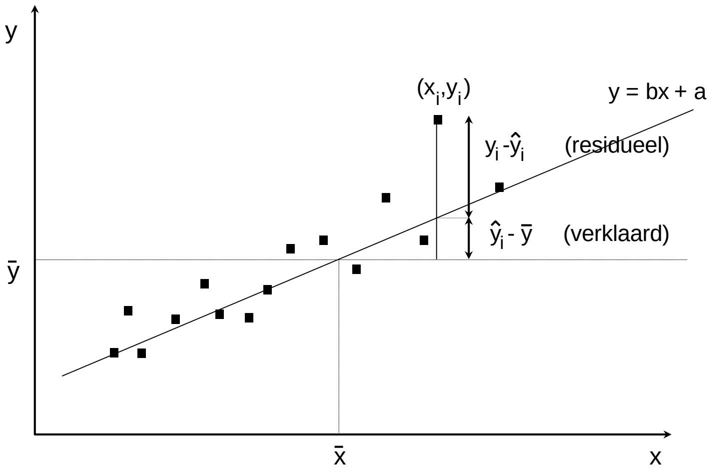
Beide geschetste oplossingen zijn gelijkwaardig ten aanzien van het gestelde criterium, doch de afbeelding rechts houdt geen rekening met de ligging van het middelpunt $(\bar{x}, \bar{y})$.



3de benadering: wanneer men voor de bepaling van de regressierechte oplegt dat de uitdrukking $\sum_i (y_i - \hat{y}_i)^2$ minimaal wordt, dan biedt dit een oplossing voor de tekortkomingen uit de twee vorige benaderingen. Dit criterium wordt het “kleinste kwadraten” principe genoemd.

W 13. Regressie: verklaarde en residuele variantie

~4.5.5



Totale deviatie (regressie van y op x): $\mathbf{y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \hat{y}), \quad \forall i}$

$$\Rightarrow \sum \mathbf{(y_i - \bar{y}) = \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]}$$

We zullen aantonen dat eveneens geldt:

$$\boxed{\sum \mathbf{(y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i) + \sum (\hat{y}_i - \bar{y})^2}}$$

Men heeft:

$$\begin{aligned}
\sum \mathbf{(y_i - \bar{y})^2} &= \sum \mathbf{\left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]^2} \\
&= \sum \mathbf{\left[(y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\right]} \\
&= \sum \mathbf{(y_i - \hat{y}_i)^2} + \sum \mathbf{(\hat{y}_i - \bar{y})^2} + 2\sum \mathbf{(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}
\end{aligned}$$

Er moet dus worden aangetoond dat de productterm

$$\sum \mathbf{(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0}$$

$$\sum \mathbf{(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})} = \sum \mathbf{(y_i - bx_i - a)(bx_i + a - \bar{y})}$$

$$\mathbf{\bar{y} = b\bar{x} + a}$$

$$\begin{aligned}
&= \sum \mathbf{(bx_iy_i + ay_i - \bar{y}y_i - b^2x_i^2 - abx_i + b\bar{y}x_i - abx_i - a^2 + a\bar{y})} \\
&= \mathbf{b\sum x_iy_i + an\bar{y} - n\bar{y}^2 - b^2\sum x_i^2 - 2nab\bar{x} + nb\bar{x}\bar{y} - na^2 + na\bar{y}} \\
&= \mathbf{b\sum x_iy_i - b^2\sum x_i^2 + 2na(\bar{y} - b\bar{x}) - n(\bar{y}^2 - b\bar{x}\bar{y} + a^2)} \\
&= \mathbf{b\sum x_iy_i - b^2\sum x_i^2 + 2na^2 - n\left[(b\bar{x} + a)^2 - b\bar{x}(b\bar{x} + a) + a^2\right]} \\
&= \mathbf{b\sum x_iy_i - b^2\sum x_i^2 + 2na^2 - n\left[b^2\bar{x}^2 + 2ab\bar{x} + a^2 - b^2\bar{x}^2 - ab\bar{x} + a^2\right]} \\
&= \mathbf{b\sum x_iy_i - b^2\sum x_i^2 + 2na^2 - nab\bar{x} - 2na^2} \\
&= \mathbf{b\left(\sum x_iy_i - na\bar{x}\right) - b^2\sum x_i^2} \\
&= \mathbf{b\left(\sum x_iy_i - n\bar{x}(\bar{y} - b\bar{x})\right) - b^2\sum x_i^2 = b\left(\sum x_iy_i - n\bar{x}\bar{y}\right) - b^2\left(\sum x_i^2 - n\bar{x}^2\right)}
\end{aligned}$$

Na substitutie van de uitdrukking voor b bekomt men dat deze uitdrukking gelijk is aan nul.

