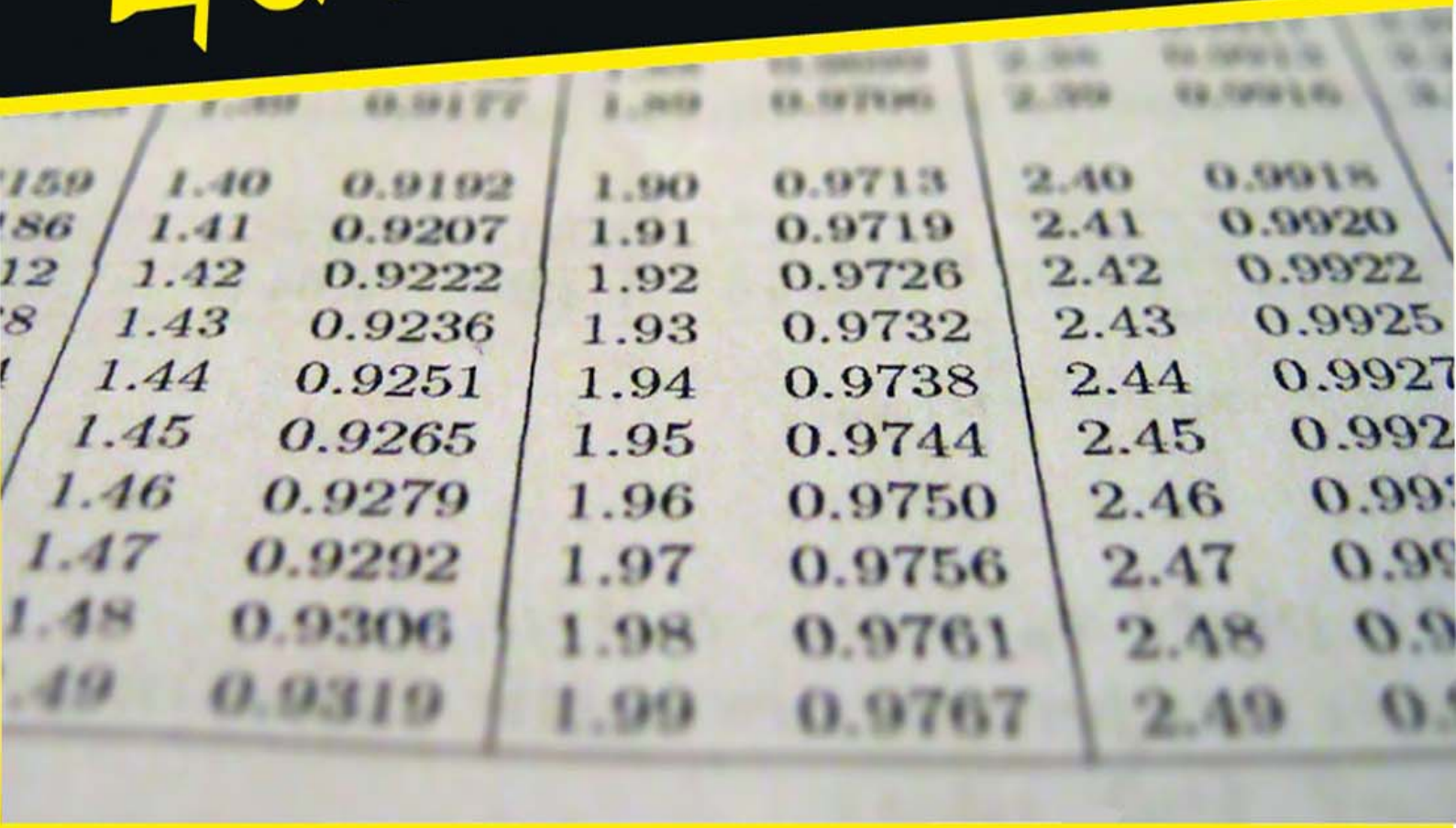
2e bach EW
2006-2007

Statistiek VOOR ECONOMISTEN



Vincent Jacobs

Beste EWers,

Omdat prof. Lauwers duidelijk niet graag heeft dat zijn studenten statistiek leren van hem heb ik voor mijn tweede zit al mijn notities van tijdens de lessen op getypt en georganiseerd.

De volgende pagina's zijn dus gebaseerd op mijn notities en die van Jasper van ‘Statistiek voor Economisten’ van 2e bachelor EW van academiejaar 2006-2007. Het staat vol met fouten, dus dit is in ieder geval geen vervanging voor het handboek en eigen nota's, maar als de prof opnieuw dezelfde lessen geeft dan zou het wel nuttig kunnen zijn.

Omdat de prof weigert om uitgewerkte oplossingen van de oefenzittingen te geven, heb ik de partiële oplossingen van Toledo en mijn notities van tijdens de oefenzittingen samengevoegd en op getypt aan het einde van dit document. Bij iedere oefenzitting ontbreken er sommige vragen (die waar hij geen antwoord voor gaf en ook niet gezien waren in de les) want in dit geval is geen antwoord is beter dan een fout antwoord. Van de gegeven oplossingen ben ik vrij zeker dat de meerderheid juist zijn.

Nog veel succes,

Vincent Jacobs

vjacobs@gmail.com

Inhoud

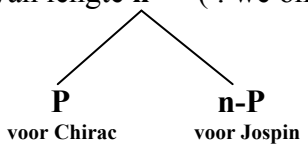
Inleidend Voorbeeld

mei 1995 – Franse presidentsverkiezing - Chirac VS Jospin

om de uitslag te "voorspellen" via een verkiezingspoll zijn er twee methodes:

- hele "populatie" (alle stemgerechtigden) ondervragen
- een "steekproef" ondervragen en de resultaten veralgemenen

"steekproef" van lengte **n** (: we ondervragen 'n' aantal stemgerechtigden)



$$p = \frac{P}{n} \qquad p^c = \frac{n-P}{n} = 1 - p$$

P = aantal stemmen in de steekproef voor Chirac
p = proportie van de stemmen in de steekproef voor Chirac
p^c = compliment van p : proportie van de stemmen in de steekproef voor Jospin = $(1-p)$

besluit omtrent de "echte" proportie van Chirac stemmers = π

Ω = de populatie : de verzameling van alle stemgerechtigden
C = groep van de populatie die op Chirac stemt : $\{ \omega \in \Omega \mid \omega$ stemt op Chirac$\}$
π = fractie van de populatie die op Chirac stemt

$$\pi = \frac{|C|}{|\Omega|}$$ (absolute waarde = aantal elementen in de verzameling)

→ $\boxed{\pi \in [p-\varepsilon, p+\varepsilon]}$
= π behoort tot het interval $(p-\varepsilon, p+\varepsilon)$
= conclusie omtrent π op basis van de steekproef

INDUCTIEVE STATISTIEK
↕
DEDUCTIEVE STATISTIEK

ε = het fout

$$\varepsilon = 1{,}96 \times \sqrt{\frac{p(1-p)}{n}}$$

95% betrouwbaarheids interval (vanuit de tabel van de normale verdeling)

Voorbeeld

een steekproef bevat 1500 mensen : n = 1500
aantal stemmen voor Chirac: P = 789
aantal stemmen voor Jospin: n-P = 711
proporties: p = 52,6% , $p^c = (1-p) =$ 47,4%

We gebruiken de formule voor het fout van hierboven ($\varepsilon = 1{,}96 \text{ x } \sqrt{\frac{p(1-p)}{n}}$) om het fout van deze voorbeeld steekproef te berekenen:

$$\varepsilon = 1{,}96 \text{ x } \sqrt{\frac{0{,}526 \text{ x } 0{,}474}{1500}} = 0{,}025$$

Wij gebruiken dit resultaat en $\pi \in \left[p-\varepsilon, p+\varepsilon \right]$ om het betrouwbaarheidsinterval te zoeken:

$$0{,}526 - 0{,}025 \le \pi \le 0{,}526 + 0{,}025$$
$$0{,}501 \le \pi \le 0{,}551$$

***** Met 95% zekerheid kunnen we concluderen dat Chirac de verkiezingen wint. Omdat π (de fractie van de populatie die op Chirac stemt) met 95% zekerheid tussen 50,1% en 55,1% ligt.

***** Er is geen 100% zekerheid dat Chirac wint (maar 95%) omdat er de mogelijkheid bestaat om een "slechte" steekproef te trekken maar dit is niet waarschijnlijk (KANSTHEORIE).

* $\varepsilon = 1{,}96 \text{ x } \sqrt{\frac{p(1-p)}{n}}$

Als de lengte van de steekproef groter wordt (dus als n ↗) dan is er meer informatie dus wordt het fout kleiner (ε ↘).

een voorbeeld van dit feit:

n=1500 → ε =0,025
n'=15000 → $\varepsilon' = \frac{0{,}025}{\sqrt{10}} = 0{,}008$

n x 10 → ε x $\frac{1}{\sqrt{10}}$ (afnemende schaal opbrengsten)
$n \to \infty$ (de volledige populatie) → ε =0 : geen fout

***** betrouwbaarheid (vanuit de tabel)
95% → 1,96
99% → 2,58

indien we 99% betrouwbaarheid gebruiken:

$$\varepsilon_{99\%} = 2{,}58 \text{ x } \sqrt{\frac{0{,}526 \text{ x } 0{,}474}{1500}} = 0{,}033$$

$$0{,}526 - 0{,}033 \le \pi \le 0{,}526 + 0{,}033$$
$$0{,}493 \le \pi \le 0{,}559$$

Statistiek → inductief (voegt conclusies toe)

↓

Ervaring
Observatie
Informatie verzamelen
} veralgemenen, besluit trekken

Problemen: onvoldoende informatie
schattingen
betrouwbaarheidsniveau

<u>Doel van de cursus:</u>

Deel I - Beschrijvende Statistiek
:ontwerpen van plannen om informatie te verzamelen en om deze samenvattend weer te geven

Deel II - Probabiliteitstheorie of kansrekenen

Deel III - Statistische Besluitvorming
:ontwerpen van procedures om beslissingen te treffen op grond van onvolledige informatie.

Deel IV – Relatie-onderzoek

→Economie als empirische wetenschap

DEEL I
BESCHRIJVENDE STATISTIEK

1. Inleidende begrippen (HB p6)

1.1 Wat is statistiek? (HB p8)

Populatie : Ω (vb. alle Franse kiesgerechtigden)
Kenmerk (hoofdletters) : X : $\Omega \to \mathbb{R}$

$\omega \to X(\omega)$ $\quad$ $X(\omega) = 0$ als voor Jospin stemt
$X(\omega) = 1$ als voor Chirac stemt

Stochastisch experiment = een actie waarvan de uitkomst bij herhaling onder dezelfde omstandigheden fluctueert.

We nemen een steekproef (S) uit de populatie (Ω):
$S = \{\omega_1, \omega_2, ..., \omega_n\} \subset \Omega$
$\omega_1, \omega_2, ..., \omega_n$ zijn de individuele kiezers uit de populatie.

Dataset : $\{X(\omega_1), X(\omega_2), ..., X(\omega_n)\}$: dit is nog niet geobserveerd
$X(\omega_1)$ is dus de stem van persoon ω_1 in de steekproef (en wordt genoteerd als 1 of 0 naargelang op wie dat ze stemmen), maar we weten nog niet op wie dat ze stemmen.

Hoofdletter 'X' = niet geobserveerde waarde = functie Kleine letter 'x' = geobserveerde waarde = meetwaarde $(x_1, x_2, ..., x_n)$ = reeks van ongeordende getallen

Deterministisch experiment = een experiment dat bij herhaling hetzelfde voorspelbaar resultaat oplevert. (vb. baksteen los laten → valt)

Steekproef : S
(merk op dat het kan dat S= Ω als de populatie heel klein is vb. 2e bach EW)

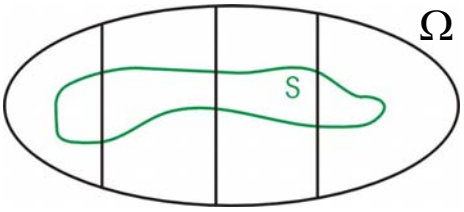
1.2 Types steekproeven (HB p12)

1. Lukraak : toevallige steekproef (met of zonder teruglegging)
= ieder lid van de populatie heeft evenveel kans om gekozen te worden.

Met teruglegging = ω uit Ω nemen, en dan terug in Ω plaatsen.
Zonder teruglegging = ω uit Ω nemen, en dan weg laten.
→ hoe groter de populatie, hoe kleiner het verschil tussen de twee methodes

vb. een verzameling van vijf knikkers; drie rood en twee wit:
$\Omega = \{R, R, R, W, W\}$
$\quad\quad\uparrow$
$X(\omega_1)$ = rood
P(rood)=3/5
Na één stap: met teruglegging: P(rood)=3/5
zonder teruglegging: P(rood)=2/4

2. Gestratifieerde steekproef
= de populatie wordt verdeeld in strata gebaseerd op een kenmerk (vb. geslacht, inkomen....)



S volgt de proporties

vb. Ω = studenten FETEW
X : $\omega \to X(\omega)$ = zakgeld van student ω
$|\Omega|$ =2800 | Jongens | = 1600 (57%)
| Meisjes | = 1200 (43%)

Een gestratifieerde steekproef (S) van lengte 100 heeft dezelfde proporties als de populatie:
→ 57 jongens en 43 meisjes

3. Clustersteekproef
= i.p.v. de steekproef één per één uit de populatie te nemen, wordt dit in groepjes gedaan.

vb. ■ president's verkiezingen → kiezers uit dezelfde straat bevragen
■ inhoud van blikjes cola → 10 blikjes uit 1 karton i.p.v. telkens 1 blikje uit 1 karton, 10 keer.

Door een steekproef te nemen, maken we een dataset.
Steekproef : $S = \{\omega_1, \omega_2, ..., \omega_n\} \subset \Omega$
Dataset : $D = \{X(\omega_1), X(\omega_2), ..., X(\omega_n)\}$

1.3 Classificatie van variabelen/data (HB p14)

Meetbaarheidsniveau

X: $\Omega \to \mathbb{R}$

Kwalitatief: ■ nominaal (geen ordening)
- geslacht
- nationaliteit
- godsdienst

■ ordinaal (wel ordening)
- graad van tevredenheid
- diploma (1. lager, 2. middelbaar, 3. hoger KT, 4. hoger LT)

Kwantitatief: ■ ratio geschaald (verhouding is vast)
- munten: EUR-USD (€0=$0)

■ interval geschaald (geen 'natuurlijk' nul punt)
- temperatuur: C-F-K

■ continu (X: $\Omega \to \mathbb{R}$)
- lengte
- rijtijd Brussel-Leuven

■ discreet (X: $\Omega \to \mathbb{N}$)
- aantal kinderen
- score op examen

Univariaat : X: $\Omega \to \mathbb{R}$
$\omega \to X(\omega)$

Bivariaat : X: $\Omega \to \mathbb{R}$ x $\mathbb{R}$ (koppelings georderd vb. lengte **en** gewicht)
$\omega \to (X_1(\omega), X_2(\omega))$

Multivariaat : X: $\Omega \to \mathbb{R}^n$ (vectoren van lengte n)

1.4 Statistische reeksen of datareeksen (HB p17)

Ω = 2e EW, 1996
X : $\Omega \to \mathbb{R}$
$\omega \to X(\omega)$ = resultaat van persoon 'ω' op Statistiek in juni

Aantal mensen in de populatie : $|\Omega|$ =26
Wij nemen de hele populatie als onze steekproef : S= Ω
en vinden de volgende dataset : D=$\{7,4,7,9,12,16,....\}$: ruwe data : een datareeks
Merk op : X is discreet (mogelijke uitkomsten : 0,1,2,….,20)

Absolute frequentie:

$i \in \{0,1,2,...,20\}$
F_i = absolute frequentie = aantal keer dat i waargenomen wordt
$\sum_{i=0}^{20} F_i = n = 26$ (= aantal personen in de populatie)

i	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
F_i	2	1	2	3	2	3	2	2	5	0	0	0	1	1	1	1
$F_{i,cum}$	2	3	5	8	10	13	15	17	22	22	22	22	23	24	25	26

$F_{i,cum}$ = absolute cumulatieve frequentie
= $F_0 + F_1 + ... + F_i$

$F_{20,cum} = 26 \leftarrow \sum_{i=0}^{20} F_i = 26$

* $F_{i,cum}$ heeft enkel zin vanaf ordinale data.

Relatieve frequentie:

$f_i = \frac{F_i}{n}$

$\Sigma f_i = 1$

relatieve cumulatieve frequentie : $f_{i,cum} = \frac{F_{i,cum}}{n}$

laatste klasse: $f_{k,cum} = 1$

2. Voorstelling data (HB p21)

Gegroepeerde data (6-8 klassen):

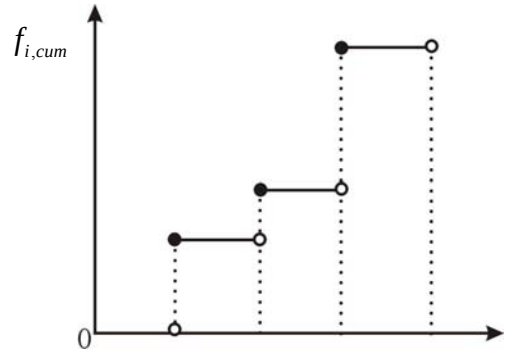
Score	#
4-6	5
7-9	8
10-12	9
13-15	0
16-18	3
19-20	1
	26

Klasse : beneden grens = 4
boven grens = 6

Klasse interval : 10-7 = 3
(het verschil tussen de benedengrenzen)

Klasse midden : $\frac{16+18}{2} = 17$

Staafdiagram (cumulatieve frequentie grafiek):



Rechts continu
Links niet continu

$X \rightarrow F_{cum}(X) =$ aantal met score $\leq X$

$F_{cum}(20) = 26$

Continue data:

Uitsluitend gegroepeerd in klassen
Ω = groep van mensen
X : $\omega \rightarrow X(\omega)$: lengte van persoon ' ω ' in cm (dit is een stochast)

....
150cm-159cm
160cm-169cm
....

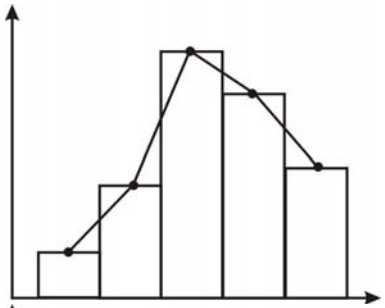
vermelde klassengrenzen (150-159)
$\neq$
ware klassengrenzen (149,5-159,5)

Klasse midden = $\frac{150+159}{2} = \frac{149,5+159,5}{2} = 154,5$

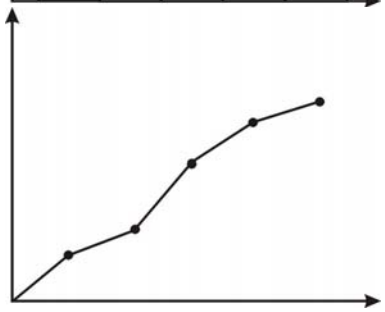
(het maakt hier niet uit of je de vermelde of ware klassengrenzen gebruikt)

Klasse interval = 160-150 = 10

Frequentie polygoon
= verbindt punten; klasse midden en frequentie
vb. voorstelling van een tijdreeks



Cumulatieve frequentie polygoon
= verbindt punten; ware bovengrens en cumulatieve frequentie



3. Datastatistieken voor 1 variabele (HB p37)
→ univariaat

3.1 Centrummaten - locatiematen (HB p38)

$\{9,9,9,...,9\}$: centrummaat = 9 (het gemiddelde is een centrummaat)
spreiding = 0 (want er is geen spreiding, alle getallen zijn hetzelfde)

3.1.1 Rekenkundig gemiddelde (HB p38)

rekenkundig gemiddelde = $\overline{x}$ (een kleine letter)

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$ (: in het formularium p1)

* als $S = \Omega$ dan wordt μ gebruikt als het rekenkundig gemiddelde van de hele populatie.