W 14. Determinatiecoëfficiënt

~4.5.5

$$\mathbf{r^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}\left(= \frac{regressie}{totale}\right)} \qquad [y = bx + a]$$

$$\rightarrow \quad \begin{array}{ll} (x_i, \hat{y}_i) & \in \text{rechte} \\ \text{en } (x, y) & \in \text{rechte} \end{array}$$

$$\sum \mathbf{(\hat{y}_i - \bar{y})^2 = \sum \hat{y}_i^{\,2} - 2\bar{y}\sum \hat{y}_i + n\bar{y}^2}$$

$$
\begin{aligned}
&= \sum (\mathbf{b}x_i + a)^2 - 2\bar{y}\sum (\mathbf{b}x_i + a) + n\bar{y}^2 \\
&= b^2\sum x_i{}^2 + 2ab\sum x_i + na^2 - 2b\bar{y}\sum x_i - 2a\bar{y}n + n\bar{y}^2 \\
&= n(a^2 - 2a\bar{y} + \bar{y}^2) + 2bn\bar{x}(a - \bar{y}) + b^2\sum x_i{}^2 \\
&= n(\bar{y} - a)^2 - 2bn\bar{x}(\bar{y} - a) + b^2\sum x_i{}^2 \\
&= nb^2\bar{x}^2 - 2nb^2\bar{x}^2 + b^2\sum x_i{}^2 \\
&= b^2\left(\sum x_i{}^2 - n\bar{x}^2\right) \qquad \text{(b is regressiecoëfficiënt)} \\
&= \frac{\left(\sum (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum x_i - \bar{x}^2\right)}\left(\sum (x_i - \bar{x}^2)\right)
\end{aligned}
$$

$$r^2 = \frac{\left(\sum (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}$$

en dit is volkomen gelijkwaardig met het kwadraat van de basisformule voor r

W 15. Regressie: kleinste kwadraten principe

~4.5.5

Algemene vergelijking regressierechte y = bx +a

Kleinste kwadraten principe: regressierechte wordt bepaald zodanig dat $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$ minimaal wordt.

Met andere woorden bereken a en b zodanig dat de impliciete functie $F(a,b) = \sum\left(y_i - \hat{y}_i\right)^2 = \sum\left(y_i - bx_i - a\right)^2$ een minimum bereikt.

Opmerking: deze uitdrukking is duidelijk een functie van a en b; voor variërende waarden voor a en b verkrijgt men een ganse waaier van rechten waarbij enkel de ‘best passende’ beoogd wordt.

Het bepalen van een minimum van een functie komt in de wiskundige analyse overeen met te stellen dat de eerste partiële afgeleide gelijk nul wordt. Afgeleid naar a bekomt men:

$$
\begin{aligned}
\frac{\partial}{\partial a}F(a,b) &= \frac{\partial}{\partial a}\sum\left(y_i - bx_i - a\right)^2 \\
&= (-1)2\sum\left(y_i - bx_i - a\right)
\end{aligned}
$$

De voorwaarden voor het bereiken van een minimum:

$$\sum\left(y_i - bx_i - a\right) = 0$$

waaruit $a = \frac{1}{n}\sum\left(y_i - bx_i\right)$ of $\boxed{a = \bar{y} - b\bar{x}}$

Opmerking: hiermee wordt aangetoond dat het middelpunt $(\bar{x}, \bar{y})$ een punt van de regressierechte is.

Partiële afgeleide naar b: $\frac{\partial}{\partial b}F(a,b) = \frac{\partial}{\partial b}\sum\left(y_i - bx_i - a\right)^2$

$$= 2\sum\left(y_i - bx_i - a\right)(-x_i)$$

$$= 2\sum (bx_i^2 - x_i y_i + ax_i)$$
$$= 2\left(b\sum x_i^2 - \sum x_i y_i + a\sum x_i\right)$$
$$= 2\left(b\sum x_i^2 - \sum x_i y_i + an\bar{x}\right)$$

De voorwaarde voor het bereiken van een minimum:

$$b\sum x_i^2 - \sum x_i y_i + an\bar{x} = 0$$
$$b\sum x_i^2 = \sum x_i y_i - (\bar{y} - b\bar{x})n\bar{x}$$
$$= \sum x_i y_i - n\bar{x}\bar{y} + bn\bar{x}^2$$
$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

W 16. Regressiecoëfficiënt b

~4.5.5

$$b_{y.x} = r\frac{s_y}{s_x}$$

$$b_{y.x} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \frac{\sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Analoog voor $b_{x..y} = r\frac{s_x}{s_y}$

Opmerking: daar waar men spreekt van de regressie van y op x (afstanden evenwijdig met de y-as) en van de regressie van x op y (afstanden evenwijdig met de x-as) is de correlatie hier onafhankelijk van. Men kan bv. vaststellen dat het voorkomen van x en y in de basisformule voor r volledig symmetrisch is, hetgeen niet het geval is voor de formule voor b.

Indien $b_{y.x} = r\frac{s_y}{s_x}$ en $b_{x..y} = r\frac{s_x}{s_y}$ dan heeft men onmiddellijk dat (na vermenigvuldiging van de overeenkomstige factoren): $b_{y.x}.b_{x.y} = r^2$

W 17. Eigenschappen van de verwachtingswaarde

~5.1.6.2

E(c) = c   (c = constante) E (cX) = c E (X) E (c + X) = c + E (X) E (a + bX) = a + b E (X)

Deze formules worden aangetoond voor een continue stochastische veranderlijke X; de afleiding voor een discrete stochastische veranderlijke is echter volledig analoog.

$$\mathbf{E(c)} = \int_a^b \mathbf{cf(x)dx} = \mathbf{c}\int_a^b \mathbf{f(x)dx} = \mathbf{c}$$

$$\mathbf{E(cX)} = \int_a^b \mathbf{cxf(x)dx} = \mathbf{c}\int_a^b \mathbf{xf(x)dx} = \mathbf{cE(X)}$$

$$\mathbf{E(c+X)} = \int_a^b \mathbf{(c+x)f(x)dx} = \mathbf{c}\int_a^b \mathbf{f(x)dx} + \int_a^b \mathbf{xf(x)dx} = \mathbf{c} + \mathbf{E(X)}$$

$$\mathbf{E(a+bX)} = \int_a^b \mathbf{(a+bx)f(x)dx} = \int_a^b \mathbf{af(x)dx} + \int_a^b \mathbf{bxf(x)dx} = \mathbf{a} + \mathbf{bE(X)}$$

W 18. Variantie van een stochastische veranderlijke X

~5.1.6.3

$$\begin{aligned}
\sigma^2 &= \mathbf{Var(X)} \\
&= \mathbf{E(X-E(X))^2} = \mathbf{E(X-\mu)^2} && \text{per definitie} \\
&= \mathbf{E(X^2 - 2X\mu + \mu^2)} \\
&= \mathbf{E(X^2)} - 2\mu\mathbf{E(X)} + \mu^2 \\
&= \mathbf{E(X^2)} - 2\mu^2 + \mu^2 = \mathbf{E(X^2)} - \mu^2 \\
&= \mathbf{E(X^2)} - \mathbf{(E(X))^2}
\end{aligned}$$

W 19. Variantie van een discrete veranderlijke X

~5.1.6.3

$$\sigma^2 = \mathbf{Var(X)} = \mathbf{E(X-\mu)^2} \qquad \text{(definitie)}$$

$$\begin{aligned}
&= \sum_i \mathbf{(x_i - \mu)^2 P[X = x_i]} \\
&= \sum_i \mathbf{x_i^2 P[X = x_i]} - 2\mu \sum \mathbf{x_i P[X = x_i]} + \mu^2 \sum_i \mathbf{P[X = x_i]} \\
&= \sum_i \mathbf{x_i^2 P[X = x_i]} - 2\mu^2 + \mu^2 \\
&= \sum_i \mathbf{x_i^2 P[X = x_i]} - \mu^2 \\
&= \sum_i \mathbf{x_i^2 P[X = x_i]} - \mathbf{(E(X))^2}
\end{aligned}$$

W 20. Variantie van een continue veranderlijke X

~5.1.6.3

$$\sigma^2 = \mathbf{Var(X)} = \mathbf{E(X-\mu)^2} \qquad \text{(per definitie)}$$

$$= \int_a^b \mathbf{(x - E(X))^2 f(x)dx}$$

$$= \int_a^b (x-\mu)^2 f(x)dx$$
$$= \int_a^b x^2 f(x)dx - 2\mu\int_a^b xf(x)dx + \mu^2\int_a^b f(x)dx$$
$$= \int_a^b x^2 f(x)dx - 2\mu^2 + \mu^2$$
$$= \int_a^b x^2 f(x)dx - \mu^2$$
$$= \int_a^b x^2 f(x)dx - (E(X))^2$$

W 21. Eigenschappen van variantie

~5.1.6.4

a.
$$\begin{aligned} Var(bX) &= E(bX - E(bX))^2 \\ &= E(b^2X^2 - 2bXE(bX) + (E(bX))^2) \\ &= b^2E(X^2) - 2b^2(E(X))^2 + b^2(E(X))^2 \\ &= b^2(E(X^2) - (E(X))^2) \\ &= b^2 VAR(X) \end{aligned}$$

b.
$$\begin{aligned} Var(a+bX) &= E(a+bX - E(a+bX))^2 \\ &= E((a+bX)^2 - 2(a+bX)E(a+bX) + (E(a+bX))^2) \\ &= E(a^2 + 2abX + b^2X^2 - 2(a+bX)(a+bE(X)) \\ &+ (a+bE(X))^2) \\ &= E(a^2 + 2abX + b^2X^2 - 2a^2 - 2abE(X) - 2abX \\ &- 2b^2XE(X) + a^2 + 2abE(X) + b^2(E(X))^2) \\ &= E(b^2(X^2 - 2XE(X) + (E(X))^2) \\ &= b^2(E(X^2) - 2E(X))^2 + (E(X))^2) \\ &= b^2(E(X^2) - (E(X))^2) \\ &= b^2 Var(X) \end{aligned}$$

W 22. Afleiden van E(X) en VAR(X)

~5.1.6.6

Voorbeeld: $f(x) = \begin{cases} \frac{x}{6} 2 \le x \le 4 \\ 0 x < 2 en x > 4 \end{cases}$

$$E(X) = \int_2^4 xf(x)dx = \int_2^4 x\frac{x}{6}dx = \frac{1}{6}\int_2^4 x^2 dx$$
$$= \frac{1}{6}\frac{x^3}{3}\Big|_2^4 = \frac{1}{6}\left(\frac{64}{3} - \frac{8}{3}\right) = \frac{56}{18} = \frac{28}{9} = 3.1$$

$$Var(X) = E(X - E(X))^2$$

$$= \int_2^4 x^2 f(x)dx - (E(X))^2$$

$$= \int_2^4 \frac{x^3}{6} dx - (3.1)^2$$

$$= \frac{1}{6}\frac{x^4}{4}\Big|_2^4 - (3.1)^2 = \frac{1}{6}\left(\frac{256}{4} - \frac{16}{4}\right) - 9.68 = \frac{240}{24} - 9.68 = 10 - 9.68 = 0.32$$