Voor gegroepeerde data:

$\overline{x} = \frac{\Sigma m_i F_i}{n}$

m_i = klassenmidden van de i-de klasse

F_i = geobserveerde frequentie van de i-de klasse

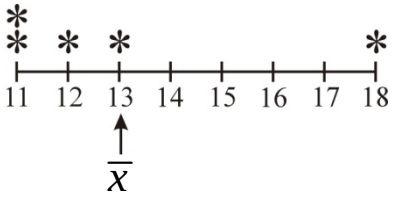
* soms wordt $\overline{m}_i$ (het gemiddelde van twee opeenvolgende beneden grenzen) gebruikt.

vb. 150-159 → m_i=154,50
160-169 → m_{i+1}=164,50
$\overline{m}_i$=155

Vier eigenschappen van het rekenkundig gemiddelde ($\overline{x}$):

1. Het gemiddelde is het massacentrum (evenwichtspunt) van de data.

vb.



2. De som van de deviaties rond het gemiddelde is gelijk aan nul :

$$\Sigma(x_i - \overline{x}) = (\Sigma(x_i)) - n\overline{x} = 0$$

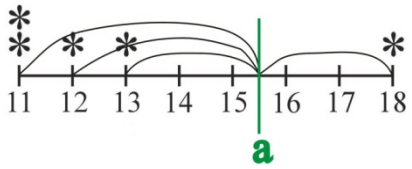
vb. $(x_1 - \overline{x}) + (x_2 - \overline{x}) + ... + (x_n - \overline{x})$: (deviaties rond het gemiddelde)
$= x_1 + x_2 + ... + x_n - n\overline{x}$
$= n\overline{x} - n\overline{x} = 0$

3. De som van de kwadratische deviaties rond een punt 'a' is minimaal als dat punt het gemiddelde is:

$$a \to \Sigma(x_i - a)^2 = \Sigma(x_i - \overline{x})^2 + n(\overline{x} - a)^2$$

dit is minimaal voor $a = \overline{x}$

vb.



Bewijs (via afgeleide):

$$\sum_{i=1}^{n}(x_i - a)^2 = \Sigma(x_i - \overline{x} + \overline{x} - a)^2$$

$$= \sum_{i=1}^{n}\left[(x_i - \overline{x})^2 + 2(x_i - a)(\overline{x} - a) + (\overline{x} - a)^2\right]$$

$$= \Sigma(x_i - \overline{x})^2 + 2(\overline{x} - a)\cancel{\Sigma(x_i - \overline{x})} + n(\overline{x} - a)^2$$

$$\cancel{\Sigma(x_i - \overline{x})} = 0$$

→ $$\sum_{i=1}^{n}(x_i - a)^2 = \Sigma(x_i - \overline{x})^2 + n(\overline{x} - a)^2$$

4. $\overline{x}$ is gevoelig voor uitschieters:

vb. 11, 11, 12, 13, **18**
$\overline{x} = 13 = \overline{x}_{100\%}$ (alle data wordt gebruikt; dit wordt sterk beïnvloed door de uitschieter (18))

Oplossing: het afgeknotte rekenkundige gemiddelde
→ bovenste en onderste geobserveerde waardes wegnemen
~~11~~, 11, 12, 13, ~~18~~ → $\overline{x}_{60\%} = 12$

3.1.2 Meetkundig gemiddelde (Geometrisch gemiddelde) (HB p40)

Het geometrisch gemiddelde van *n* data ($x_1, x_2, ..., x_n > 0$):
$g = \sqrt[n]{x_1 x_2 .....x_n}$ (: in het formularium p1)

Twee eigenschappen van het geometrisch gemiddelde (g):
1. Het log-geometrisch gemiddelde is het gewoone gemiddelde van de logdata:
$$\log g = \frac{(\log x_1 + \log x_2 + ... + \log x_n)}{n}$$

2. Het geometrisch gemiddelde is kleiner of gelijk aan het gewoon gemiddelde: $g \le \overline{x}$

vb. 2 , 32 : $\overline{x} = \frac{2+32}{2} = 17$
$g = \sqrt[2]{(2)(32)} = \sqrt{64} = 8$

vb. (toepassing):
$r_1, r_2, ..., r_n$ zijn jaarlijkse groeiritmes. Wat is de gemiddelde groei?
$(1+r^*)^n = (1+r_1)(1+r_2)(1+r_3)....(1+r_n)$
$(1+r^*) = \sqrt[n]{(1+r_1)(1+r_2)(1+r_3)....(1+r_n)}$

* $1+r^*$ is het geometrisch gemiddelde van $\{1+r_1, 1+r_2, ..., 1+r_n\}$

3.1.3 Mediaan : *Me* (HB p42)
= het punt dat 50% van de data links laat

Stap 1: ordenen van data van klein naar groot:
D : $\{x_1, x_2, ..., x_n\}$
→ $x_{<1>}, x_{<2>}, ..., x_{<n>}$ (geordende data)
$\le \ \le \ \le$

Stap 2 : kijken naar de middelste observatie:
vb. D=$\{7,5,1,8,12\} \to \{1,5,7,8,12\} \to Me = 7$

* als er een even aantal gegevens zijn, gebruikt men de $\frac{n+1}{2}$ de observatie als de mediaan

vb. 1, 5, 7, 8, 12, 20
3de observatie = 7 , 4de observatie = 8 → *Me* =7,5

men gebruikt lineaire interpolatie van de $\frac{n}{2}$ de en de $(\frac{n}{2}+1)$ de observatie in de dataset

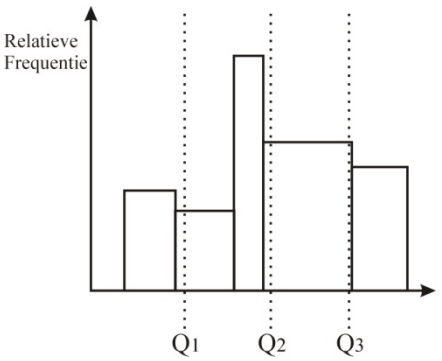
* de mediaan is ongevoelig voor uitschieters.

3.1.4 Kwartielen (HB p45)

Q_1 op 25% $\quad \frac{n+1}{4}$ de observatie

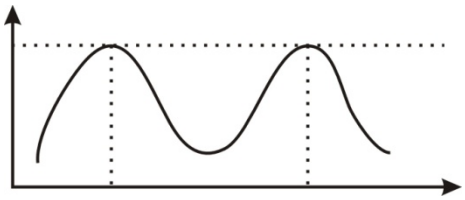
Q_2 op 50% = Me $\quad \frac{n+1}{2}$ de observatie

Q_3 op 75% $\quad \frac{3(n+1)}{4}$ de observatie



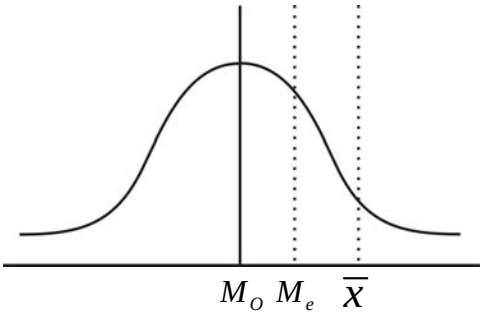
3.1.5 Modus : Mo (HB p47)
= de observatie in de dataset (D) die het meeste voorkomt (vanaf kwalitatieve data)
→ 'de' modus is niet altijd uniek, je kan er twee of meer hebben

vb. 2 bij bimodale data



3.1.6 Empirische relatie tussen de locatiematen

De relatie tussen modus en mediaan:
$\overline{x} - Mo \approx 3(\overline{x} - Me)$ (: in het formularium p1)



3.2 Spreidingsmaten (HB p47)

1. gemiddelde afwijking : GA

gemiddelde afwijking tot $\overline{x}$: $GA = \frac{\Sigma|x_i - \overline{x}|}{n}$ (: in het formularium p1)

gemiddelde afwijking tot Me : $GA = \frac{\Sigma|x_i - Me|}{n}$

2. variatie : V
= de som van de gekwadrateerde afwijkingen van $\overline{x}$:

$V = \sum_{i=1}^{n}(x_i - \overline{x})^2$, maar dit kunnen we nog herschrijven:

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

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

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

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

$$\boxed{V = \sum_{i=1}^{n}x_i^2 - n\overline{x}^2}$$

3. variantie : $\tilde{s}^2$ of s^2
= de gemiddelde gekwadrateerde afwijking van $\overline{x}$

$\tilde{s}^2 = \frac{\Sigma(x_i - \overline{x})^2}{n}$: variantie met n-weging (: in het formularium p1)

$s^2 = \frac{\Sigma(x_i - \overline{x})^2}{n-1}$: steekproefvariantie met (n-1) weging (: in het formularium p1)

* als $S = \Omega$ (de steekproef bevat de hele populatie)
dan $\overline{x} = \mu$
en $\tilde{s}^2 = \sigma^2$

* Let op : bij kleine steekproeven: $s^2 \geq \tilde{s}^2$

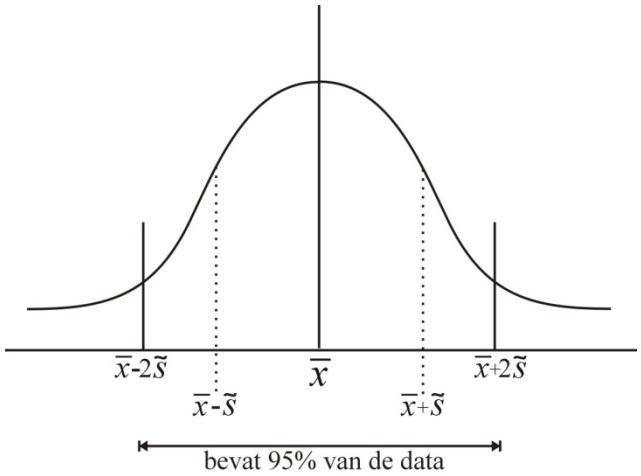
Stelling van Tchebychev (1821-1894):

Zij X: $\Omega \to \mathbb{R}$ een kenmerk
We nemen een steekproef uit Ω : $S \in \Omega$
En bekomen een dataset: $D = \{x_1, x_2, ..., x_n\}$

Dan : bevat het interval $[\overline{x} - k\tilde{s}, \overline{x} + k\tilde{s}]$ ten minste $(1 - \frac{1}{k^2})$ x 100% van de data.
= dit geldt voor eender welke dataset

vb. $[\overline{x} - 2\tilde{s}, \overline{x} + 2\tilde{s}]$ bevat ten minste 75% van de data
$[\overline{x} - 3\tilde{s}, \overline{x} + 3\tilde{s}]$ bevat ten minste 88,9% van de data

* voor bijzondere datasets levert de stelling een zwakkere schatting
vb. normaal verdeelde data:



4. range
= de hoogste geobserveerde waarde min de laagste geobserveerde waarde

5. interkwartiele afstand (IQR)
= $Q_3 - Q_1$

Al deze informatie kan worden voorgesteld als een boxplot:

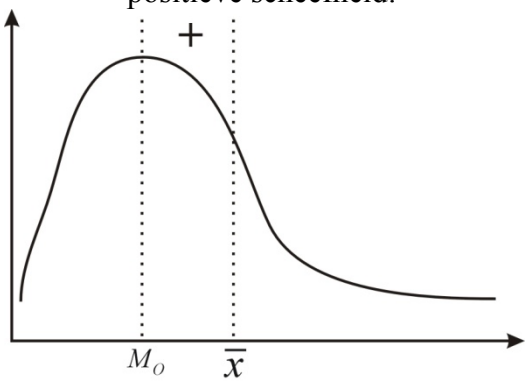
* 1 t.e.m. 5 zijn allemaal absolute spreidingsmaten.
Voorbeeld van een relatieve spreidingsmaat (dimentieloos):

variatie coëfficiënt : VC = $\frac{\tilde{s}}{\overline{x}}$ (: in het formularium p1)

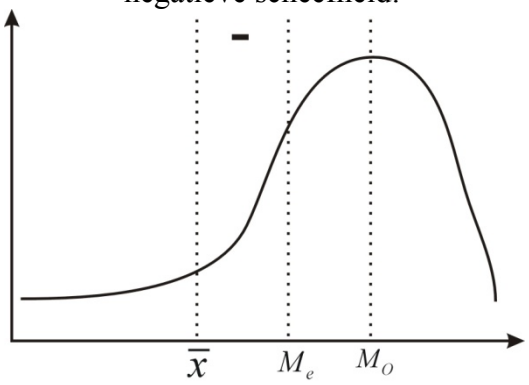
3.4 Vormmaten (scheefheidsmaten) (HB p55) (: allemaal in het formularium p1)

1. Pearson$_1$ = $\frac{\overline{x} - Mo}{s}$ en Pearson$_2$ = $\frac{3(\overline{x} - Me)}{s}$

positieve scheefheid:



negatieve scheefheid:

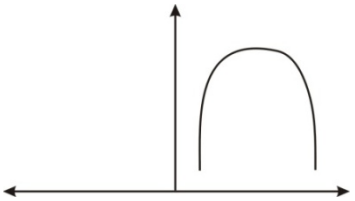


2. Skewness : Sk = $\frac{\Sigma(x_i - \overline{x})^3}{n\tilde{s}^3}$

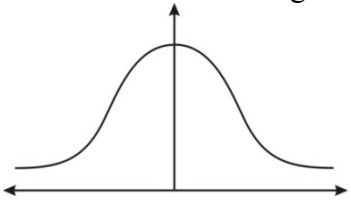
3. Kurtosis = $\frac{\Sigma(x_i - \overline{x})^4}{n\tilde{s}^4} - 3$

= mechanisme om data te toetsen om te zien of het normaal verdeeld is

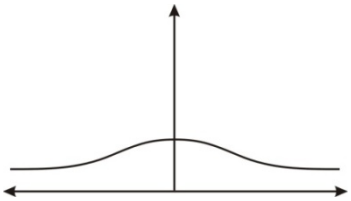
Kurtosis < 0 :
zwakke staarten



Kurtosis = 0 :
normale verdeling

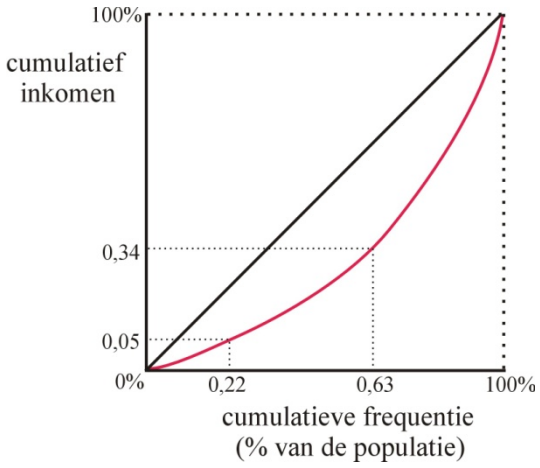


Kurtosis > 0 :
zware staarten



3.5 Concentratiematen (HB p57)

Lorenz-curve (inkomensverdeling)



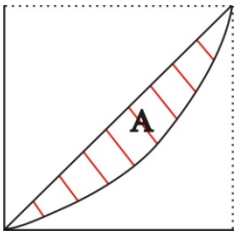
Inkomensklasse	$f_{i,cum}$	Cumulatief gedeelte van totaal inkomen
0-50000	0,22	0,05
50000-125000	0,63	0,34

→ lees: de 'onderste' 22% van de bevolking beschikt over 5% van het inkomen
→ 100% van de bevolking beschikt over 100% van het inkomen

* Indien constante en perfecte verdeling zal de Lorenz-curve samen vallen met de diagonaal.

Gini-coefficient : g

$$g = \frac{\text{oppervlakte A}}{\text{oppervlakte driehoek onder diagonaal (5000)}}$$



$$g = \frac{\Sigma x_j y_{j+1} - x_{j+1} y_j}{10000}$$ (: in het formularium p1)

$$g = \frac{\Sigma \det \begin{pmatrix} x_j & y_j \\ x_{j+1} & y_{j+1} \end{pmatrix}}{10000}$$

(x_j en y_j zijn cumulatieve percentages)

* extreme gevallen:
g = 0 : perfecte gelijkheid
g = 1 : extreme ongelijkheid (99% heeft niets, 1% heeft alles)

4. Datastatistieken voor twee of meer variabelen (HB p59)

bi-variaat :

$(x, y) : \Omega \to \mathbb{R}^2$ (aan de populatie gaan we twee kenmerken koppelen vb. lengte **en** gewicht)

$\omega \to \big(X(\omega), Y(\omega)\big)$

$S \in \Omega$, $S = \{\omega_1, \omega_2, ..., \omega_n\}$

$D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$

met: $(x_1, y_1) = (X(\omega_1), Y(\omega_1))$ etc.

* per dimensie : locatiematen, spreidingsmaten, vormmaten → zie vorige
voor twee dimensies:

Covariantie (samenbeweging):

= meten van het samenbewegen van de veranderlijke (vb. relatie tussen lengte en gewicht)

$$\tilde{s}_{x,y} = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{n} \quad \text{(met n weging)}$$

$$s_{x,y} = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{n\tilde{s}_{x,y}}{n-1} \quad \text{(met n-1 weging)}$$

NB. als $S = \Omega$:
$\tilde{s}_{x,y} = \sigma_{x,y}$
$\tilde{s} = \sigma$
$\overline{x} = \mu$

* indien $S = \Omega$ dan wordt $\sigma_{x,y}$ gebruikt in de plaats van $\tilde{s}_{x,y}$

Correlatiecoëfficiënt:

$$r = r_{x,y} = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\left(\sqrt{\Sigma(x_i - \overline{x})^2}\right)\left(\sqrt{\Sigma(y_i - \overline{y})^2}\right)}$$

$$r_{x,y} = \frac{\tilde{s}_{x,y}}{\tilde{s}_x \tilde{s}_y} = \frac{s_{x,y}}{s_x s_y}$$

* De correlatiecoëfficiënt ligt altijd tussen -1 en 1 $(-1 \le r_{x,y} \le 1)$

En $r_{x,y} = \pm 1$ als en alleen als de data perfect colineair zijn.

→ de correlatiecoëfficiënt is dus een maat voor een lineaire trend in de data.

Bewijs:

Beschouw de vectoren; $\vec{\upsilon}$ en $\vec{\omega}$:

$\vec{\upsilon} = (x_1 - \overline{x}, x_2 - \overline{x}, ..., x_n - \overline{x})$

$\vec{\omega} = (y_1 - \overline{y}, y_2 - \overline{y}, ..., y_n - \overline{y})$

We kunnen dan de correlatiecoëfficiënt schrijven als een product van deze vectoren:

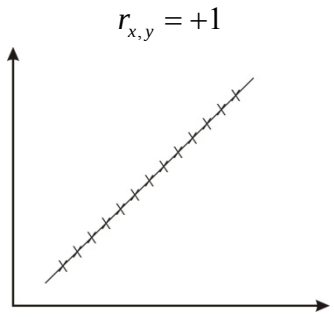
$$r_{x,y} = \frac{\vec{\upsilon}}{\|\vec{\upsilon}\|} \cdot \frac{\vec{\omega}}{\|\vec{\omega}\|} = \cos(\vec{\upsilon}, \vec{\omega})$$

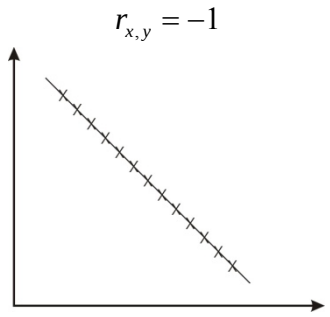
en bijgevolg weten we dat: $-1 \le r_{x,y} \le 1$

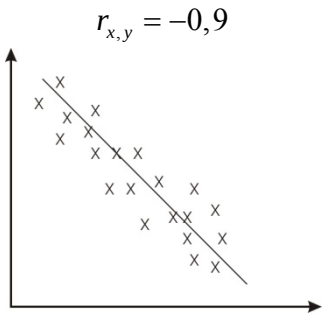
Als $r_{x,y} = \pm 1$ dan $\vec{\upsilon} = \pm a\vec{\omega}$ met $a > 0$ of $x_i - \overline{x} = \pm a(y_i - \overline{y})$ voor elke i

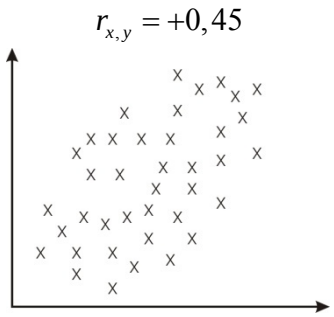
$x_i \pm ay_i = \overline{x} + a\overline{y}$ (hangt niet af van i)

Conclusie: als $r_{x,y} = \pm 1$ dan liggen de datapunten op een rechte.









5. Indexen (HB p69) = geen leerstof

DEEL II
KANSREKENING

6. Kansrekening (HB p81)

6.1 Wat is kanstheorie? (HB p83)

Stochastische experiment = een actie waarvan de uitkomst bij herhaling onder dezelfde omstandigheden fluctueert. Je kan dus niet op voorhand de uitslag voorspellen maar je kan wel iets zeggen over de uitkomsten verzameling (Ω).

Voorbeelden:

a) trek een kaart uit een kaartspel van 52 kaarten:
$\Omega = \{\heartsuit 1, ..., \clubsuit H\}$: de uitkomsten verzameling is de 52 kaarten
→ wat is de kans dat de getrokken kaart een hartendrie is?

een gebeurtenis is een deelverzameling van Ω . vb.:
D : kaart is een drie
H : kaart is een harten
$D \cap H$: kaart is een hartendrie
$D \cup H$: kaart is een drie of een harten
$H^c = \Omega / H$: kaart is geen harten
$D \mid H$: kaart is een drie, gegeven dat ze een harten is

men zoekt de kans van een gebeurtenis. vb.:
$P(D), P(H), P(D \cap H), P(D \cup H), P(H^c), P(D \mid H)$
lees: $P(D)$: "de kans van gebeurtenis D"

b) gooi een dobbelsteen en noteer het aantal ogen:
$\Omega = \{1,2,3,4,5,6\}$
→ wat is de kans dat het aantal ogen even is?

c) tos een munt en noteer de zijde:
$\Omega = \{K, M\}$
→ wat is de kans dat de uitkomst 'M' is?

d) rij van Leuven naar Brussel en noteer de rijtijd:
$\Omega \in \mathbb{R}^+$
→ wat is de kans dat de rijtijd minder dan 40 minuten is? Meer dan 2 uur?
→ bij continue kenmerken (vb. tijd) zijn gebeurtenissen intervallen

e) noteer de wachttijd in een telefoon centrale tot de volgende oproep:
$\Omega \in \mathbb{R}^+$
→ wat is de kans dat de wachttijd; kleiner dan? groter dan? tussen twee?

f) iemand fietst van thuis naar werk en passeert 3 verkeerslichten; noteer de stand:
$\Omega = (ggg, ggr, grg, rgg, grr, rrg, rgr, rrr)$
→ wat is de kans dat het drie keer rood is?

6.2 Kansruimte (HB p85)

Kansruimte : (Ω, G, P)

1. Uitkomstenverzameling (populatie, universum) : Ω : een niet-lege verzameling, de verzameling van alle mogelijke uitkomsten van een experiment.

2. Gebeurtenissenverzameling : G : een verzameling van deelverzamelingen van Ω
* axioma's voor G (HB p92): lezen maar geen leerstof

3. Kansmaat : P : (probability) is een functie: $P : G \to \mathbb{R}$
$A \to P(A)$
die aan elke gebeurtenis A een getal P(A), de kans van gebeurtenis A, toekent. Deze functie heeft de eigenschappen:

$$[\text{ de klassieke kansdefinitie : } P(A) = \frac{\text{aantal gunstige uitkomsten}}{\text{aantal mogelijke uitkomsten}} \text{]}$$

P(1) : $P(\Omega) = 1$
→ De kans dat we een uitkomst van Ω trekken uit Ω is altijd 100%. Want Ω bevat alle uitkomsten en we kunnen met zekerheid zeggen dat we uit Ω trekken.

P(2) : $P(A) \geq 0$ voor elke gebeurtenis (A) in de gebeurtenissenverzameling (G).
→ de kans is altijd positief, nooit negatief.

P(3) : indien $A, B \in G$ voor elkaar uitsluitende gebeurtenissen A en B (i.e. $A \cap B = \varnothing$) dan:
$\boxed{P(A \cup B) = P(A) + P(B)}$

(* $A \cap B = \varnothing$ wil zeggen dat de verzameling $A \cap B$ leeg is)

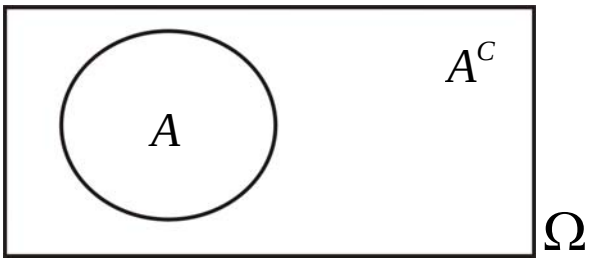
vb. $P(A) = 0,25$ en $P(B) = 0,25$ als $A \cap B = \varnothing$, dan:
$P(A \cup B) = P(A) + P(B) = 0,5$

P(3)' : voor een rij disjuncte gebeurtenissen ($A_1, A_2, ..., A_n, ... \in G$, $A_i \cap A_j = \varnothing$ dus $i \neq j$)
geldt: $P(\cup A_i) = \Sigma P(A_i)$

Eigenschappen van een kansruimte - (Ω, G, P)

1. $P(A) \leq 1$ voor elke gebeurtenis (A) in de gebeurtenissenverzameling (G).

Bewijs : zij A in G
definieer $A^C = \Omega - A$
dan $A \cap A^c = \varnothing$ (leeg)

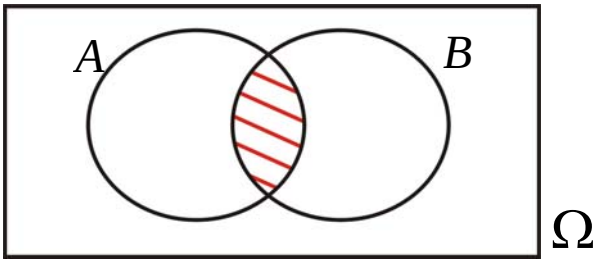


Som regel: $P(A \cap A^c) = P(\Omega)$
eigenschap P(3): $= P(A) + P(A^c)$
eigenschap P(1): $= 1$

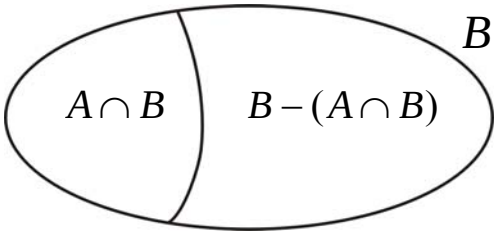
→ Twee getallen $P(A)$ en $P(A^c)$ tellen op tot 1 (kansen mogen niet groter zijn dan 1).
eigenschap P(2): $P(A)$ en $P(A^c) \geq 0$

2. optelregel uitgebreid (met deze regel mogen er wel elementen in de doorsnede zitten)
zij A en B in G:
$\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$ (: in het formularium p2)
→ de doorsnede wordt afgetrokken om de dubbeltelling te voorkomen

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



som regel:
$P(A \cap B) + P(B - (A \cap B)) = P(B)$



(Ω, G, P) is een kansruimte, Ω is eindig (vb. kaart trekken uit een kaartspel; maar 52 kaarten)

Uniforme kansverdeling : de kans van de gebeurtenis is hetzelfde voor alle elementen in de verzameling → $P(\{\omega\})$ is dezelfde voor elke ω in Ω

$$P(\{\omega\}) = \frac{1}{|\Omega|} = \frac{1}{\text{aantal elementen in de verzameling}}$$

vb. Ω is een spel van 52 kaarten:
$P(\{\omega\}) = \frac{1}{52}$ = kans dat een kaart ω wordt getrokken
omdat elke kaart evenveel kans maakt om getrokken te worden
= uniforme kansverdeling

Oefeningen/Voorbeelden

Oefening 1 : verjaardagen probleem

Wat is de kans dat in een groep van 'm' personen er twee of meer personen zitten met dezelfde verjaardag?

$\Omega = k^m$ =k x k x k x ...x k
(m keer k)
m = lengte van de groep
k = dag uit de kalender van 365 dagen

A is de gebeurtenis van ω in Ω waarbij $\omega_i = \omega_j$ voor ten minste één i en j verschillend van elkaar. (lees: A is de gebeurtenis waar twee verschillende mensen (i en j) op dezelfde dag hun verjaardag hebben.)

$\omega = (\omega_1, \omega_2, \omega_3, ..., \omega_m)$

Kans van gebeurtenis A? : $P(A) = ? = \frac{|A|}{|\Omega|} = \frac{|A|}{365^m}$

P is uniform verdeeld over Ω

Kans dat een persoon hun verjaardag heeft op één specifieke dag: $P(\{\omega\}) = \frac{1}{|\Omega|}$

$|A| = ?$ = aantal elementen in Ω waar $\omega_i = \omega_j$

Het is gemakkelijker om A^c te berekenen dan A. A^c (A compliment) is de gebeurtenis van ω in Ω waarbij alle coördinaten van ω van elkaar verschillen. (= waar er geen twee personen in de groep zijn met dezelfde verjaardag.)

$|A^c| = 365$ x 364 x 363 x...x $(365(m-1))$: elk individu heeft een verschillende verjaardag

$$P(A^c) = \frac{|A^c|}{|\Omega|} = \frac{365 \text{ x } 364 \text{ x } 363 \text{ x...x } (365(m-1))}{365^m}$$

Met deze kans kunnen we nu gemakkelijk de kans van A berekenen:

$P(A) = 1 - P(A^c)$

$$P(A) = 1 - \frac{|A^c|}{|\Omega|} = 1 - \frac{365 \text{ x } 364 \text{ x } 363 \text{ x } ... \text{ x } (365(m-1))}{365^m}$$

We kunnen $P(A)$ nu berekenen voor verschillende waardes van 'm' (voor verschillende groottes van de groep):

als $m = 23 \rightarrow P(A) = 0,507$
$m = 24 \rightarrow P(A) = 0,536$
$m = 30 \rightarrow P(A) = 0,706$
$m = 40 \rightarrow P(A) = 0,891$
$m = 64 \rightarrow P(A) = 0,997$

→ vanaf 64 personen ben je bijna zeker dat er 2 personen in de groep zijn met dezelfde verjaardag.

Oefening 2:

Wat is de kans dat in een groep van m personen, iemand op "mijn" verjaardag verjaart?

$\Omega = k^m$

B = de gebeurtenis dat iemand op mijn verjaardag verjaart

$|B^c|$ = uitkomst waar geen enkele samenvalt met mijn verjaardag

$= 364 \text{ x } 364 \text{ x } 364 \text{ x..} = 364^m$ (je mag kiezen eender welke dag, behalve die van mij)

$$P(B^c) = \frac{|B^c|}{|\Omega|} = \frac{364^m}{365^m}$$

$$P(B) = 1 - P(B^c)$$

$$P(B) = 1 - \frac{364^m}{365^m}$$

We kunnen opnieuw m invullen om de kans te berekenen.

→ vanaf m = 253 hebben we $P(B) \geq 0{,}5$

→ indien we met tenminste 253 mensen zijn kan je met een kans van 50% zeggen dat er iemand op "mijn" verjaardag verjaart.

6.3 Voorwaardelijke kans en onafhankelijke gebeurtenissen (HB p94)

Voorwaardelijke kans : $\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}}$ met $P(B) \neq 0$ (: in het formularium p2)

vb. een kansruimte (Ω, G, P) met Ω een spel van 52 kaarten en P uniforme kans.

drie gebeurtenissen : H = kaart is harten

D = kaart is drie

R = kaart is rood

$$P(H) = \frac{13}{52} = \frac{1}{4}$$

$$P(D) = \frac{4}{52} = \frac{1}{13}$$

$$P(R) = \frac{26}{52} = \frac{1}{2}$$

$P(H|D)$ = de kans dat de getrokken kaart een harten is gegeven dat de kaart een drie is

$|$ = "gegeven dat"

$$\{\heartsuit 3, \diamondsuit 3, \clubsuit 3, \spadesuit 3\} = P(H|D) = \frac{1}{4} = \frac{P(H \cap D)}{P(D)} = \frac{\text{kans harten en drie}}{\text{kans drie}} = \frac{1/52}{4/52}$$

$P(H|R)$ = de kans dat de getrokken kaart een harten is gegeven dat de kaart rood is

$|R| = 26$

$|H| = 13$

$$P(H|R) = \frac{13}{26} = \frac{1}{2} = \frac{P(H \cap R)}{P(R)} = \frac{1/4}{1/2} = \frac{1}{2}$$

We stellen dus dat:
$P(H) = P(H|D)$ → de informatie is niet relevant.
$P(H) \neq P(H|R)$ → de informatie is wel relevant dus moeten we er rekening bij houden.

Definities :

■ Zij (Ω, G, P) een kansruimte en zij A en B twee gebeurtenissen:
De voorwaardelijke kans van A gegeven B is het getal $P(A|B)$ gegeven door:

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}} \text{ mits } P(B) \neq 0$$

Opmerking: $P(A|\Omega) = \frac{P(A \cap \Omega)}{P(A)} = P(A)$

■ Zij (Ω, G, P) een kansruimte en zij A en B twee gebeurtenissen:
A en B zijn onafhankelijke gebeurtenissen als: $\boxed{P(A|B) = P(A)}$

vb. * $P(H|D) = P(H) = 1/4$ → H is dus onafhankelijk van D
* $P(H|R) = 1/2 > P(H) = 1/4$ → H is niet onafhankelijk van R, de informatie is relevant

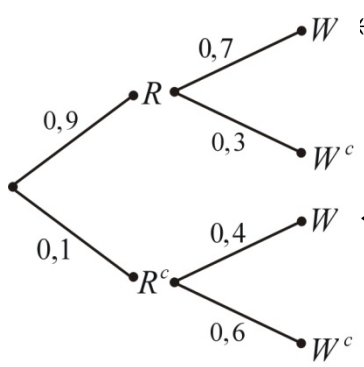
Opmerking : indien A onafhankelijk is van B, dan is B ook onafhankelijk van A:
$P(A|\,B) = \frac{P(A \cap B)}{P(B)} = P(A)$
waaruit $P(A \cap B) = P(A) + P(B)$
of $\frac{P(B \cap A)}{P(A)} = P(B|A) = P(B)$

Bepalen van kansen via kansbomen:

Oefening 1:
Woensdag, als je wakker wordt en opstaat om 8u30 dan ben je op tijd voor de les van 9u30.
Je zet je wekker om 8u30.
Indien je wekker afloopt dan wordt je wakker met een kans 0,7
Indien je wekker niet afloopt dan wordt je wakker met een kans 0,4
De kans dat de wekker afloopt is gelijk aan 0,9

Vraag : Bepaal de kans dat je op tijd wakker bent.
R is de gebeurtenis “de wekker rinkelt om 8u30”
W is de gebeurtenis “je wordt wakker om 8u30”

We tekenen een kansboom met al deze gegevens:



$\leftarrow P(R)P(W|R) = 0,9 \text{ x } 0,7 = 0,63$

$\leftarrow P(R^c)P(W|R^c) = 0,1 \text{ x } 0,4 = 0,04$

Kans dat je wakker wordt om 8u30 : $P(W) = 0,63 + 0,04 = 0,67$ (via de optelregel)

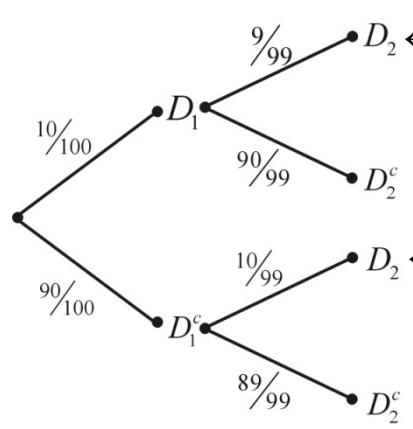
Oefening 2:
Een partij van 100 radio's bevat 10 defecte radio's.
We trekken lukraak twee radio's (zonder teruglegging)
Bepaal de kans dat de tweede radio defect is.