W 23. Dichtheidsfunctie van de normale distributie

~5.2.2

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (-\infty < x < +\infty)$$

[genoteerd als X ~ **N**(μ, σ^2)]

voldoet aan de voorwaarden van een dichtheidsfunctie:

$$\begin{cases} f(x) \geq 0 \\ \int_{-\infty}^{+\infty} f(x)dx = 1 \end{cases}$$

$$\int_{-\infty}^{+\infty} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

substitutie: $z = \dfrac{x-\mu}{\sigma}$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}z^2} \sigma dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz = 1$$

immers, $\displaystyle\int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$

W 24. Karakteristieken van de normaalverdeling

~5.2.3

De toevalsveranderlijke X is normaalverdeeld

X ~ **N**(μ,σ^2)

dan

E (X) = μ, Var (X) = σ^2

Bewijs

$$\mathbf{E(X)} = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} \mathbf{x}\mathbf{e}^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mu)^2}\mathbf{dx}$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} (\sigma\mathbf{z}+\mu)\mathbf{e}^{-\frac{1}{2}\mathbf{z}^2}\sigma\mathbf{dz} \qquad \text{(substitutie } \mathbf{z} = \frac{\mathbf{x}-\mu}{\sigma}\text{)}$$

$$= \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \mathbf{z}\mathbf{e}^{-\frac{\mathbf{z}^2}{2}}\mathbf{dz} + \frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \mathbf{e}^{-\frac{\mathbf{z}^2}{2}}\mathbf{dz}$$

$$= \mu$$

Immers: $\int_{-\infty}^{+\infty} \mathbf{e}^{-\frac{\mathbf{z}^2}{2}}\mathbf{dz} = \mathbf{1} \text{ en } \int_{-\infty}^{+\infty} \mathbf{z}\mathbf{e}^{-\frac{\mathbf{z}^2}{2}}\mathbf{dz} = \mathbf{0}$

$$\mathbf{Var(X)} = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} (\mathbf{x}-\mu)^2\mathbf{e}^{\frac{-(\mathbf{x}-\mu)^2}{2\sigma^2}}\mathbf{dz}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \mathbf{z}^2\mathbf{e}^{-\frac{\mathbf{z}^2}{2}}\mathbf{dz} \qquad \text{(substitutie } \mathbf{z} = \frac{\mathbf{x}-\mu}{\sigma}\text{)}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}}\left[-\mathbf{z}\mathbf{e}^{-\frac{\mathbf{z}^2}{2}}\Big|_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} \mathbf{e}^{-\frac{\mathbf{z}^2}{2}}\mathbf{dz}\right] \qquad \text{(partiële integratie)}$$

$$= \frac{\sigma^2}{\sqrt{2\pi}}\left[0 + \sqrt{2\pi}\right]$$

$$= \sigma^2$$

Zie ook standaardnormale verdeling;

E(Z) = 0
Var(Z) = 1

W 25. Binomium van Newton

~5.3.4

Binomiale distributie:
voor r = 0, 1, ..., n geldt:

$$\mathbf{P(X = r) = P(r)} = \binom{\mathbf{n}}{\mathbf{r}}\pi^{\mathbf{r}}(\mathbf{1}-\pi)^{\mathbf{n-r}}$$

Men heeft: P(0) + P(1) + ... + P(n-1) + P(n) =1

(de som van alle probabiliteiten is steeds gelijk aan 1)

of $\sum_{\mathbf{r=0}}^{\mathbf{n}}\mathbf{P(r)} = \mathbf{1}$

met andere woorden $\sum_{r=0}^{n}\binom{n}{r}\pi^r(1-\pi)^{n-r} = 1$

Dit is een afleiding van het binomium van Newton

$$\boxed{\sum_{i=0}^{n}\binom{n}{i}a^{n-i}b^i = (a+b)^n}$$

W 26. Karakteristieken van de binomiale verdeling

~5.3.4

$$E(r) = \sum_{r=0}^{n} r\binom{n}{r}\pi^r(1-\pi)^{n-r} = \sum_{r=1}^{n} r\binom{n}{r}\pi^r(1-\pi)^{n-r}$$

$$= \sum_{r=1}^{n} n\binom{n-1}{r-1}\pi^r(1-\pi)^{n-r}$$

[want voor r ≥ 1: $r\binom{n}{r} = n\binom{n-1}{r-1}$]

$$= n\pi\sum_{r=1}^{n}\binom{n-1}{r-1}\pi^{r-1}(1-\pi)^{n-r}$$

$$= n\pi\sum_{m=0}^{n-1}\binom{n-1}{m}\pi^{m}(1-\pi)^{n-1-m}$$

(m = r – 1, dan voor r gaat naar n,

gaat m naar n - 1)

$$= n\pi(\pi+(1-\pi))^{n-1}$$

(Newton, W 25)

$$= n\pi$$

$$Var(r) = \sum_{r=0}^{n} r^2\binom{n}{r}\pi^r(1-\pi)^{n-r} - (n\pi)^2$$

$$= \sum_{r=1}^{n} r^2\binom{n}{r}\pi^r(1-\pi)^{n-r} - (n\pi)^2$$

$$= \sum_{r=1}^{n} r(r-1)\binom{n}{r}\pi^r(1-\pi)^{n-r} + n\pi - (n\pi)^2$$

$$= \sum_{r=2}^{n} r(r-1)\binom{n}{r}\pi^r(1-\pi)^{n-r} + n\pi - (n\pi)^2$$

$$= n(n-1)\sum_{r=2}^{n}\binom{n-2}{r-2}\pi^r(1-\pi)^{n-r} + n\pi - (n\pi)^2$$

immers voor r ≥ 2: $r(r-1)\binom{n}{r} = n(n-1)\binom{n-2}{r-2}$

$$= n(n-1)\pi^2\sum_{r=2}^{n}\binom{n-2}{r-2}\pi^{r-2}(1-\pi)^{n-r} + n\pi - (n\pi)^2$$