D_1 : gebeurtenis dat de eerste radio defect is

D_2 : gebeurtenis dat de tweede radio defect is

Ω : verzameling van vectoren van lengte r

$\{(r_1, r_2) | r_1 \neq r_2\}$ zijn twee radio's uit de partij van 100



$\leftarrow$ kans $= \frac{10}{100} \text{ x } \frac{9}{99} = \frac{90}{9900} = P(D_2|D_1) \text{ x } P(D_1)$

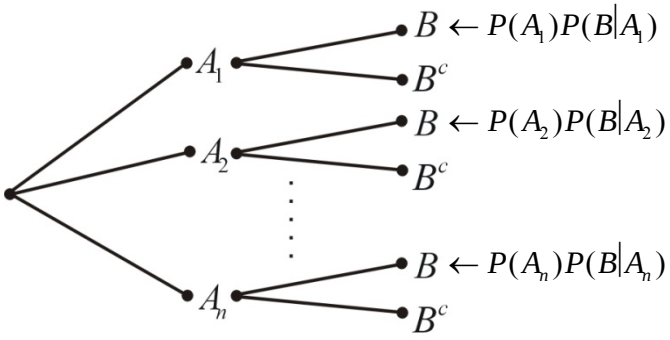
(we zijn geïnteresseerd in D_2)

$\leftarrow$ kans $= \frac{90}{100} \text{ x } \frac{10}{99} = \frac{900}{9900} = P(D_2|D_1^c) \text{ x } P(D_1^c)$

$$\Rightarrow P(D_2) = \frac{90}{9900} + \frac{900}{9900} = \frac{990}{9900} = \frac{1}{10}$$

Wet van de totale kans (stratifiëringsregel) (HB p96)

Zij (Ω, G, P) een kansruimte
Zij B een gebeurtenis
Zij $\Omega = A_1 \cup A_2 \cup ... \cup A_n$ een partitie van Ω in onderling disjuncte gebeurtenissen
(elkaar onafhankelijke gebeurtenissen $A \cap B = \varnothing$: de verzameling is leeg)
Dan: $P(B) = \left[P(A_1)P(B|A_1)\right] + \left[P(A_2)P(B|A_2)\right] + ... + \left[P(A_n)P(B|A_n)\right]$



$$\Rightarrow P(B) = \sum_k P(B \cap A_k) = \sum_k P(A_k).P(B|\,A_k)$$

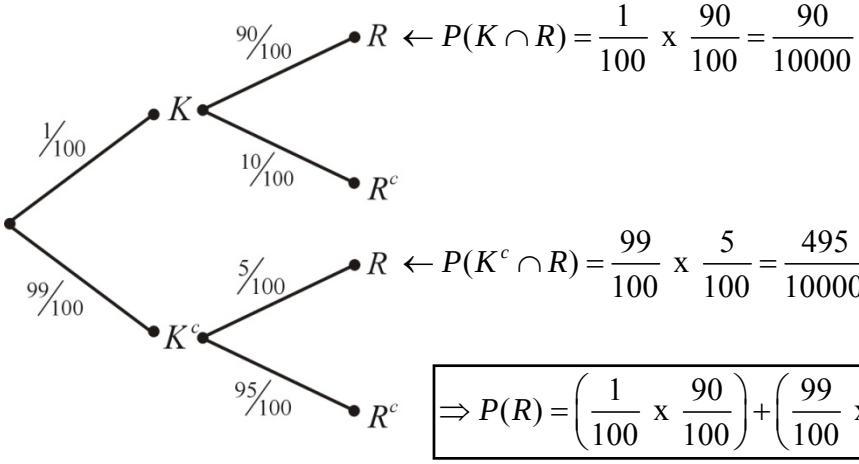
<u>Stelling van Bayes</u> (HB p96)

Zij (Ω, G, P) een kansruimte
Zij B een gebeurtenis
Zij $\Omega = A_1 \cup A_2 \cup ... \cup A_n$ een partitie van Ω in onderling <u>disjuncte</u> gebeurtenissen

Dan: $$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{P(B)}$$ (: in een andere vorm in het formularium p2)

<u>vb.</u> ontwikkelen van een kankertest:
90% van de personen <u>met</u> kanker reageert "positief"
5% van de personen <u>zonder</u> kanker reageert "positief"
1% van de bevolking die op consultatie komt heeft kanker

<u>Vraag</u> : is dit een "goede" test?
K = gebeurtenis; persoon heeft kanker
R = gebeurtenis; persoon reageert "positief"



$$\Rightarrow P(R) = \left(\frac{1}{100} \text{ x } \frac{90}{100}\right) + \left(\frac{99}{100} \text{ x } \frac{5}{100}\right) = \frac{117}{2000}$$

We maken gebruik van de stelling van Bayes om de kans dat een persoon die "positief" reageert op de test ook kanker heeft ($P(K|R)$) te berekenen:

$$P(K|R) = \frac{P(K)P(R|K)}{P(R)} = \frac{1/100 \text{ x } 90/100}{(1/100 \text{ x } 90/100) + (99/100 \text{ x } 5/100)}$$
$$= \frac{90}{90 + (99 \text{ x } 5)}$$
$$= \frac{2}{2+11} = \frac{2}{13} \approx 15\%$$

→de kans dat een persoon die "positief" reageert op de test ook kanker heeft in 15%

Hoe kunnen we deze test best verbeteren? Door 90% te verhogen of 5% te verlagen?

90% ↑ 92% → $P(K|R) = \frac{92}{92 + (99 \text{ x } 5)} \approx 15\%$

5% ↓ 3% → $P(K|R) = \frac{90}{90 + (99 \text{ x } 3)} \approx 23\%$ = de betere oplossing

7. Stochastische variabelen (HB p108)

7.1 Stochastische variabele (HB p109)

Een stochastische variabele (toevalsveranderlijke) is de toevalsuitkomst van een stochastisch experiment.

* trek een student uit het eerste jaar en meet de lengte in cm
* trek een blik opgevuld door een welbepaald machine en meet de inhoud in cl
* trek een gloei lamp van een bepaald merk en meet de levensduur in uren

Zij (Ω, G, P) een kansruimte
Een stochastische variabele is een variabele:

$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$

De stochastische variabele X is discreet indien de beeldverzameling (image / uitkomstenverzameling): ImX = $\{x_1, x_2, ..., x_n\}$ eindig is (discreet).
(om het discrete te benadrukken gebruiken we ImX in plaats van Ω)

vb. tos een "eerlijke" munt twee keer en noteer het aantal keer 'M':

$\Omega = \{(K,K),(K,M),(M,K),(M,M)\}$

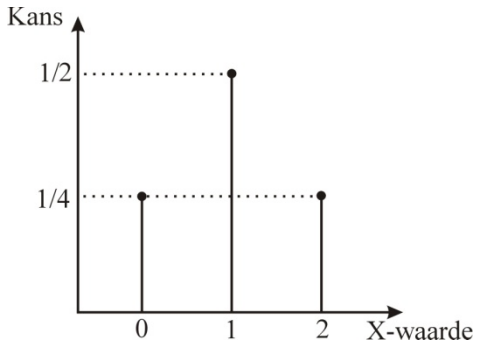
$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$ = aantal keer M in ω

Uitkomsten verzameling: ImX = $\{0,1,2\}$: discrete stochast, omdat de verzameling eindig is ($\Omega =$)

$P\left(\{\omega | X(\omega) = k\}\right)$ met k = 0, 1, of 2
$= P_k(X = k)$: (verkorte notatie)

Grafiek van de kansdichtheid van de stochast X:

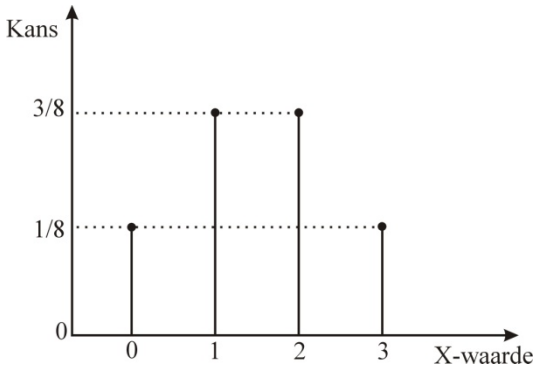
kans dat je nooit M gooit: $P_0 = 1/4$
kans dat je 1 keer M gooit: $P_1 = 1/2$
kans dat je 2 keer M gooit: $P_2 = 1/4$
$\Sigma = 1$



vb. een fietser legt een traject af en passeert drie verkeerslichten. Tel het aantal groen.
$\Omega = \{ggg, ggr, grg, rgg, grr, rgr, rrg, rrr\}$
$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$ = aantal keer g(roen) in ω
Uitkomsten verzameling : Im$X = \{0,1,2,3\}$

Grafiek van de kansdichtheid van de stochast X:

kans 0 groen : $P_0 = \frac{1}{8}$
kans 1 groen : $P_1 = \frac{3}{8}$
kans 2 groen : $P_2 = \frac{3}{8}$
kans 3 groen : $P_3 = \frac{1}{8}$
$\Sigma = 1$



Zij $X : \Omega \to \mathbb{N}$ een discrete stochastische veranderlijke
De kansdichtheid van X is de afbeelding van $p : \mathbb{N} \to \mathbb{R}$

$$k \to p_k = P(\{\omega | X(\omega) = k\}) = P_k(X = k)$$

Voorwaarden: * $p_k \geq 0$: (de kans mag niet negatief zijn)
* $\Sigma p_k = p_0 + p_1 + ... + p_n = 1$: (de kansen moeten optellen tot 1)

Zij (Ω, G, P) een kansruimte
stochastische variabele $X : \Omega \to \mathbb{R}$
→ dit is een <u>discrete</u> stochastische variabele
: de uitkomsten verzameling van X is eindig (of aftelbaar: Im $= \mathbb{N}$).

Dichtheidsfunctie: $p_0, p_1, p_2, ..., p_n$

$$p_k = P(X = k) = P(\{\omega | X(\omega) = k\})$$

Eigenschappen : * $p_k \geq 0$: (de kans mag niet negatief zijn)
* $\Sigma p_k = 1$: (de kansen moeten optellen tot 1)

P : kansmaat
p : dichtheidsfunctie
X : stochastische variabele
x : waarde : reëel getal

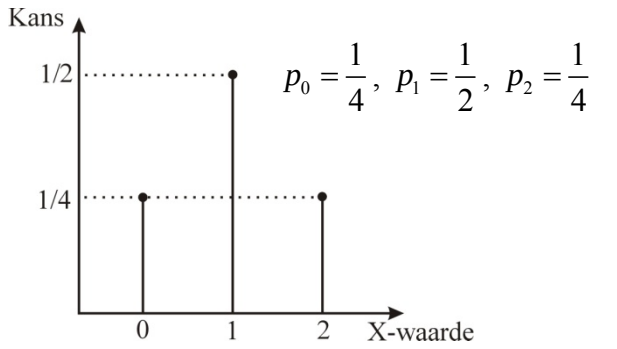
cumulatieve verdelingsfunctie: $F : \mathbb{R} \to \mathbb{R}$

$$X \to F(x) = P(X \leq x) = P(\{\omega | X(\omega) \leq x\})$$

vb. twee keer tossen van een munt :

$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$: aantal keer M in ω
$\text{Im(X)} = \{0,1,2\}$

→ discrete dichtheid p

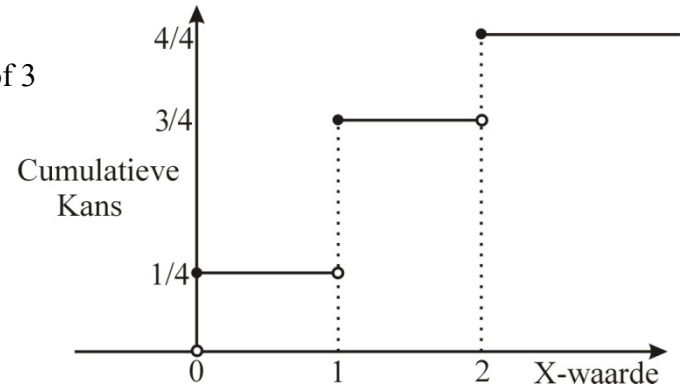


$p_0 = \frac{1}{4}$, $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{4}$

Cumulatieve verdelingsfunctie :

$F : \mathbb{R} \to \mathbb{R}$
$x \to P(X \leq x)$: de kans dat $X \leq 1,2$ of 3

Kenmerken:
* F is nergens dalend
* F is rechts continu
* $F(-\infty) = 0$: (horizontale asymptoot)
* $F(+\infty) = 1$



De cumulatieve verdelingsfunctie maakt disjuncte bewegingen (sprongen):
De functie is 0 voor negatieve waardes;
in 0 verspringt de functie naar ¼ ;
in 1 verspringt de functie naar ¾ ;
in 2 verspringt de functie naar 1.

7.2 Typische verdelingen (HB p116)

7.2.1 Uniforme discrete verdeling : $U\{1,...,N\}$

uniforme = de kans van de gebeurtenis is hetzelfde voor alle elementen in de verzameling.
discrete = de verdeling is eindig.

Verband tussen dichtheid en de verdeling:
* $p_K = F(k) - F(k-1)$
* $F(k) = p_0 + p_1 + ... + p_K$

7.2.2 Uniforme continue verdeling : $U[a,b]$

continue = de verdeling is niet eindig

Zij (Ω,G,P) een kansruimte

De stochastische variabele $X:\Omega\to\mathbb{R}$ is een continue stochastische variabele indien er een continue dichtheidsfunctie $p:\mathbb{R}\to\mathbb{R}$ bestaat zodat:

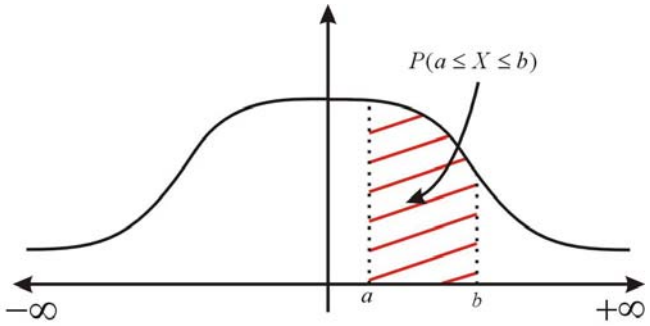
* $p(x)\geq 0$ voor elke x in $\mathbb{R}$: de functie p mag nergens negatief zijn
* $\int_{-\infty}^{+\infty} p = 1$: de totale kans is gelijk aan 1
* $P(\{\omega | a \leq X(\omega) \leq b\})$ voor elke a, b in $\mathbb{R}$

  $= \int_{a}^{b} p = F(b) - F(a)$

  = wat is de kans at onze stochast tussen a en b ligt?

→ we maken nu gebruik van de integraal om de oppervlakte onder de grafiek te berekenen

* de oppervlakte onder de grafiek tussen a en b is gelijk aan $P(a \leq X \leq b)$
* $P(-\infty < X < +\infty) = 1$: de totale oppervlakte onder de curve is gelijk aan één



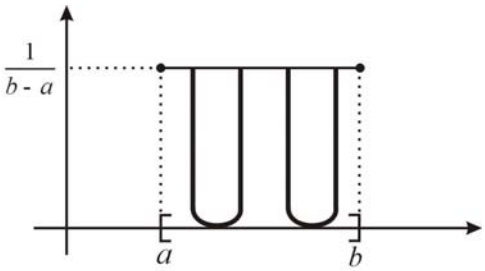
! Zij X een continue stochastische variabele dan $P(X=a) = \int_{a}^{a} p = 0$ $(\neq P(a)$: geen kans)

→ een punt is geen kans : de oppervlakte moet je interpreteren als een kans

Zij (Ω,G,P) een kansruimte

$X:\Omega\to\mathbb{R}$ (: een continue stochastische variabele)

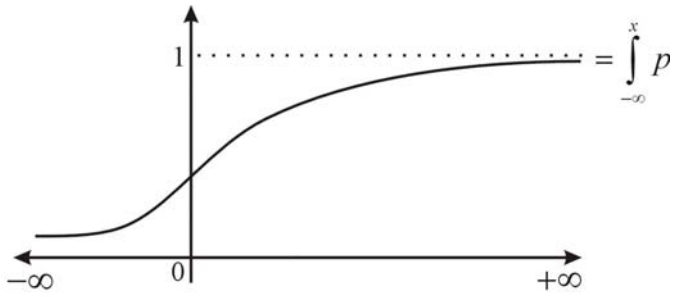
P is de uniforme verdeling over $[a,b]$



: als de boogjes even lang zijn dan zijn de kansen hetzelfde (gevolg van lukraak te trekken)

$X \sim U([a,b])$: X is uniform verdeeld over het interval $[a,b]$

We definiëren de verdeling $F : \mathbb{R} \to \mathbb{R} \ : X \to F(x) = P(X \leq x)$



Discreet = dichtheid is een kans	**Continu** = dichtheid : gebruiken van integralen
F is nergens dalend F is alleen <u>rechts</u> continu $F(-\infty) = 0$ $F(+\infty) = 1$	F is nergens dalend F is <u>overal</u> continu $F(-\infty) = 0$ $F(+\infty) = 1$

X is continu (C^{tu})

$F'(X) = p(x)$ voor elke X in $\mathbb{R}$

Bewijs: $F(X) = \int_{-\infty}^{x} p(x)\,dx$

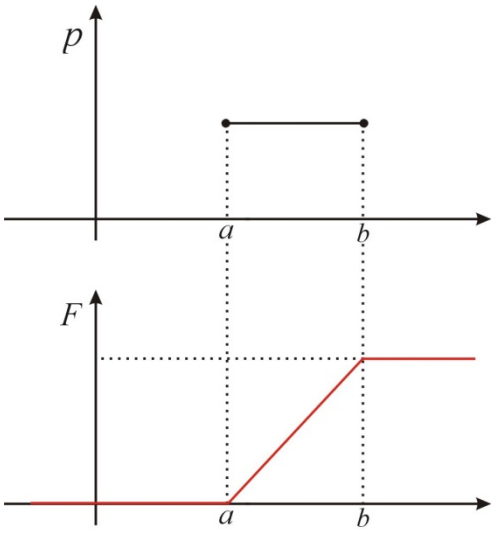
Hoofdstelling van integraal rekenen:

Zij f een continue f' van $\mathbb{R}$ naar $\mathbb{R}$, zij $F(x) = \int_{a}^{x} f$

Dan $(\int_{a}^{x} f)' = f(x)$

Dan $F'(x) = f(x)$

Uniforme verdeling : $X \sim U([a,b])$

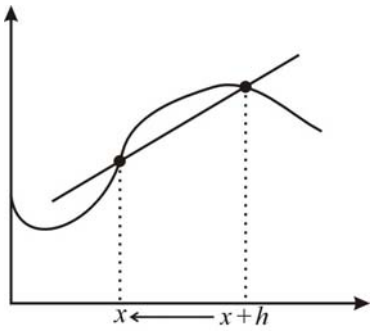


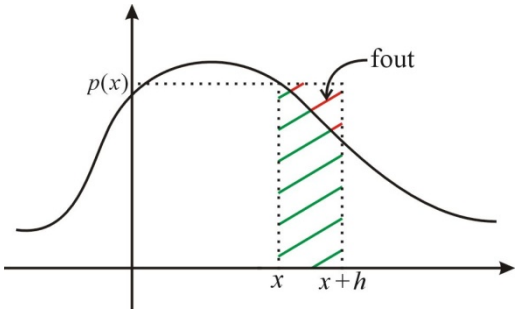
Wiskunde van vorig jaar : Wat is een afgeleide?

$$\lim_{h \to 0} \frac{F(x+h)-F(x)}{h}$$

$$\frac{F(x+h)-F(x)}{h} \approx p(x)$$

$$F(x+h)-F(x) \approx p(x).h$$





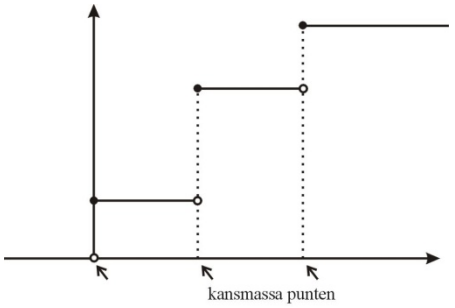
→ hoe kleiner h, hoe kleiner het fout

Merk op :

$\boxed{p(x \le X \le x+h) = P(x+h)-F(x) \approx p(x).h}$ (we gebruiken ≈ omdat er een fout van 'h' is)

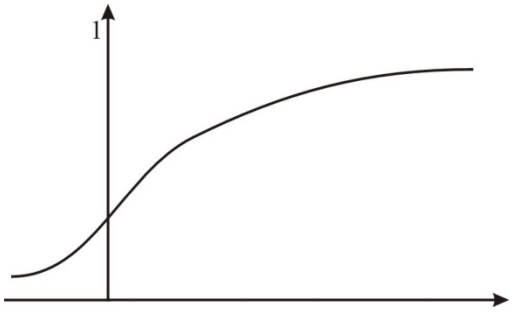
Stelling (geen bewijs) :
Zij F een functie van $\mathbb{R}$ naar $\mathbb{R}$
Dan is F een verdelingsfunctie als een slechts als: * F is rechts continu
* F is nergens dalend
* $F(-\infty) = 0$
* $F(+\infty) = 1$

X is discreet:

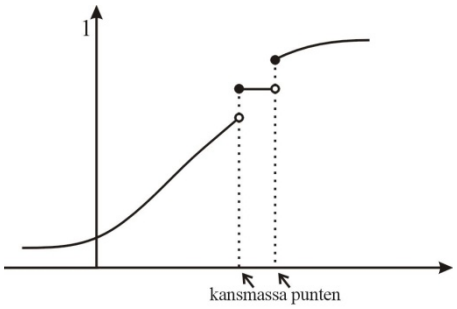


= geeft de cumulatieve verdeling van de stochast weer

X is continu:



Gemengde stochast: (voldoet aan alle eigenschappen)



Voorbeelden van stochastische variabelen

Uniforme verdeling:

$X \sim U([a,b])$: X is (lukraak) uniform verdeeld over $[a,b]$

$X \sim U(\{1,2,...,k\})$: X is uniform verdeeld over $\{1,2,3,...,k\}$

$$p_1 = p_2 = ... = p_k = 1/k$$

7.2.4 Bernoulli-verdeling (HB p119)

Een Bernoulli-experiment is een experiment met twee mogelijke uitkomsten, doorgaans succes (S) en mislukking (M) genoemd.

Een Bernoulli stochastische variabele met parameter p is de uitkomst X van een Bernoulli-experiment, met waarde 1 voor succes en 0 voor mislukking met een succeskans p.

$X : \Omega \to \{0,1\}$: een stochast met twee mogelijke uitkomsten

$p_0 = P(X = 0)$

$p_1 = P(X = 1) = q = 1 - p$

$p_0 + p_1 = 1$ (omdat er maar twee mogelijke uitkomsten zijn)

Zij $Y_1, Y_2, ..., Y_n$ een stochastische variabele met $Y \sim b(1, p)$

(= het experiment wordt n maal uitgevoerd)

Definieer $X = Y_1, Y_2, ..., Y_n$

→ algemene notatie : $\boxed{X \sim b(n, p)}$

n = hoeveel keer het Bernoulli-experiment wordt uitgevoerd

p = kans op succes

q = p^c = 1 - p = de kans op mislukking

---

* $X \sim b(1, p)$: Bernoulli stochastische variabele (st.v.)

* Bernoulli experiment wordt n keer onafhankelijk herhaald

(= alle experimenten staan los van elkaar)

$X_1, X_2, ..., X_n \sim b(1, p)$

$Y = X_1 + X_2 + ... + X_n$ (: het aantal successen optellen) $Y \sim b(n, p)$

vb. tos een munt ‘n’ keer (: er zijn twee mogelijke uitkomsten)
(* de eerste vraag die je moet stellen is of deze stochast discreet of continu is)

$Y \sim b(n, p)$
Y is discreet: Im(Y)$=\{0,1,2,...,n\}$
$\Omega = \left(\{\omega_1, \omega_2, ..., \omega_n\} | \omega_i \in \{0,1\}\right)$
$Y : \Omega \to \{0,1,2,...,n\}$
$\omega \to Y(\omega)$ = aantal ééntjes in ω

$$p_0 = P(Y=0) = P(X_1 = 0 \text{ en } X_2 = 0 \text{ en } X_n = 0)$$
$$= P(X_1 = 0) \text{ x } P(X_2 = 0) \text{ x ... x } P(X_n = 0)$$
$$= (1-p)^n$$

$$p_1 = P(Y=1) = P(X_1 + X_2 + ... + X_n = 1)$$
$$= \mathbf{n} \text{ x } P(X_1 = 1 \text{ en } X_2 = 0 \text{ en } X_3 = 0 \text{ en } X_4 = 0 \text{ en ... en } X_n = 0)$$
$$= n.p(1-p)^{n-1}$$

maal n : omdat je de 1 op n verschillende plaatsen kan zetten

$$p_2 = P(Y=2) = P(Y = X_1 + X_2 + ... + X_n = 2)$$
$$= \frac{n(n-1)}{2} \text{ x } P(X_1 = 1 \text{ en } X_2 = 1 \text{ en } X_3 = 0 \text{ en } X_4 = 0 \text{ en...en } X_n = 0)$$

delen door 2 om de dubbeltelling te vermijden

$$p_n = P(Y=n)$$

7.2.5 Binomium verdeling (HB p120)

Binomium coëfficiënt:
$\binom{n}{k}$: het aantal deelverzamelingen van lengte k getrokken uit een verzameling van lengte n.

n! = aantal manieren om n voorwerpen te ordenen
$= n \text{ x } (n-1) \text{ x } (n-2) \text{ x ... x } 2 \text{ x } 1$

Door het binomium coëfficiënt en n! te combineren:
$n! = \binom{n}{k} \text{ x } k! \text{ x } (n-k)!$
en dan te herschrijven:
$\boxed{\binom{n}{k} = \frac{n!}{k!(n-k)!}}$: in het formularium p3

$p_1 = np(1-p)^{n-1}$

$\binom{n}{1} = \frac{n!}{1!(n-1)!} = \frac{n \text{ x } (n-1)!}{(n-1)!} = n$

$p_2 = \binom{n}{2} = p^2(1-p)^{n-2}$

$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n \text{ x } (n-1) \text{ x } (n-2)!}{2!(n-2)!} = \frac{n(n-1)}{2!}$

$p_k = P(Y = X_1 + X_2 + ... + X_n = k)$

$= \binom{n}{k} p^k (1-p)^{n-k}$ voor elke k = 0, 1, ...n

→ dichtheid van de binomium verdeling: $\boxed{p_k = \binom{n}{k} p^k (1-p)^{n-k}}$: in het formularium p3

vb. voor k = 0 :

$\binom{n}{0} = \frac{n!}{n!0!} = 1$ $\left(\frac{n!}{k!(n-k)!} = \frac{n!}{0!(n-0)!} = \frac{n!}{0!n!}\right)$ Afspraak : $0! = 1$

in $Y \sim b(n, p)$ zit alle gegevens nodig om de stochast vast te leggen.
n = hoeveel keer het Bernoulli-experiment wordt uitgevoerd
p = kans op succes

Toetsen op de twee voorwaarden voor een dichtheid:

1. $p_k \geq 0$: OK

2. $p_1 + p_2 + ... + p_n = 1$

$= \sum_{k=0}^{n} P_k = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$: OK

(deze formule is hetzelfde als de binomium van Newton, zie verder.)

Binomium van Newton

$(a+b)^n = (a+b) \text{ x } (a+b) \text{ x } ... \text{ x } (a+b)$

$= 1a^n + n(ba^{n-1}) + \binom{n}{2} b^2 a^{n-2} + ... + \binom{n}{k} b^k a^{n-k} + ... + \binom{n}{n} b^n a^0$

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

(binominaal wordt soms als synoniem gebruikt voor Bernoulli)

$Y \sim b(n, p)$ = Bernoulli binominaal

vb1. $Y \sim b(3,0.7)$ (: Bernoulli-experiment 3 maal uitgevoerd met kans op succes 0,7)

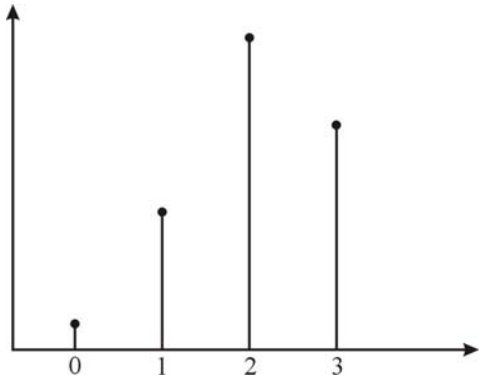
Kans dat je 3 keer mislukt : $P_0 = \binom{3}{0}(0,3)^3 = 0,027$

Kans dat de 2 keer mislukt en 1 keer slaagt : $P_1 = \binom{3}{1}(0,3)^2(0,7)^1 = 0,189$

Kans dat de 1 keer mislukt en 2 keer slaagt : $P_2 = \binom{3}{2}(0,3)^1(0,7)^2 = 0,441$

Kans dat je 3 keer slaagt : $P_3 = \binom{3}{3}(0,7)^3 = 0,343$

$\Sigma = 1$

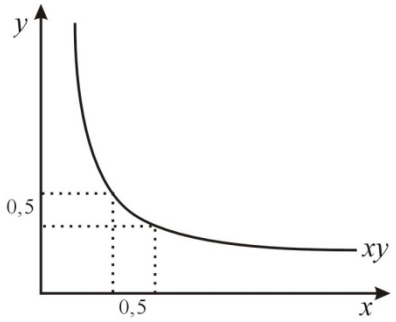


vb2.

Ω = Franse kiesrechtigde

$X : \Omega \to \{0,1\}$

$\omega \to X(\omega)$ = 0 indien Jospin
1 indien Chirac



$p = 0,52$: de kans op succes is 52%

$X \sim b(1,0.52)$: het Bernoulli-experiment wordt één keer uitgevoerd

We nemen nu n = 25 (we herhalen het experiment 25 maal) :
$Y \sim b(25,0.52)$

Wat is de kans dat minder dan of gelijk aan 5 mensen uit de 25 op Chirac stemmen?:

$P(Y \le 5) = P(Y = 0) + P(Y = 1) + ... + P(Y = 5)$

= kans 0 successen + kans 1 succes +...+ kans 5 successen

$= \binom{25}{0}0,48^{25} + \binom{25}{1}0,48^{24}0,52^1 + ... + \binom{25}{5}0,48^{20}0,52^5$

We ronden af naar boven om een uitkomst te bekomen :

$\le 0,50^{25}\left[\binom{25}{0} + \binom{25}{1} + ... + \binom{25}{5}\right]$

$= 0,50^{25}\left[1 + 25 + ... + 53130\right]$

$= \left(\frac{1}{2}\right)^{25} 68406 = \underline{0,002}$

$X \sim b(1, p)$

$X \sim b(n, p)$

Y_i is <u>multinominaal</u> = er zijn meer dan twee mogelijke uitkomsten

<u>vb.</u> Verkeerslichten : P_1 = groen , P_2 = oranje , P_3 = rood

	Kans	Trekkingen
Groen	p_1	n_1
Oranje	p_2	n_2
Rood	p_3	n_3
		n

n trekkingen = experiment n maal herhalen
$P(\mathrm{n}_1$ groen, n_2 oranje, n_3 rood)

$Z_1 =$ aantal keer groen
$Z_2 =$ aantal keer oranje
$Z_3 =$ aantal keer rood

$$Z_1 = \text{ aantal keer groen } = P(Z_1 = n_1 \text{ en } Z_2 = n_2 \text{ en } Z_3 = n_3)$$
$$= \binom{n}{n_1}\binom{n_2+n_3}{n_2} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

vereenvoudigen:

$$= \frac{n!}{n!(n_2+n_3)!} \text{ x } \frac{(n_2+n_3)!}{n_2!n_3!}$$
$$= \frac{n!}{n_1!n_2!n_3!}$$

<u>Y is multinominaal:</u>

Y is het aantal keer dat een Bernoulli experiment herhaald moet worden om een <u>eerste</u> keer succes te bekomen → **discrete stochast**

<u>vb.</u> hoeveel stemmen moeten we tellen om eerst een stem voor Chirac te bekomen?
hoeveel examens moet je afleggen om tot dat je er voor de eerste keer door bent?

$$\Omega = \left\{(\omega_1, \omega_2, ..., \omega_n, ...) \middle| \omega_1 \in \{0,1\}\right\}$$
$$Y : \Omega \to \mathbb{N}$$
$$\omega \to Y(\omega) \text{ = de plaats van de eerste 1-coordinaat}$$

~~p_0~~ (: schrappen om dat het geen uitkomst heeft)

$p_1 = P(Y=1) = p$

$p_2 = P(Y=2) = P(X_1 = 0 \text{ en } X_2 = 1) = P(X_1 = 0) \text{ x } P(X_2 = 1) = p(1-p) \geq 0$ (onafhankelijk)

⋮

$p_k = P(Y=k) = P(X_1 = X_2 = ... = X_{k-1} = 0 \text{ en } X_k = 1) = p(1-p)^{k-1} \geq 0$

$$= p_1 + p_2 + ... + p_k + ... = 1$$
$$= p + p(1-p) + ... + p(1-p)^{k-1} + ... = 1$$
$$= p\left[1 + (1-p) + (1-p)^2 + ... + (1+p)^{k-1} + ...\right]$$
$$= p\left(\frac{1}{1-(1-p)}\right) = 1$$

Let op : in economie werd dit gebruikt voor de consumptiequota (c).
dan had je $c\frac{1}{1-(1-c)} = 1$ (dus $(1-p) = c$)

Hypergeometrische Stochast

vb. een doos (Ω) bevat 4 witte knikkers en 5 rode knikkers:

$X : \Omega \to \{0,1\}$
$\omega \to X(\omega)$ = 0 als ω wit is
= 1 als ω rood is
X $\sim$ Bernoulli

Trek 5 knikkers (zonder teruglegging) uit een zak die 9 knikkers bevat waarvan er 4 'mislukkingen' zijn en 5 'successen' zijn.
Noteer het aantal successen met de stochastische variabele Y.

Vraag: Wat is de kans op twee successen?
$P(Y = 2) = p_2$

$$p_2 = P(0,0,0,1,1) = \left(\frac{4}{9}\right)\left(\frac{3}{8}\right)\left(\frac{2}{7}\right)\left(\frac{5}{6}\right)\left(\frac{4}{5}\right)$$

: kans 4 op 9 dat je mislukt maal kans 3 op 8 dat je mislukt.....(zonder teruglegging)
: dit is de kans voor de configuratie 3 keer wit en 2 keer rood te bekomen.
Om te veralgemenen moet je nog vermenigvuldigen met $\binom{5}{2}$:

$$p_2 = \binom{5}{2} \text{ x } P(0,0,0,1,1)$$

(: nu kan je alle volgorden toelaten vb. ook (1,0,0,1,0) etc..)

$$= \binom{5}{2} \text{ x } \left(\frac{4}{9}\right)\left(\frac{3}{8}\right)\left(\frac{2}{7}\right)\left(\frac{5}{6}\right)\left(\frac{4}{5}\right)$$

$$p_2 = \frac{\binom{5}{2}\binom{4}{3}}{\binom{9}{5}}$$

$\binom{5}{2}$: van de 5 successen dat er zijn moet je 2 successen trekken
$\binom{4}{3}$: uit de 4 mislukkingen moet je 3 mislukkingen trekken
$\binom{9}{5}$: uit de 9 knikkers in de populatie moet je er 5 trekken

$$p_2 = \frac{\binom{5}{2}\left(\frac{4!}{3!1!}\right)}{\left(\frac{9!}{5!4!}\right)} = \binom{5}{2} \text{ x } \frac{(4)(5!)}{\left(\frac{9!}{4!}\right)}$$

Formule : $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

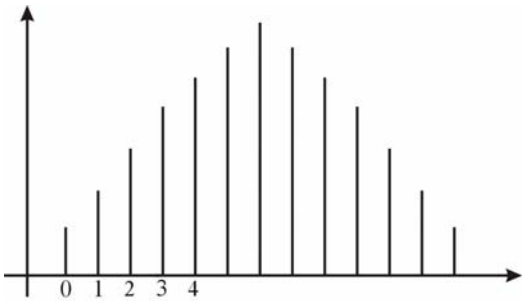
$X \sim hyp(n,a,b)$
n = aantal herhalingen
a = aantal successen in de populatie
b = aantal mislukkingen in de populatie
k = aantal successen waarvan je de kans wil weten

dichtheid hypergeometrische : $\boxed{p(k) = \binom{a}{k}\binom{b}{n-k} \Big/ \binom{a+b}{n}}$ (: in het formularium p3)

In dit voorbeeld:
$\boxed{Y \sim hyp(5;5,4)}$: 5 herhalingen en er zijn 5 successen en 4 mislukkingen

7.2.6 Poisson-verdeling (HB p125)

X = aantal ongevallen op een bepaald kruispunt per maand
= aantal telefoon oproepen in een bepaalde centrale per dag
= aantal bacterie in water per m^3
= aantal tik fouten per pagina
→ tel stochastische variabelen per volume



discrete stochastische variabele:
$p_k = P(X = k)$
$= \frac{\lambda^k e^{-\lambda}}{k!}$, k=0,1...

$\boxed{X \sim Pois(\lambda)}$

Let op :
in het HB gebruiken ze een andere notatie:
$p(k) = e^{-\mu}\frac{\mu^k}{k!}$

Is dit een dichtheid?
→ voldoet het aan de twee voorwaarden?

1. $p_k \geq 0$: OK

2. $\sum p_K = 1$: ?

$= p_0 + p_1 + p_2 + ... + p_k + ...$

$= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!}$

$= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$: OK

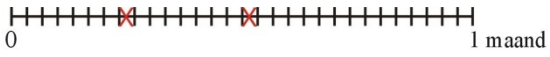
$0! = 1$

($\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}$ komt van : $e^x = 1 + X + \frac{X^2}{2!} + ... + \frac{X^k}{k!} + ... = \sum_{k=0}^{\infty}\frac{X^k}{k!}$)

Stelling : Poisson verdeling als benadering voor binomiale $b(n, p)$ voor n groot.

vb. X = aantal ongevallen op een bepaald kruispunt per jaar
Stel : 6 ongevallen per jaar; deze ongevallen zijn "uniform" over tijd verdeeld:
: elke dag heeft evenveel kans op een ongeval.
Succes : ongeval in maand

per maand : $X_1 \sim b(1;\frac{6}{12})$ = $X_1 \sim b(1;0,5)$



per halve maand : $X_2 \sim b(2;\frac{0,5}{2})$

per dag : $X_{30} \sim b(30;\frac{0,5}{30})$

per 'n' : $X_n \sim b(n;\frac{0,5}{n})$ (: als $n \to \infty$ zal er max één succes per periode zijn)

(uur, second...)