$$= n(n-1)\pi^2 + n\pi - (n\pi)^2$$

$$= -n\pi^2 + n\pi$$

$$= n\pi(1-\pi)$$

W 27. Poisson-distributie afgeleid uit de binomiale distributie

~ 5.4.1

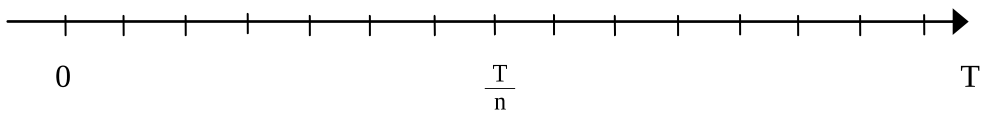
Voor de binomiale distributies geldt voor r = 0,1,...., n

$$P(X = r) = P(r) = \binom{n}{r}\pi^r(1-\pi)^{n-r}$$
$$= \frac{n!}{r!(n-r)!}\pi^r(1-\pi)^{n-r}$$

<u>Poisson-distributies</u>:

λ = gemiddeld optreden van een zeldzame gebeurtenis binnen een beschouwend tijdsinterval.

M.a.w. aantal optredende gebeurtenissen in tijdsinterval t is gelijk aan λ t



$$\mu = \lambda T$$
$$\lambda\frac{T}{n} = \frac{\mu}{n}$$

binomiaal: $r \rightarrow \binom{n}{r}\pi^r(1-\pi)^{n-r}$

Poisson: $x \rightarrow \binom{n}{x}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x}$

Binomiaal:

$$P[X = x] = \binom{n}{x}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{n}\right)^{n-x}$$
$$= \frac{n(n-1)(n-2)....(n-x+1)}{x!}\left(\frac{\mu}{n}\right)^x\left(1-\frac{\mu}{x}\right)^{n-x}$$

Voor n → ∞ en x een zeldzame gebeurtenis:

$$n(n-1)(n-2)\ ...\ (n-x+1) \cong n^x$$
$$\left(1-\frac{\mu}{n}\right)^{n-x} \cong \left(1-\frac{\mu}{n}\right)^n$$

Men heeft: $\lim_{n\to\infty}\left(1-\frac{\mu}{n}\right)^n = e^{-\mu}$ (zie hieronder, het getal e)

M.a.w. P[X = x] kan na overgang van de binomiaal naar de Poisson-verdeling geschreven worden als:

$$P[X=x]=\lim_{n\to\infty}\binom{n}{x}\left(\frac{\mu}{n}\right)^{x}\left(1-\frac{\mu}{n}\right)^{n-x}$$
$$=\frac{n^{x}}{x!}\left(\frac{\mu}{n}\right)^{x}e^{-\mu}=\frac{\mu^{x}}{x!}e^{-\mu}$$

Het getal e

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^{n}=e^{x}$$

Vooreerst: $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n}=e$

Volgens binomium van Newton (W 25):

$$\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n}=\lim_{n\to\infty}\sum_{i-0}^{n}\binom{n}{i}1^{n-i}\left(\frac{1}{n}\right)^{i}$$

$$=\lim_{n\to\infty}\left[\binom{n}{0}+\binom{n}{1}\frac{1}{n}+\binom{n}{2}\frac{1}{n^{2}}+\ldots\ldots+\binom{n}{n}\frac{1}{n^{n}}\right]$$

$$=\lim_{n\to\infty}\left[\frac{n!}{0!n!}+\frac{n!}{(n-1)!1!}\frac{1}{n}+\frac{n!}{2!(n-2)!}\frac{1}{n^{2}}+\ldots+\frac{n!}{n!0!}\frac{1}{n^{n}}\right]$$

$$=\lim_{n\to\infty}\left[1+n\frac{1}{n}+\frac{n(n-1)}{2n^{2}}+\frac{n(n-1)(n-2)}{6n^{3}}+\ldots+\frac{1}{n^{n}}\right]$$

$$=\lim_{n\to\infty}\left[1+1+\frac{n-1}{2n}+\frac{(n-1)(n-2)}{6n^{2}}+\ldots+\frac{1}{n^{n}}\right]$$

$$=2+\frac{1}{2}+\frac{1}{6}+\ldots+0$$

$$=2+\frac{1}{2!}+\frac{1}{3!}+\frac{1}{4!}+\ldots+$$

Deze reeks convergeert naar het getal 2.71828 gedefinieerd als het getal e

<u>Algemeen</u>: $\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^{n}=e^{x}$

Na analoge uitwerking bekomt men:

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^{n}=1+x+\frac{x^{2}}{2}+\frac{x^{3}}{6}+\ldots$$

$$= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$$

$$= \sum_{i=0}^{\infty} \frac{x^i}{i} = e^x$$

(MacLaurin reeksontwikkeling)