$X_n \to X$ met $X \sim Pois(0,5)$

$X_n \sim b(n;\frac{0,5}{n})$

$\boxed{X_n \sim b(n;\frac{\lambda}{n})}$: λ is de Poisson parameter

$$P(X_n = k) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \left(\frac{n(n-1)...(n-k)}{k!}\right)\left(\frac{\lambda^k}{n^k}\right)\left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \left(\frac{n(n-1)...(n-k)}{n\ n\ ...\ n}\right)\frac{\lambda^k}{k!}\left\{\left(1-\frac{\lambda}{n}\right)^n\right\}^{\frac{n-k}{n}}$$

↓ als n naar ∞ gaat

$$= 1 \text{ x } 1 \text{ x } 1 \text{ x } ... \text{ x } 1 \text{ x } \frac{\lambda^k}{k!}e^{-\lambda}$$

De uiteindelijke formule : $\boxed{P(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}}$ (= dichtheid van Poisson : formulariumpagina 3)

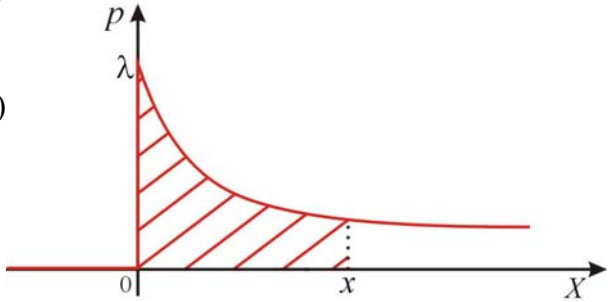
* $X \sim b(n, p)$ voor n groot $X \sim Pois(\lambda = p.n)$

$e^x = \lim_{x \to \infty}\left(1+\frac{x}{n}\right)^n$ voor $x = -\lambda$
(: in het formularium p4)

7.2.7 Exponentiële verdeling (HB p128)
→ voor continue stochastische variabelen vb. wachttijd of levensduur.

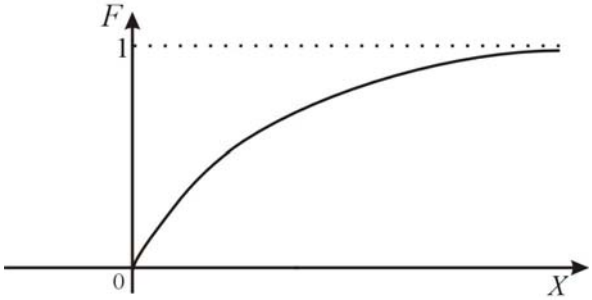
vb. X = tijd tussen 2 opeenvolgende telefoon oproepen
→ dit is een continue stochastische variabele

$P(X) = \lambda e^{-\lambda x}$ met $\lambda > 0$ (dichtheidsfunctie)
$P(X) = 0$ voor $X < 0$



Cumulatieve verdeling:

$$F(X) = P(X \le x) = 1 - e^{-\lambda x}$$
$$= \int_0^x p$$
$$= \int_0^x \lambda e^{-\lambda t}\ \mathrm{dt} = {}_0^x\left[-e^{-\lambda t}\right] = -e^{-\lambda t} + e^0$$



(we beginnen in 0 en eindigen in 1)

8. Gezamenlijke verdelingen en onafhankelijkheid (HB p135)
→ stochastische vectoren

(Ω, G, P) is een kansruimte
$(X, Y) : \Omega \to \mathbb{R}^2$
$\omega \to \big(X(\omega), Y(\omega)\big)$ (vb. lengte **en** gewicht van persoon ω)

vb. discreet
Ω is een bevolking
$X(\omega)$ = bloedgroep van persoon $\omega \in \{\overset{1}{O}, \overset{2}{A}, \overset{3}{B}, \overset{4}{AB}\}$

$Y(\omega)$ = rhesusfactor van persoon $\omega \in (\overset{1}{+}, \overset{2}{-})$

→ dit zijn gezamenlijke discrete stochastische variabelen

Kruistabel:

		Y	
		1	2
X	1	p_{11}	p_{12}
	2	p_{21}	p_{22}
	3	p_{31}	p_{32}
	4	p_{41}	p_{42}

p_{ij} = gezamenlijke discrete dichtheid van X en Y
$= P(X = i \text{ en } Y = j)$
$= P_{XY}(i, j)$ (: andere notatie)

De twee voorwaarden voor een dichtheid:
1. $P_{ij} \geq 0$: OK
2. $\sum_{i,j} P_{ij} = 1$: OK

$p_{XY}(2,1) = P(X = 2 \text{ en } Y = 1)$
$= p_{21}$ $(= A^+)$

$p_X(1) = P(X = 1)$: alle mensen met bloed groep 'O' ongeacht de rhesusfactor
$= P(X = 1 \text{ en } Y = 1) + P(X = 1 \text{ en } Y = 2)$
$= p_{11} + p_{12}$ (: zie kruistabel)

$p_X(2) = A = p_{21} + p_{22} = A^+ + A^-$
$p_X(3) = B = p_{31} + p_{32} = B^+ + B^-$
$p_X(4) = AB = p_{41} + p_{42} = AB^+ + AB^-$

* de marginale variabelen zijn de afzonderlijke variabelen X en Y
* de marginale dichtheden zijn de dichtheden van X en Y; p_X en p_Y.

marginale dichtheid van X:
$p_X : k \to p_X(k)$

Berekening : $p_X(x) = \sum_y p_{XY}(x, y)$
$p_Y(y) = \sum_x p_{XY}(x, y)$
$p_{XY}(x, y) = p_X(x).p_Y(y)$

gezamelijke dichtheid:
$p_{i,j} = P_{XY}(i, j)$

marginale dichtheid:
$p_X(i)$

voorwaardelijke dichtheid : $\left[P(A|B) \right]$: kans 'A' gegeven dat 'B'

$P_{X|Y=y}(x) = \frac{P_{X,Y}(x, y)}{p_Y(y)} = \frac{\text{gezamelijke dichtheid in (x,y)}}{\text{marginale dichtheid in y}}$

$P_{X|Y=1}$: $1 \to P(X=1|Y=1) = p_{11}/p_{11}+p_{21}+p_{31}+p_{41} \geq 0$

$2 \to P(X=2|Y=1) = p_{21}/p_{11}+p_{21}+p_{31}+p_{41} \geq 0$

$3 \to P(X=3|Y=1) = p_{31}/p_{11}+p_{21}+p_{31}+p_{41} \geq 0$

$4 \to P(X=4|Y=1) = p_{41}/p_{11}+p_{21}+p_{31}+p_{41} \geq 0$

Formule:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Stochastische vectoren $(X,Y): \Omega \to \mathbb{R}^2$ **<u>continu</u>**

marginale dichtheid : $p_X(x) = \int_{-\infty}^{+\infty} p_{XY}(x,y)\,dy$

voorwaardelijke dichtheid : $P_{X|Y=y}(x)$

<u>gezamelijke dichtheid</u> : $p_{XY} = \mathbb{R}^2 \to \mathbb{R}$

Voorwaarden: $p_{xy} \geq 0$ en $\int\int p_{xy} = 1$ zijn voldaan

$P((x,y) \in \mathbb{R}) = P\left(\{\omega | X(\omega), Y(\omega)) \in R\}\right)$

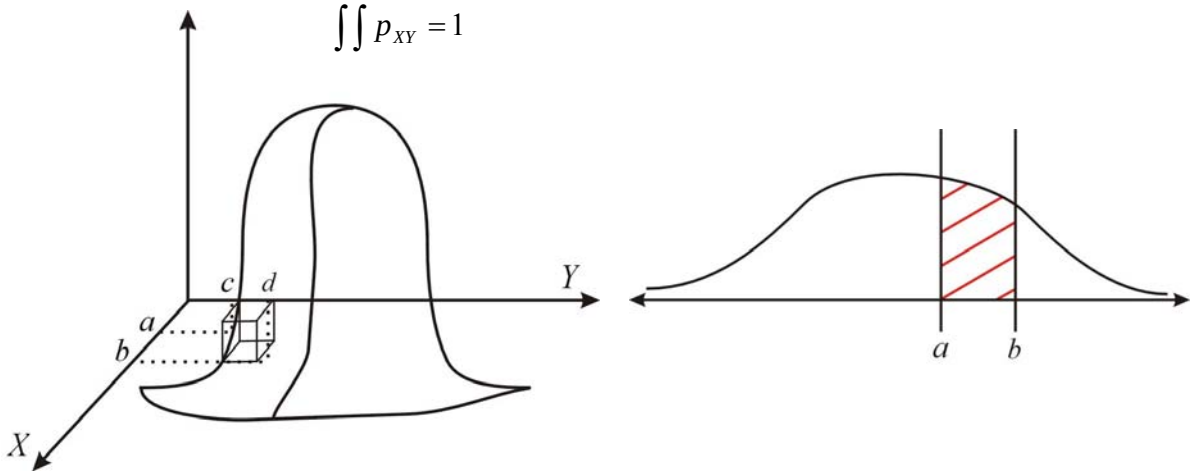
$= \int_R \int p_{XY}$ voor elke rechthoek R in $\mathbb{R}^2$

stel $R = [a,b]$ x $[c,d]$: volume onder de dichtheids oppervlakte A

$= \int_{x=a}^{b} \int_{y=c}^{d} p_{XY}$: dubbele integraal

Volume onder de figuur is gelijk aan 1:

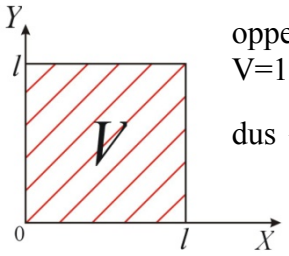
$\int\int p_{XY} = 1$



<u>vb1.</u> Kies lukraak een punt uit $V = [0,l]$ x $[0,l]$

<u>gezamelijke dichtheid</u>: $P_{XY}(x,y) = 1/l^2$ indien $(x,y) \in V$

(uniform in V) $= 0$ indien $(x,y) \notin V$



oppervlakte = l^2

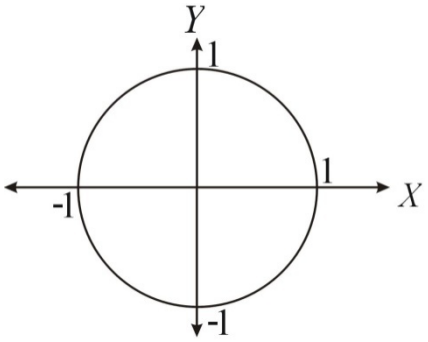
V=1

dus $\frac{l^2}{l^2} = 1$

marginale dichtheid : we zijn enkel geïnteresseerd in X (a en b) en niet in Y (c en d)

$P_X(x) = 0$ indien $x \notin [0,l]$

uniform in $[0,l]$ $\qquad \int_{y=-\infty}^{+\infty} P_{XY}(x,y) = \int_{y=0}^{l} \frac{1}{l^2} \text{ dy} = \frac{1}{l}$



vb2. kies lukraak een punt uit een schijf 'S' met straal 1

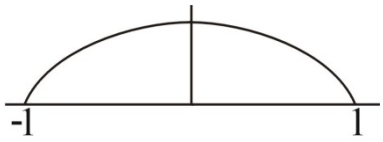
$S = \{(x,y) | x^2 + y^2 \leq 1\}$: oppervlakte schijf met straal 1

gezamelijke dichtheid : $P_{XY}(x,y) = 1/\pi$

$= 0$ indien $(x,y) \in S$

marginale dichtheid : $P_X(x) = \int_{y=-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} \frac{1}{\pi} \text{ dy} = \frac{2\sqrt{1-x^2}}{\pi}$

$= 0$ indien $X \notin [-1,1]$



* We kunnen de marginale dichtheid zien als een oppervlakte onder een bepaalde curve.

$P_X(x) = \int_{y=-\infty}^{+\infty} P_{XY}(x,y) \text{ dy}$

Cumulatieve verdeling (cumulatieve frequentiefunctie)

$F_{XY}(x,y) = P(X \leq x \text{ en } Y \leq y) = \int_{y=-\infty}^{y} \int_{t=-\infty}^{x} P_{XY}(t,s) \text{ dt ds}$

Bewijs : $\frac{\partial}{\partial x} . \frac{\partial}{\partial y} F_{XY}(x,y) = P_{XY}(x,y)$

$F_X(x) = P(X \leq x)$

$= P(X \leq x \text{ en } Y \leq +\infty)$

$= F_{XY}(x, +\infty)$

$F_X'(x) = P_X(x)$

Twee variabelen zijn onafhankelijk indien; 3 voorwaarden: (HB p147)
vb. is de bloedgroepverdeling onafhankelijk van de rhesusfactor?

1. Stelling: Zij $(X,Y) : \Omega \to \mathbb{R}^2$ een stochastische vector
X en Y zijn onafhankelijk als en slechts als voor elke paar A en B van intervallen;

$P(X \in A \text{ en } Y \in B) = P(X \in A).P(Y \in B)$

[m.aw de gebeurtenissen $\{\omega | X(\omega) \in A\}$ en $\{\omega | X(\omega) \in B\}$ zijn onafhankelijk van elke A en B]
[dit is hetzelfde als: indien onafhankelijk: $P(A \cap B) = P(A)$ x $P(B)$]

2. Stelling: stochasten X en Y zijn onafhankelijk als en slechts als;

$\boxed{p_{XY} = p_x . p_y}$

→ de gezamelijke dichtheid is het product der marginalen

Bewijs : (X en Y zijn discreet)

$\downarrow \quad P_{XY}(x,y) = P(X = x \text{ en } Y = y)$

onafhankelijkheid $= P(X = x).P(Y = y)$

$= p_X(x).p_Y(y)$ (voor elke x,y)

$\uparrow \; P(X \in A \text{ en } Y \in B) = \sum_{\substack{X \in A \\ Y \in B}} P(X = x, Y = y)$: gezamelijke dichtheid

$= \sum_{\substack{X \in A \\ Y \in B}} p_X(X = x).p_Y(Y = y)$

$= \sum_{X \in A} p_X(X = x).\sum_{Y \in B} p_Y(Y = y)$

$= P(X \in A).P(Y \in B)$

→ X en Y zijn onafhankelijk

3. Stelling: X en Y zijn onafhankelijk als en slechts als;

$\boxed{\forall_{x,y} : p_{X|Y=y}(x) = p_{X|Y=y'}(x)}$

Bewijs: $\downarrow \; p_{X|Y=y}(x) = P(X = x|Y = y)$

$= \frac{P(X = x \text{ en } Y = y)}{P(Y = y)}$

$= \frac{p_{XY}(x,y)}{p_Y(y)}$: definitie gezamenlijke en marginale dichtheid

$= \frac{p_X(x).p_Y(y)}{p_Y(y)} = p_X(x)$: en is dus onafhankelijk van y

9. Verwachtingswaarden (verwachte waarde) (HB p149)

Voorbeeld :
Spel : twee keer een eerlijke dobbelsteen tossen.
Winstregel : als 2 zessen dan €1000
als 1 zes dan €10
anders dan €0

Vraag: Hoeveel ben je bereid te betalen om één keer mee te spelen?

Bereid te betalen : P(€1000 winst) x 1000 + P(€10 winst) x 10 = €30,60

= P(2 zessen) = P(1 zes)

= 1/36 = 10/36

De berekening:

$\Omega = \{(1,1),(1,2),...,(6,6)\}$

$|\Omega| = 36$: er zijn 36 mogelijke uitkomsten

10 combinaties met 1 zes : (1,6), (2,6), (3,6), (4,6), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5)
1 combinatie met 2 zessen : (6,6)

De kansen zijn uniform verdeeld :

$X : \Omega \to \mathbb{R}^{(0,1,2)} \to \mathbb{R}$

$\omega \to X(\omega) =$ aantal zessen in ω

$p(0) = P(0 \text{ zessen}) = 25/36$	$0 \to 0$	$\pi(0) = 0$
$p(1) = P(1 \text{ zes}) = 10/36$	$1 \to 10$	$\pi(1) = 10$
$p(2) = P(2 \text{ zessen}) = 1/36$	$2 \to 1000$	$\pi(2) = 1000$

$E(\pi(x)) =$ verwachte waarde van het spel (expected value) = verwachte winst

$= \pi(0).P(X=0) + \pi(1).P(X=1) + \pi(2).P(X=2)$

$= \left(0.{}^{25}\!/_{36}\right) + \left(10.{}^{10}\!/_{36}\right) + \left(1000.{}^{1}\!/_{36}\right)$

$= 30{,}60$

Definitie: Zij $X : \Omega \to \mathbb{R}$ een stochastische variabele en (Ω, G, P) een kansruimte
Zij $g : \mathbb{R} \to \mathbb{R}$ een afbeelding

De verwachtingswaarde $E[g(x)] =$

- $\sum_k g(k)P(X=k)$: discreet
- $\int_{-\infty}^{+\infty} g(x).p(x)\,dx$: continu

Opmerking: Indien $g : \mathbb{R} \to \mathbb{R} : X \to x$ (zichzelf afbeeldt)
dan $\mu = E(x) = E(g(x))$: de verwachte waarde van X

Verwachte waarde: $EX =$

- $\sum_x x.p(x)$ als x discreet is
- $\int_{-\infty}^{+\infty} x.p(x)\,dx$ als x continu is

De verwachtingswaarde: $E[g(x)] =$

- $\sum_x g(x).p(x)$ discreet geval
- $\int_{-\infty}^{+\infty} g(x).p(x)\,dx$ continu geval

Verwachtingswaarde voor een continue stochast

(Ω, G, P) is een kansruimte

$\Omega \xrightarrow{X} \mathbb{R} \xrightarrow{g} \mathbb{R}$

$$E\big(g(x)\big) = \int_{-\infty}^{+\infty} g(x).p(x)\ \text{dx} \quad \text{(continu)}$$
$$= \sum_{k} g(k).p(k) \quad \text{(discreet)}$$

NB. als $S = \Omega$:
$\overline{x} = \mu$
$\tilde{s}_{x,y} = \sigma_{x,y}$
$\tilde{s}^2 = \sigma^2$

Speciale gevallen (HB p154)

Gemiddelde van x = $\mu = E(x)$
- $\sum_{x} x.p(x)$ voor x discreet
- $\int_{-\infty}^{+\infty} x.p(x)\ \text{dx}$ voor x continu

variantie van x = var x = $\sigma^2 = E\left[(x-\mu)^2\right]$
- $= \sum_{k} (k-\mu)^2.p(k)$ discrete stochast
- $= \int_{\mathbb{R}} (x-\mu)^2.p(x)\ \text{dx}$ continue stochast

Variantie: $\sigma^2 = \mu_2 = E(X-\mu)^2 = \mu'_2 - (\mu'_1)^2 = E(X^2) - \big(E(X)\big)^2$

Geometrische verdeling

$\boxed{X \sim Geo(p)}$

→ wanneer heb je voor de eerste keer succes? vb. na hoeveel worpen?

$p_k = (1-p)^{k-1}.p = q^{k-1}p$ (: dichtheid Geometrische verdeling : formularium p3)

$q = 1 - p$

k = aantal worpen dat je moet doen voor de eerste keer succes

$$\mu = E(k)$$
$$= \sum_{k=1}^{\infty} k.p(k)$$
$$= \sum_{k=1}^{\infty} kq^{k-1}p$$
$$= p\sum_{k=1}^{\infty} kq^{k-1} = p.\frac{1}{p^2} = \frac{1}{p}$$

Dus $E(k) = \frac{1}{p}$

vb. dobbelsteen $P(Y=6)=p(x)=1/6$

$E(X)=6$

→ je hebt een 1/6 kans dat je een 6 smijt dus $p=1/6$

Bewijs: (: in het formularium p4)

$$1+q+q^2+q^3+q^4+\ldots=\frac{1}{1-q}=(1-q)^{-1} \text{ met } 0\le q\le 1$$

↓ afleiden

$$1+2q+3q^2+4q^3+\ldots=1.(1-q)^{-2}=\frac{1}{(1-q)^2}=\frac{1}{p^2}$$

$$E(X^2)=\sum_{k=1}^{\infty}k^2q^{k-1}p$$

$$=p\sum_{k=1}^{\infty}k^2q^{k-1}$$

$$=p(1+4q+9q^2+16q^3+\ldots)$$

vermenigvuldigen met q
en dan afleiden

Dus: $q+2q^2+3q^3+4q^4+\ldots=\frac{q}{(1-q)^2}=\frac{1}{p^2}$

$$1+4q+9q^2+16q^3+\ldots=\frac{(1-q)^2-2(1-q).(-1).q}{(1-q)^4}=\frac{1-q+2q}{(1-q)^3}=\frac{1-q}{(1-q)^3}=\frac{1+q}{p^3}$$

Dus: $E(x^2)=p.\frac{1+q}{p^3}=\frac{1+q}{p^2}$

$$\sigma^2=E\left[(x-\mu)^2\right]$$

$$=E\left[x^2-2\mu x+\mu^2\right]$$

Integraal van de som: (eigenschappen HB p150)
$E(X+Y)=E(X)+E(Y)$

Dus: $\sigma^2=E\left[x^2\right]-E\left[2\mu x\right]+E\left[\mu^2\right]$

$$=E\left[x^2\right]-2\mu E\left[x\right]+\mu^2$$

$$=E\left[x^2\right]-2\mu^2+\mu^2$$

$$=E\left[x^2\right]-\mu^2$$

$$=E\left[x^2\right]-\left(E(x)\right)^2$$

$$=\frac{1+q}{p^2}-\frac{1}{p^2}$$ (: zie de vorige pagina)

$$\boxed{\sigma^2=\frac{q}{p^2}}$$

vb. Dobbelsteen $X \sim Geo\left(\frac{1}{6}\right)$ (: $p = 1/6$ = de kans dat je een zes gooit; uniform verdeeld)

$\mu = 6$

$\sigma^2 = \frac{q}{p^2} = \frac{1-p}{p^2} = \frac{5/6}{1/36} = \frac{5}{6}.\frac{36}{1} = 30$

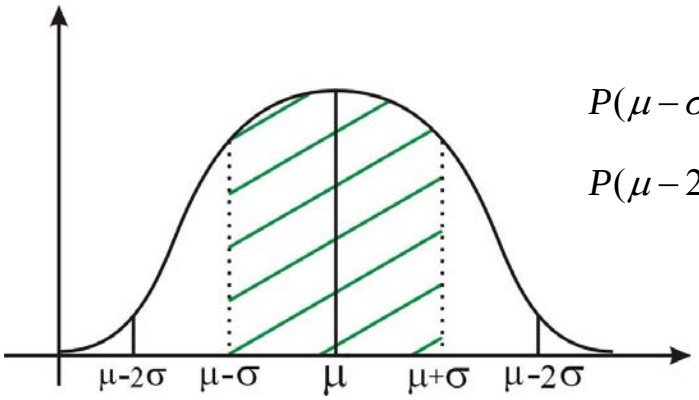
* stochastische variabelen: locatie (μ) en spreiding (σ^2) zijn de twee belangrijkste maten.

Stelling van Tchebychev (HB p162): bij symmetrische intervallen

Zij (Ω, G, P) een kansruimte

$X : \Omega \to \mathbb{R}$, μ, σ^2

dan $P(\mu - k\sigma < x < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$



$P(\mu - \sigma < x < \mu + \sigma) \geq 0$

$P(\mu - 2\sigma < x < \mu + 2\sigma) \geq \frac{3}{4}$

Beslissingsproblemen : Mean-variance problemen : Payoff matrix (HB p153)

Gevraagd: zoek de kritische kans waarbij beide acties een gelijke gemiddelde payoff bieden.

Payoff matrix van een spel:

Economische toestand: / Actie van bedrijf:	Expansie θ	Stagnatie $1-\theta$	← kans
a1	$\pi = 100$	$\pi = -20$	
a2	$\pi = 60$	$\pi = -10$	

Analoog aan Nash evenwicht zoeken:

$E(\pi|a_1) = 100\theta - 20(1-\theta)$

$= 120\theta - 20$

$E(\pi|a_2) = 60\theta - 10(1-\theta)$

$= 70\theta - 10$

$$E\left(\pi|a_1\right) \overset{?}{><} E\left(\pi|a_2\right)$$
$$120\theta - 20 >< 70\theta - 10$$
$$50\theta >< 10$$
$$\theta >< 1/5$$

<u>Besluit:</u> kies a_1 indien $\theta \in [0,2;1]$

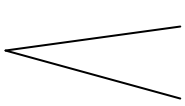
[Merk op: we veronderstellen dat het bedrijf geen invloed heeft op θ : dat θ exogeen is.]

* Stel $\theta = 1/5$ dan is de verwachte winst van beide acties hetzelfde.

Op basis van de verwachte waarde alleen kan je dus <u>niet</u> beslissen. We kijken wel naar de spreiding $\sigma^2\ E\left(\pi|a_1\right) = E\left(\pi|a_2\right)$
→ bijkomende informatie:
op basis van $\sigma^2\left(\pi|a_1\right) \rightarrow\ \sigma^2\left(\pi|a_2\right)$

<u>Beslissingsproblemen - testen van bloedstalen</u> (HB p151)

n = aantal bloedstalen dat worden getest ('n' is groot)
p = kans op besmetting in het bloedstaal ('p' is klein)



individuele testen : kost prijs $= n$ x c (: c is de kost van één test)

gegroepeerde testen = $n = m$ x k (: m groepen van k stalen)

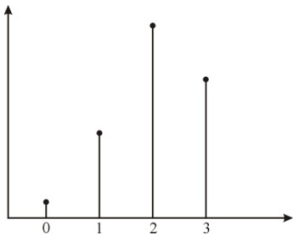
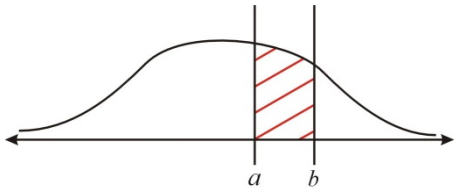
Kost prijs (tweede strategie) = $(m \text{ x } c) + \sum (k \text{ x } c)$
(: de besmette groepen moeten opnieuw individueel worden getest)

Dus: "a-priori" expected value van de kost prijs
Belangrijkste stap : welke stochasten gaan we definiëren?
We weten dus niet of een groep besmet is of niet en indien wel heeft het extra tests nodig.

$X_1, X_2, ..., X_n$
X_i = aantal analyses in groepje i → 1 als groepje i <u>niet</u> besmet is
→ 1+k als groepje i <u>wel</u> besmet is

kans geen besmette stalen in het groepje : $p_1 = (1-p)^k$
(: $1-p$ is de kans voor 1 persoon, wij willen de kans weten voor k personen)

kans wel besmette stalen in het groepje : $p_{k+1} = 1-(1-p)^k$
→ de kansen tellen op tot 1

Discrete stochastische variabele	Continue stochastische variabele
$p:\{0,1,...,n\}\to\mathbb{R}$ $k\to p_k \quad (=p(k))$	$p:\mathbb{R}\to\mathbb{R}$ $x\to p(x)$
	

$$E(X)=\sum_X x.p(x)\ :\ E(X_i)=\overset{\text{niet besmet}}{1.}\ (1-p)^k+\overset{\text{besmet}}{(k+1)}(1-(1-p)^k)$$

$$=k+1-k(1-p)^k \quad (:\text{ we zien dat het niet afhangt van i})$$

E is lineair

$$E(x_1+x_2+...+x_m)=E(x_1)+E(x_2)+...+E(x_m)$$

$$=m.E(x_i)$$

$$=m(k+1-k(1-p)^k)$$

$m=\frac{n}{k}$: $\frac{\text{totaal aantal testen}}{\text{aantal stalen}}=$ m groepen

Expected value van de kost prijs $=m(k+1-k(1-p)^k).c$

$$=\frac{n}{k}(k+1-k(1-p)^k)$$

$\frac{1}{n}.E(x_1+x_2+...+x_n)$: het verwachte aantal testen onder gegroepeerde testen als percent van het aantal testen onder individuele testen

$$=\frac{1}{n}.m(k+1-k(1-p)^k) \qquad (:\frac{1}{n}.m=\frac{1}{k})$$

$$=\frac{1}{k}(k+1-k(1-p)^k)$$

$$=1+\frac{1}{k}-(1-p)^k$$

Rekenvoorbeeld:

$p=0,01$: kans op besmetting = 1/100

$k=10$: groepjes van 10

We vullen deze getallen in de formule van hier boven:

$$=1+\frac{1}{k}-(1-p)^k=1+\frac{1}{10}-(0,99)^{10}\approx \underline{19,56\%}$$

$$\min_k\text{aliseer}\ :\ \frac{E(x_1+...+x_m)}{n}$$

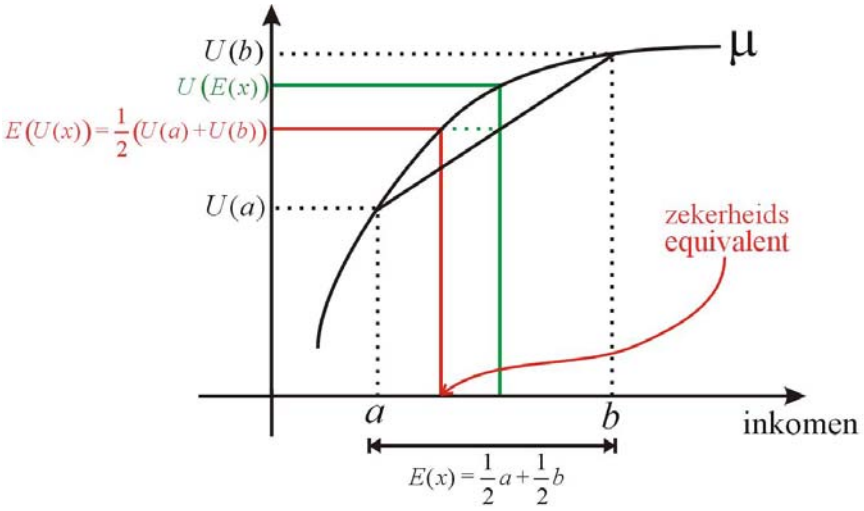
* p en n zijn exogeen (niet beïnvloedbaar)

vb.

individu → werkloos, inkomen a, kans p
individu → werkend, inkomen b, kans 1-p

Vraag: hoe beoordelen?
nutsfunctie : $U : \mathbb{R} \to \mathbb{R}$
discrete stochast : X = inkomen a, kans p
= inkomen b, kans 1-p

$E(X) = p.a + (1-p).b$ → verwachte waarde
$E[U(x)] = p.u(a) + (1-p).U(b)$ → verwachtingswaarde



Hier zien we goed dat $U\left(E(inkomen)\right) \geq E\left(U(inkomen)\right)$

→ Indien de functie U negatieve kromming heeft
dan $E(U(x)) \leq U(E(x))$
: ongelijkheid heeft te maken met de kromming

Ongelijkheid van Jensen

Zij $f : \mathbb{R} \to \mathbb{R}$ convex (: negatieve kromming zoals de nutsfunctie : 2de afgeleide is negatief)
Zij X een stochastische variabele met $E(X)$ in $\mathbb{R}$
Dan $E\left(f(x)\right) \leq f\left(E(X)\right)$

Bewijs: (X is een continue stochastische variabele en f is twee keer afleidbaar)

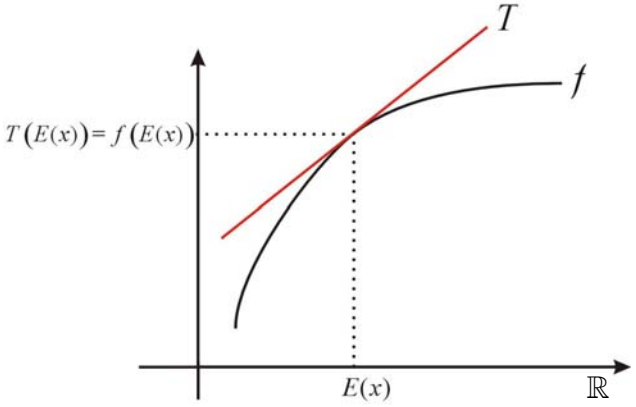
T is een eerste graadsfunctie dus; $ax + b = T(x)$

$f(x) \le T(x)$

$P_x(x) f(x) \le P_x(x) T(x)$

$\int p_x(x) f(x) \le \int p_x(x) T(x)$

De grafiek ligt onder de raak lijn:



Verwachtingswaarde:
(integreren bewaart de ongelijkheid)

$E(f(x)) \le \int p_x(x)(ax+b)$

$E(f(x)) \le a\int x p_x(x) + b\int p_x(x)$

$E(f(x)) \le aE(x) + b(1)$

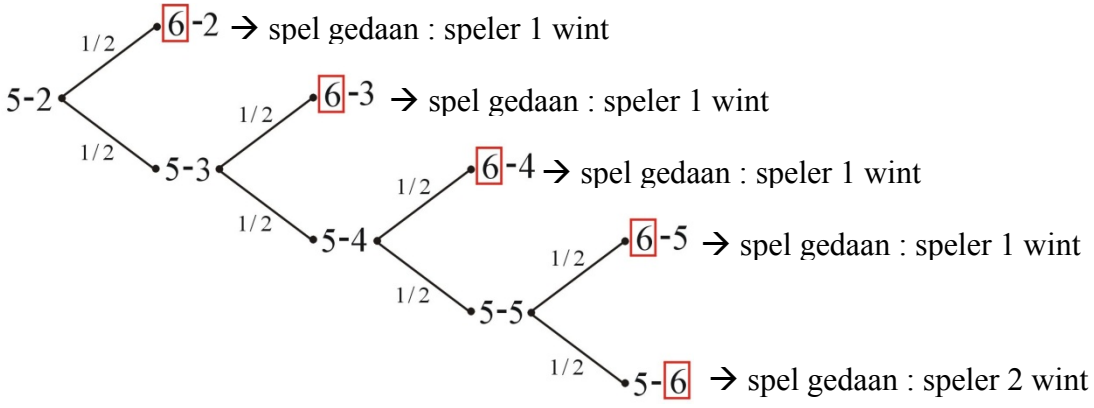
$E(f(x)) \le T(E(x))$

$E(f(x)) \le f(E(x))$

vb: Tornooi tussen 2 spelers: "de beste van 11 herhalingen wint" (wie eerst 6 haalt wint)
Het spel wordt afgebroken bij een stand 5-2

Vraag: hoe de prijs verdelen?

Antwoord: Kansboom : de stand is 5-2, hoe zal het spel verder lopen?



$P(\text{speler 2 wint}) : P(5-6) = \left(\frac{1}{2}\right)^4 = \frac{1}{16}$

$P(\text{speler 1 wint}) : P^C(2de) = \left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4 = \frac{15}{16}$

$X : \Omega \to \mathbb{R}$

$\mu = E(x)$

variantie : var(x) : $\boxed{\sigma^2 = E\left[(x-\mu)^2\right] = E(x^2) + \left(E(x)\right)^2}$

Stel: stochast X is uniform verdeeld over het interval $[a,b]$:

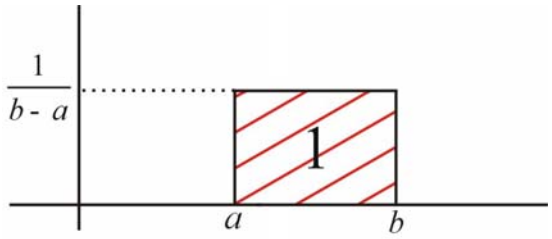
→ $X \sim U[a,b]$ (continu)

$$E(x) = \int_a^b \frac{1}{b-a} X \text{ dt}$$

$$= \frac{1}{b-a} \int_a^b X$$

$$= \frac{1}{b-a}\left(\frac{1}{2}\right)\left(b^2 - a^2\right)$$

$$= \frac{a+b}{2} = Y$$



Stel: $X \sim Pois(\lambda)$ (discrete stochast) $\quad p_k = e^{-\lambda}\frac{\lambda^k}{k!}$

$E(X) = \sum_0^\infty kP_k$ (: definitie verwachte waarde)

$$= \sum_0^\infty ke^{-\lambda}\left(\frac{\lambda^k}{k!}\right)$$ (: i.p.v. te sommeren vanaf 0 gaan we sommeren vanaf 1)

$$= \sum_{k=1}^\infty ke^{-\lambda}\left(\frac{\lambda^{k-1}\lambda}{k(k-1)!}\right)$$

$$= \sum_{k=1}^\infty e^{-\lambda}\left(\frac{\lambda^{k-1}\lambda}{(k-1)!}\right)$$

$$= e^{-\lambda}\lambda\sum_{k=1}^\infty\left(\frac{\lambda^{k-1}}{(k-1)!}\right)$$

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

$E(X) = \mu = \lambda$

Formularium :

$$1 + c + c^2 + \ldots = \frac{1}{1+c}$$

$$e^t = 1 + t + \frac{t^2}{2!} + \ldots + \frac{t^n}{n!}$$

Eigenschappen van de variantie (HB p157)

$X : \Omega \to \mathbb{R}$

$\mu = E(X)$

$\text{var}(X) = \sigma^2 = E\left[(x-\mu)^2\right] = E(x^2) - (E(X))^2$

Schaalgevoeligheid van de variantie:

a) $\text{var}(aX) = a^2 \text{ var } X$

$E\left[(ax - a\mu)^2\right] = E(a^2(x - Ex)^2) = E(a^2(x-\mu)^2) = a^2E((x-\mu)^2)$

b) $E(x+b) = E(x) + b$

$\text{var}(X+b) = \text{var}(X)$

* k-de ruwe moment : $E(X^k)$

$\mu = E(x)$

$E(x^2)$

.....

$E(x^k)$: k-de ruwe moment

* k-de centrale moment : $E\left[(x-\mu)^k\right]$

$E(x) - E(\mu) = E((x-\mu)) = 0$

$\mu \quad \mu$

1ste centrale moment is gelijk aan 0

$E\left[(x-\mu)^2\right]$

.....

$E\left[(x-\mu)^k\right]$: k-de centrale moment

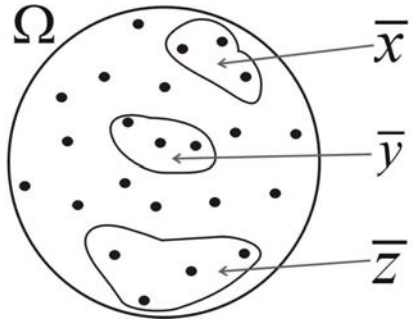
Momentgenererende functies (Mgf)

Motivatie 1: Indien X en Y twee stochastische variabelen zijn met dezelfde ruwe momenten (k=1,2,3,4...). Dan zijn X en Y ongeveer gelijk.
→ handmaten aan de hand van de momentgenererende functie

Motivatie 2: $X : \Omega \to \mathbb{R}$

$\omega \to X(\omega)$: lengte (ω)

$E(x) = \mu \in \mathbb{R}$: μ is het reëel gemiddelde van de hele populatie



: observaties van de stochastische variabele

$\overline{x} = \frac{x_1 + ... + x_n}{n}$ ← verdeling van $\overline{x}$?