W 28. Karakteristieken van de Poisson-verdeling

~ 5.4.2

$$E(X) = \sum_{i=0}^{\infty} xP(X = x) = \sum x \frac{\mu^x e^{-\mu}}{x!}$$

$$= \mu \sum \frac{\mu^{x-1} e^{-\mu}}{(x-1)!} = \mu e^{-\mu} \sum_{i=0}^{\infty} \frac{\mu^i}{i!}$$

$$= \mu e^{-\mu} e^{\mu} = \mu$$

$$\boxed{E(X) = \mu}$$

$$Var(X) = E(X^2) - (E(X))^2$$

er geldt: $E(X(X-1)) = E(X^2) - E(X)$

met andere woorden

$$Var(X) = E(X(X-1)) + E(X) - (E(X))^2$$

$$= E(X(X-1)) + \mu - \mu^2$$

$$E(X(X-1)) = \sum x(x-1)P(X = x)$$

$$= \sum x(x-1) \frac{\mu^x e^{-\mu}}{x!}$$

$$= \mu^2 e^{-\mu} \sum \frac{\mu^{x-2}}{(x-2)!}$$

$$= \mu^2 e^{-\mu} e^{\mu} = \mu^2$$

M.a.w.: $Var(X) = \mu^2 + \mu - \mu^2 = \mu$

$$\boxed{Var(X) = \mu}$$

W 29. Lineaire combinaties van twee toevalsveranderlijken

~ 6.2.1

Beschouw de som 'Z' van twee toevalsveranderlijken X en Y

$$Z = X + Y$$

(Voorbeeld: X = inkomen van de echtgenoot)
Y = inkomen van de echtgenote, Z = gezamenlijk inkomen)

Gemiddelde:

$$\boxed{E(X+Y) = E(X) + E(Y)}$$

Immers $\mathbf{E(X+Y)} = \sum_{x}\sum_{y}\mathbf{(X+Y)P(x,y)}$

$$= \sum_{x}\sum_{y}\mathbf{XP(x,y)} + \sum_{x}\sum_{y}\mathbf{YP(x,y)}$$

(marginale distributies)

$$= \sum_{x}\mathbf{X}\sum_{y}\mathbf{P(x,y)} + \sum_{y}\mathbf{Y}\sum_{x}\mathbf{P(x,y)}$$

$$= \sum_{x}\mathbf{XP(x)} + \sum_{y}\mathbf{YP(y)} = \mathbf{E(X)} + \mathbf{E(Y)}$$

Variantie:

$$\boxed{\text{Var }(X+Y) = \text{Var } X + \text{Var } Y + 2 \text{ Cov }(X,Y)}$$

$$\text{Var }(X + Y) = \text{Var } Z = E(Z - \mu_s)^2$$
$$= E((X + Y) - (\mu_x + \mu_y))^2$$

(De zonet aangetoonde uitdrukking E(X + Y) = E(X) + E(Y) kan ook worden geschreven $E(Z) = \mu_Z = \mu_x + \mu_y$)

$$= E[(X-\mu_x) + (Y-\mu_y)]^2$$

$$= E[(X-\mu_x)^2 + 2(X-\mu_x)(Y-\mu_y)+(Y-\mu_y)^2]$$

$$= E(X-\mu_x)^2+2E(X-\mu_x)(Y-\mu_y)+E(Y-\mu_y)^2$$

$$= \text{VarX} + 2\text{Cov }(X,Y) + \text{Var } Y$$

Indien X en Y onafhankelijk zijn van elkaar dan heeft men dat de covariantie nul is.
M.a.w.

$$\boxed{\text{Var }(X+Y) = \text{Var }(X) + \text{Var }(Y)}$$

Karakteristieken van het steekproefgemiddelde

$$E(\overline{X}) = E\left(\frac{\sum x}{n}\right) = \frac{1}{n}E\left(\sum X\right) = \frac{1}{n}\sum E(X) = \frac{n}{n}E(X) = \mu$$

$$\boxed{E(\overline{X}) = \mu}$$

$$Var(\overline{X}) = Var\left(\frac{\sum X}{n}\right) = Var\left(\frac{X_1 + X_2 + X_3 + ... + X_n}{n}\right)$$

$$= \frac{1}{n^2} Var(X_1 + X_2 + ... + X_n)$$

De steekproefwaarden (simple random sample) zijn onafhankelijk van elkaar, m.a.w.

$$Var(\overline{X}) = \frac{1}{n^2}[Var(X_1) + Var(X_2) + ... + Var(X_n)]$$

(alle X_i hebben zelfde kansverdelingsfunctie)

$$= \frac{n}{n^2}\sigma^2$$

$$\boxed{Var(\overline{X}) = \frac{\sigma^2}{n}}$$

$$\sigma(\overline{X}) = SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

W 30. Zuivere schatters

~ 6.4.1

a. $\overline{X}$ **is een zuivere schatter voor μ**

b. $s^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$ **= een zuivere schatter voor σ²**

Zuivere schatters ⇔ $E(\hat{\theta}) = \theta$ **(**$\hat{\theta}$ **is schatter van de parameter θ)**

a. $\overline{X} = \frac{1}{n}\sum X_i$ is een zuivere schatter voor μ omdat $E(\overline{X}) = \mu$

b. In de formule voor de variantie met deling door (n-1) werd reeds aangegeven dat dit een betere schatter zou opleveren voor de populatie variantie, zoals inderdaad mag blijken uit volgende afleiding:

$$E(s^2) = E\left[\frac{\sum (X_i - \overline{X})^2}{n-1}\right] = \frac{1}{n-1}E\left[\sum (X_i - \overline{X})^2\right]$$

$$= \frac{1}{n-1}E\left[\sum ((X_i - \mu) - (\overline{X} - \mu))^2\right]$$

$$= \frac{1}{n-1} E\left[\sum((X_i-\mu)^2 - 2(X_i-\mu)(\overline{X}-\mu) + (\overline{X}-\mu)^2)\right]$$
$$= \frac{1}{n-1} E\left[\sum(X_i-\mu)^2 - 2\sum(X_i-\mu)(\overline{X}-\mu) + \sum(\overline{X}-\mu)^2\right]$$
$$= \frac{1}{n-1} \left[E[\sum(X_i-\mu)^2] - 2\left[E(\overline{X}-\mu)\sum(X_i-\mu)\right] + nE(\overline{X}-\mu)^2\right]$$
$$= \frac{1}{n-1} \left[\sum E(X_i-\mu)^2 - 2nE(\overline{X}-\mu)^2 + nE(\overline{X}-\mu)^2\right]$$
$$= \frac{1}{n-1} (n\sigma^2 - nVar(\overline{X}))$$
$$= \frac{n}{n-1} (\sigma^2 - \frac{\sigma^2}{n}) = \frac{n}{n-1} \frac{(n-1)\sigma^2}{n}$$
$$= \sigma^2$$

W 31. Betrouwbaarheidsintervallen

~ 6.4.2.1

$$P\left[-1.96 \le \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le 1.96\right] = 0.95$$
$$P\left[-1.96\frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu < 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$
$$P\left[\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

of uit de voorlaatste uitdrukking mits afzondering van μ:

$$P\left[-\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le -\mu \le 1.96\frac{\sigma}{\sqrt{n}} - \overline{X}\right] = 0.95$$

mits omkering van teken en van de ongelijkheid

$$P\left[\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

als 95%-betrouwbaarheidsinterval voor μ

W 32. Waarschijnlijkheidsverdeling van verschillen

~ 6.4.12

Uit W 29 volgt onmiddellijk:

$$E(X_1 - X_2) = E(X_1) - E(X_2)$$

Afleiding van: $Var(X_1 - X_2) = Var(X_1) + Var(X_2)$

Stel $X_1 - X_2 = D$ dan

$$
\begin{aligned}
\mathbf{Var}(\mathbf{X}_1 - \mathbf{X}_2) = \mathbf{VarD} &= \mathbf{E}(\mathbf{D} - \mu_D)^2 \\
&= \mathbf{E}((\mathbf{X}_1 - \mathbf{X}_2) - (\mu_1 - \mu_2))^2 \\
&= \mathbf{E}((\mathbf{X}_1 - \mu_1) - (\mathbf{X}_2 - \mu_2))^2 \\
&= \mathbf{E}((\mathbf{X}_1 - \mu_1)^2 - 2(\mathbf{X}_1 - \mu_1)(\mathbf{X}_2 - \mu_2) + (\mathbf{X}_2 - \mu_2)^2) \\
&= \mathbf{E}(\mathbf{X}_1 - \mu_1)^2 - 2\mathbf{E}(\mathbf{X}_1 - \mu_2)(\mathbf{X}_2 - \mu_2) + \mathbf{E}(\mathbf{X}_2 - \mu_2)^2 \\
&= \mathbf{Var}(\mathbf{X}_1) - 2\mathbf{Cov}(\mathbf{X}_1, \mathbf{X}_2) + \mathbf{Var}(\mathbf{X}_2)
\end{aligned}
$$

Aangezien X_1 en X_2 twee onafhankelijke toevallige steekproeven zijn is de covariantie $Cov(X_1, X_2) = 0$

$$\Rightarrow Var(X_1 - X_2) = Var(X_1) + Var(X_2)$$

$E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu$
$Var(X_1 - X_2) = \sigma_1^2 + \sigma_2^2$

W 33. Verschil tussen twee gemiddelden

~ 6.4.12

Uit W 29 volgt onmiddellijk:

$$\mathbf{E}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = \mathbf{E}(\overline{\mathbf{X}}_1) - \mathbf{E}(\overline{\mathbf{X}}_2) = \mu_1 - \mu_2$$

en

$$\mathbf{Var}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = \mathbf{Var}(\overline{\mathbf{X}}_1) + \mathbf{Var}(\overline{\mathbf{X}}_2)$$
$$= \frac{\sigma_1^2}{\mathbf{n}_1} + \frac{\sigma_2^2}{\mathbf{n}_2}$$

$$\textbf{steekproeven} \quad \mathbf{n}_1, \overline{\mathbf{X}}_1, \mathbf{s}_1 \qquad \mathbf{n}_2, \overline{\mathbf{X}}_2, \mathbf{s}_2$$

$$\textbf{wanneer} \quad \sigma_1^2 \cong \sigma_2^2 \ (= \sigma^2)$$

$$\mathbf{s}_1^2 = \frac{\sum_{i=1}^{n_1} (\mathbf{x}_i - \overline{\mathbf{x}}_1)^2}{\mathbf{n}_1 - 1}; \quad \mathbf{s}_2^2 = \frac{\sum_{i=1}^{n_2} (\mathbf{x}_i - \overline{\mathbf{x}}_2)^2}{\mathbf{n}_2 - 1}$$

$$\textbf{gezamelijk} \Rightarrow \textbf{alles in 1 pot}$$

$$\mathbf{s}^2 = \frac{\sum_{i=1}^{n_1} (\mathbf{x}_i - \overline{\mathbf{x}}_1)^2 + \sum_{i=1}^{n_2} (\mathbf{x}_i - \overline{\mathbf{x}}_2)^2}{(\mathbf{n}_1 - 1) + (\mathbf{n}_2 - 1)}$$

$$\mathbf{s} = \sqrt{\frac{(\mathbf{n}_1 - 1)\mathbf{s}_1^2 + (\mathbf{n}_2 - 1)\mathbf{s}_2^2}{\mathbf{n}_1 + \mathbf{n}_2 - 2}}$$

$$\mathbf{SE}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = \mathbf{s}\sqrt{\frac{1}{\mathbf{n}_1} + \frac{1}{\mathbf{n}_2}}$$

$$E(X_1 - X_2) = E(X_1) - E(X_2)$$
$$Var(X_1 - X_2) = Var(X_1) + Var(X_2)$$
$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2)$$
$$Var(\overline{X}_1 - \overline{X}_2) = Var(\overline{X}_1) + Var(\overline{X}_2)$$
$$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$
$$SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
wanneer $\sigma_1^2 \cong \sigma_2^2 \ (= \sigma^2)$
$$SE(\overline{X}_1 - \overline{X}_2) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

W 34. Verschil tussen twee proporties

~ 6.4.13

$$E(p_1 - p_2) = E(p_1) - E(p_2) = \pi_1 - \pi_2$$

en
$$Var(p_1 - p_2) = Var(p_1) + Var(p_2)$$
$$= \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

$$\pi_A - \pi_B : p_A - p_B$$
$$p_A - p_B :$$
$$SE(p_A - p_B) = \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$$

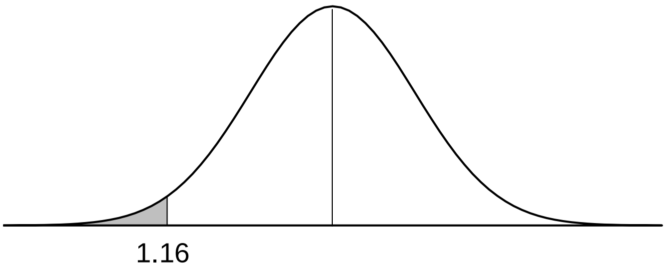
W 35. α en β-kansen

~ 7.6.3.4

Zie stap 3 in het gedachte-experiment: indien gegeven zou zijn dat de gemiddelde overlevingsduur 30 maanden bedraagt, dan nog kan men eventueel een steekproef bekomen die steekproefgemiddelden oplevert die lager zijn dan 27.1.

De kans hiertoe is niets anders dan β: de kans om de nulhypothese niet te verwerpen indien H_0 nochtans ongeldig is.

Deze kans wordt afgeleid uit: $\left(\dfrac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) : \dfrac{27.1 - 30}{\frac{25}{\sqrt{100}}} = -1.16$



De P waarde overeenkomstig met z = -1.16 is gelijk aan P = 0.123

W 36. Tabel poweranalyse – afleiding K-waarden

~ 7.8.1

Vereiste steekproefaantallen in het geval van de vergelijking van twee populatiegemiddelden:

$$\mathbf{n} > \frac{\mathbf{2Ks^2}}{\Delta^2}$$

Afleiding van de tabel: $\mathbf{K} = (\mathbf{z}_{1-\frac{\alpha}{2}} + \mathbf{z}_{1-\beta})^2$

		significantieniveau α	
(tweezijdig toetsen)		**0.05**	0.01
	0.95	$(1.959964+1.644854)^2$	$(2.575829+1.644854)^2$
1 - β (power)	0.90	$(1.959964+1.281552)^2$	$(2.575829+1.281552)^2$
	0.80	$(1.959964+0.8416212)^2$	$(2.575829+0.8416212)^2$

W 37. Verkorte formule voor 2×2-tabellen

~ 8.3.1.3

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Algemene voorstelling van een 2 × 2-tabel:

		(uitkomst of groep, …)		
		A	B	
criterium	I	a	b	a + b
	II	c	d	c + d
		a + c	b + d	N = a + b + c + d

Verwachte aantallen:

		(uitkomst of groep,…)		
		A	B	
	I	$\frac{(a+b)(a+c)}{N}$	$\frac{(a+b)(b+d)}{N}$	a + b
criterium	II	$\frac{(c+d)(a+c)}{N}$	$\frac{(c+d)(b+d)}{N}$	c + d
		a + c	b + d	N = a + b + c + d

$$\chi^2 = \frac{\left(a - \frac{(a+b)(a+c)}{N}\right)^2}{\frac{(a+b)(a+c)}{N}} + \frac{\left(b - \frac{(a+b)(b+d)}{N}\right)^2}{\frac{(a+b)(b+d)}{N}}$$

$$+ \frac{\left(c - \frac{(c+d)(a+c)}{N}\right)^2}{\frac{(c+d)(a+c)}{N}} + \frac{\left(d - \frac{(c+d)(b+d)}{N}\right)^2}{\frac{(c+d)(b+d)}{N}}$$

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$