← via momentgenererende functie

Definitie: (Ω, G, P)

$X : \Omega \to \mathbb{R}$: een stochastische variabele

$M_X : \mathbb{R} \to \mathbb{R}$

$$t \to M_x(t) = E(e^{tx}) = \begin{cases} \sum e^{tk}\, p(k) & \text{: discreet (: definitie verwachtings waarde)} \\ \int e^{tx}\, p(x)\ \mathrm{dx} & \text{: continu} \end{cases}$$

Eigenschap van motivatie 1:

$E(X^k) = \frac{d^k}{dt^k} M_x(t)$ (: k keer afleiden) of $E(x^k) = M_X^k(t)$

Bewijs:

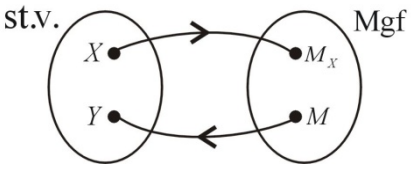
$$M_X(t) = M(0) + M'(0)t + M''(0)\frac{t^2}{2!} + \ldots + M^k(0)\frac{t^k}{k!} + \ldots \quad (: \text{Taylor reeks})$$

$$= E(e^{tx})$$

$$= E\left(1 + tx + \frac{(tx)^2}{2!} + \ldots + \frac{(tx)^k}{k!} + \ldots\right)$$

$$= E(1) + tE(X) + \frac{t^2}{2!}E(X^2) + \ldots + \frac{t^k}{k!}E(X^k) + \ldots$$

Eigenschap: (geen bewijs)
één-één verband tussen momentgenererende functie (Mgf) en de stochastische variabele:



Momentgenererende functies:

motivatie < momentum
transformatie van een stochastische variabele

vb. $X \sim b(n, p)$: is discreet (HB p160)

$$M_X(t) = E(e^{tx})$$

$$= \sum_{k=0}^{n} e^{tk}\, pk \qquad \boxed{pk = \binom{n}{k} p^k q^{n-k}}$$

$$= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k q^{n-k} \qquad q = 1 - p$$

$$= \sum_{k=0}^{n} \binom{n}{k}\binom{t}{ep}^k q^{n-k}$$

$M_X(t) = (pe^t + q)^n$: in het formularium

Binomium van Newton :
$$(a+b)^n = \sum_{0}^{n} \binom{n}{k} a^k b^{n-k}$$

Dus : $E(X) = M'_X(t)$ (: 1ste ruwe moment)

$$= n(pe^t + q)^{n-1}.(pe^t)$$

$$= n(p+q)^{n-1}.p \qquad (\ p + q = 1 \text{ want het zijn complimenten})$$

$$= np$$

$E(X^2) = M''(t)$ (: 2de ruwe moment)

$$= n(n-1)(pe^t + q)^{n-2}.(pe^t)^2 + n(pe^t + q).pe^t$$

$$= n(n-1)p^2 + np$$

$$\text{var}(X) = E(X^2) - (E(X))^2$$
$$= n(n-1)p^2 + np - n^2p^2$$
$$= np(-p+1)$$
$$= npq$$ (: n = hoeveel keer het experiment herhaald werd)

Oefening 1: Bernoulli verdeling

- $X \sim b(n, p)$

  ($M_X(t) = (pe^t + q)^n$)
- $Y \sim b(m, p)$
- X en Y zijn onafhankelijk

Te bewijzen: $X + Y \sim ?$

Intuïtief: $X + Y \sim b(m+n, p)$
→ je herhaalt n keer het Bernouilli experiment en daarna herhaal je m keer een ander Bernouilli experiment, onafhankelijk van elkaar.

Hulpstelling (bewijs):

Zij $(X,Y) : \Omega \to \mathbb{R}^2$ een stochastische variabele met verdeling $P_{X,Y}$
Zij X en Y onafhankelijk
Dan $M_{X+Y}(t) = E(e^{t(x+y)})$: definitie verwachte waarde

$$= \int_x \int_y e^{t(x+y)} p_{X,Y}(x,y)$$
$$= \int_x \int_y e^{tx} e^{ty} p_X(x) p_Y(y)$$
$$= \int_x e^{tx} p_X(x) \int_y e^{ty} p_Y(y)$$

Onafhankelijkheid heeft te maken met de productregel.
Analoog: $P(A \cap B) = P(A).P(B)$ indien A en B onafhankelijk.

Dus: $\boxed{M_{X+Y}(t) = M_X(t).M_y(t)}$

$$M_{X+Y}(t) = M_X(t).M_Y(t)$$
$$= (pe^t + q)^n.(pe^t + q)^m$$
$$= (pe^t + q)^{m+n}$$: dit is de momentgenererende functie van X en Y

We concluderen dat onze intuïtieve benadering correct was:
$x + y \sim b(m+n, p)$
$$\Updownarrow$$
$$M_{X+Y}(t) = (pe^t + q)^{m+n}$$

Oefening 2: Poisson verdeling

$X \sim Pois(\lambda)$: aantal telefoon oproepen per uur in centrale X
$Y \sim Pois(\mu)$: aantal telefoon oproepen per uur in centrale Y
$\lambda,\ \mu =$ verwachte aantal oproepen per uur
X en Y zijn onafhankelijk

Vraag: wat is de verdeling: $X + Y \sim ?$
Intuïtief: $X + Y \sim Pois(\lambda + \mu)$

1ste stap: momentgenereerende functie zoeken van de Poisson verdeling:
$X \sim Pois(\lambda)$

$p_K = e^{-\lambda}\frac{\lambda^k}{k!}$: definitie dichtheid
$M_X(t) = E(e^{xt}) = e^{tk}\,pk$: definitie verwachte waarde

$$= \sum_{k=0}^{\infty} e^{tk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(e^t\lambda)^k}{k!}$$: zie Taylor reeks in het formularium

$$= e^{-\lambda}e^{(\lambda e^t)} \boxed{= e^{-\lambda+\lambda e^t}}$$

$$M_{X+Y}(t) = M_X(t).M_Y(t)$$
$$= e^{-\lambda(1-e^t)}.e^{-\mu(1-e^t)}$$
$$M_{X+Y}(t) = e^{-(\lambda+\mu)(1-e^t)}$$

Normale vorm :
$M_X(t) = e^{-\lambda(1-e^t)}$

Conclusie: Het intuïtieve resultaat was juist:
$X + Y \sim Pois(\lambda + \mu)$
$\Updownarrow$
$M_{X+Y}(t) = e^{-(\lambda+\mu)(1-e^t)}$

Oefening : stochastische vectoren van lengte 2 (staat niet in het formularium)

$$(X,Y):\Omega \to \mathbb{R}^2$$
$$\omega \to \big(X(\omega), Y(\omega)\big)$$

covariantie : $\operatorname{cov}(x,y) = \sigma_{XY} = E\big((X-EX)(Y-EY)\big)$
$$= \int_y\int_x (x-E(x))(y-E(Y))P_{XY}(x,y)$$: indien continu stochast

correlatiecoëfficiënt : $P_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \in [-1,1]$: dit is wel schaalgevoelig

Eigenschappen:

1. $\sigma_{XY} = \sigma_{YX}$
2. $\operatorname{cov}(x+a,Y) = \operatorname{cov}(X,Y)$
3. $\operatorname{cov}(aX,bY) = ab\operatorname{cov}(X,Y)$

4. $\text{cov}(X,Y) = E((X-EX)(Y-EY))$

$= E(XY - XEY - YEX + (EX)(EY))$

$= E(XY) - E(XEY) - E(YEX) + E((EX)(EY))$

$= E(XY) - (EY)(EX) - (EX)(EY) + (EX)(EY)$

$\boxed{\text{cov}(X,Y) = E(XY) - E(X)E(Y)}$

5. indien X en Y onafhankelijk zijn dan is de verwachte waarde van het product gelijk aan het product van de verwachte waardes : dit is de productregel.

$$E(XY) = \int_X \int_Y xyp_{XY}(x,y)$$

$$= \int_X \int_Y xyp_X(x)p_Y(y)$$

$$= \int_X xp_X(x) \int_Y yp_Y(y)$$

$\boxed{E(XY) = E(x).E(y)}$

Indien X en Y <u>onafhankelijk</u> zijn dan is er geen verband tussen de twee en dus is de covariantie gelijk aan 0: er is geen verband tussen de twee:
$\text{cov}(X,Y) = E(XY) - (EX)(EY) = 0$

Maar: de redenering als cov = 0 dan zijn ze onafhankelijk geldt <u>niet</u>.

<u>FOUTE redenering:</u> $\text{cov}(X,Y) = 0$ → X en Y zijn onafhankelijk
indien cov = 0 dan kan je NIET zeggen dat ze onafhankelijk zijn
<u>vb.</u> $X \sim U\{-1,1\}$ (: X is uniform verdeeld over het interval -1,1)
$Y = X^2$ → zijn afhankelijk
$\text{cov}(X,Y) = E(XY) - E(X)E(Y)$

$= 0 - 0.E(Y) = 0$: en toch is de cov = 0

$XY = \begin{cases} -1 & p = 1/2 \\ 1 & p = 1/2 \end{cases}$

<u>10. Normale verdeling</u> (HB p167)

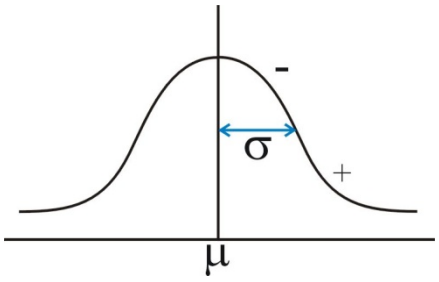
X is een stochastische variabele. <u>vb.</u> lengte, gewicht, IQ
werktijd van een bepaalde taak
opbrengst van een aardappelstruik....

$X = Y_1 + Y_2 + ... + Y_n$ (som van een groot aantal stochastische variabelen)

<u>Wat hebben deze stochasten gemeenschappelijk?</u>
- tendens naar een gemiddelde
- positieve en negatieve afwijkingen t.o.v. het gemiddelde zijn even waarschijnlijk
- weinig grote afwijkingen t.o.v. het gemiddelde

Dichtheidsfunctie : $x \to p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

: in het formularium p4



Ter controle: is p een dichtheidsfunctie? : de twee condities:

- $p \geq 0$: OK
- $\int p = 1$ (stelling zonder bewijs) : oppervlakte onder de curve = 1 : OK

Definitie:

Zij (Ω, G, P) een kansruimte

Zij $Z : \Omega \to \mathbb{R}$ een stochastische variabele: een continue stochast

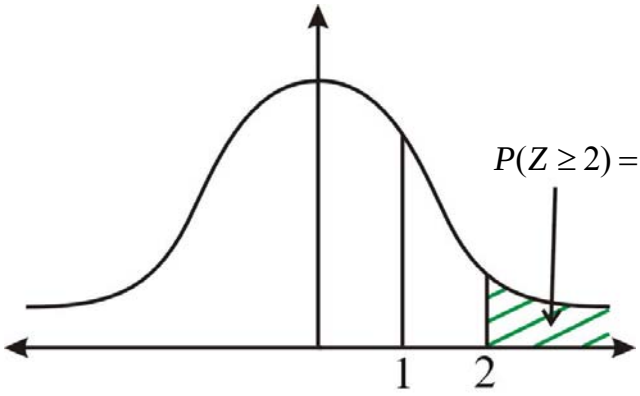
dan is Z standaard normaal verdeeld : $Z \sim N(0,1)$

: Z heeft dus een gemiddelde van 0 ($\mu = 0$) en een variantie van 1 ($\sigma^2 = 1$).

indien $p_Z(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$

Hoe maak je gebruik van de normale verdeling en de tabel?

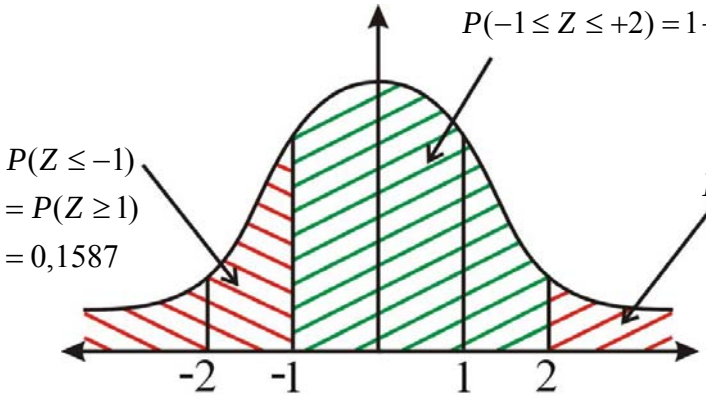
1.



$P(Z \geq 2) = 0,0228$: zie tabel

* De tabel genereert bovenstaart kansen.

2.



$P(-1 \leq Z \leq +2) = 1 - 0,1587 - 0,0228$

$P(Z \leq -1)$
$= P(Z \geq 1)$
$= 0,1587$

$P(Z \geq 2) = 0,0228$
: zie 1

* De normale verdeling is symmetrisch : $P(Z \le -1) = P(Z \ge 1)$
  → de onderstaart kans staat niet in de tabel maar de bovenstaart kans wel.

$M_Z(t) = E(e^{tz})$: standaard normale verdeling (: definitie verwachte waarde)

$$= \int_{-\infty}^{+\infty} e^{tZ} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} \quad \text{: de constante mag voor de integraal}$$

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

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(Z-t)^2+\frac{1}{2}t^2} \text{ dz}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(Z-t)^2} e^{\frac{1}{2}t^2} \text{ dz}$$

Tussen berekening :

$$-\frac{1}{2}Z^2 + tZ = -\frac{1}{2}(Z-t)^2 + \frac{1}{2}t^2$$
$$= -\frac{1}{2}\left(Z^2 - 2tZ + t^2\right) + \frac{1}{2}t^2$$
$$= -\frac{1}{2}Z^2 + tZ - \frac{1}{2}t^2 + \frac{1}{2}t^2$$
$$= -\frac{1}{2}(Z-t)^2 + \frac{1}{2}t^2$$

$$M_Z(t) = e^{\frac{1}{2}t^2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(Z-t)^2} \text{ d(z-t)}$$

$$\boxed{M_Z(t) = e^{\frac{1}{2}t^2}}$$

In het formularium: $M_Z(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$ maar we weten dat $\mu = 0$ en $\sigma = 1$ dus $M_Z(t) = e^{\frac{1}{2}t^2}$

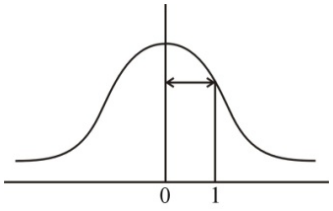
Normale verdeling : $X \sim (\mu, \sigma^2)$

Definitie 1 : Standaard normale verdeling:

$Z \sim N(0,1)$: Z is normaal verdeeld met gemiddelde = $\mu = 0$ en variantie = $\sigma^2 = 1$

$$p_2(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$

$$M_Z(t) = e^{\frac{t^2}{2}}$$



Definitie 2 : Willekeurige normale verdeling

Zij (Ω, G, P) een kansruimte

$X : \Omega \to \mathbb{R}$ een stochastische variabele

Dan $X \sim N(\mu, \sigma^2)$

$E(Z) = M_Z'(t) = 0$ (verwachte waarde = 0)

$E(Z^2) = M_Z''(t)$

$M_Z'(t) = e^{\frac{t^2}{2}} . t = 0$ (: afgeleide van $M_Z(t)$)

$M_Z''(t) = e^{\frac{t^2}{2}} . t^2 + e^{\frac{t^2}{2}} = 1$

indien ofwel $Z = \frac{x-\mu}{\sigma} \sim N(0,1)$

: de willekeurige verdeling is herschaald: we brengen $x-\mu$ op 0 en σ op 1

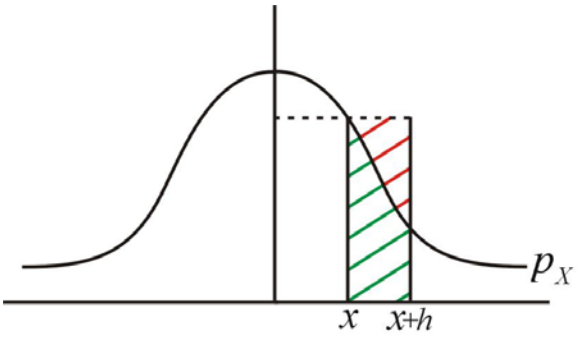
ofwel $p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$: de dichtheid van X

Zijn die 2 criteria hetzelfde? gelijkaardig?

Te bewijzen : $\frac{x-\mu}{\sigma} \sim N(0,1)$ (: continu stochast)

als en slechts als

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



: je maakt een fout, maar als h klein is dan is het fout ook klein.

$p_X(x).h = P(x \le X \le x+h)$

$-\mu$ bijvoegen:

$= P(x-\mu \le X-\mu \le x+h-\mu)$

delen door σ:

$= P(\frac{x-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{x+h-\mu}{\sigma})$

$= P(\frac{x-\mu}{\sigma} \le Z \le \frac{x+h-\mu}{\sigma})$

$= P(\frac{x-\mu}{\sigma} \le Z \le \frac{x-\mu}{\sigma} + \frac{h}{\sigma})$

$= p_Z\left(\frac{x-\mu}{\sigma}\right).\frac{h}{\sigma}$

$p_X(x).h = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} . \frac{h}{\sigma}$

h weg delen:

$$\boxed{p_X(x) = \frac{1}{\sqrt{2\pi}} . \frac{1}{\sigma} . e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}$$

Voor Normale Verdeling : definitie momentgenereerende functie
→ voor het genereren van momenten en transformeren van stochasten

Bewijs:
$M_X(t) = E\left(e^{tx}\right)$: eerder al gedaan voor Poisson en Bernouilli, nu voor Normale verdeling

$(x = \sigma Z + \mu$:)

$$= E\left(e^{t(\sigma Z+\mu)}\right)$$
$$= E\left(e^{t\sigma Z}.e^{t\mu}\right)$$
$$= e^{t\mu}E\left(e^{t\sigma Z}\right)$$
$$= e^{t\mu}M_Z(\sigma t) \quad \leftarrow \; M_Z(t) = E(e^{tZ}) = e^{\frac{t^2}{2}}$$

$$M_X(t) = e^{\mu t + \frac{(\sigma t)^2}{2}}$$

$X \sim N(\mu, \sigma^2)$

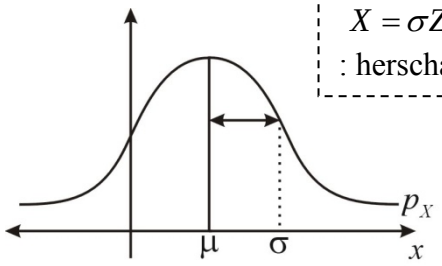
$Z = \dfrac{x-\mu}{\sigma} \sim N(0,1)$

$X = \sigma Z + \mu$: herschalen

Bewijs:

$$M_X(t) = e^{\mu t + \sigma^2 \frac{t^2}{2}}$$

$$E(X) = E(\sigma Z + \mu)$$
$$= \sigma E(Z) + E(\mu)$$
$$= \mu$$



$X = \sigma Z + \mu$
: herschalen van de normale verdeling

Dus : $\mu = \mu$

Bewijs:

$\text{var}\, X = \text{var}(\sigma Z + \mu)$ (: constanten optellen (de $+\mu$) heeft geen effect op de spreiding)

$$= \sigma^2 \,\text{var}\, Z$$
$$= \sigma^2$$

Dus : $\sigma^2 = \sigma^2$

Kijken naar sommen van Normale Verdelingen aan de hand van momentgenereerende functies (HB p179)
→ analoog zie eerder Poisson en Bernouilli

$X_1 \sim N(\mu_1, \sigma_1^2)$

$X_2 \sim N(\mu_2, \sigma_2^2)$

X_1 en X_2 zijn onafhankelijk

Vraag: $X_1 + X_2 \sim ?$

Intuïtief : $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

met behulp van de momentgenereerende functie:

$M_{X_1+X_2}(t) = E\left(e^{t(X_1+X_2)}\right)$ (: exponentieel dus sommen van producten)

$= E\left(e^{tX_1}.e^{tX_2}\right)$ (: X_1 en X_2 zijn onafhankelijk, anders mag deze stap niet)

$= E\left(e^{tX_1}\right)E\left(e^{tX_2}\right)$

$= M_{X_1}(t).M_{X_2}(t)$ (: definitie: $= e^{t\mu}M_Z(\sigma t) = e^{\mu t+\frac{(\sigma t)^2}{2}}$)

$= e^{\mu_1 t+\sigma_1^2\frac{t^2}{2}}.e^{\mu_2 t+\sigma_2^2\frac{t^2}{2}}$

$= e^{(\mu_1+\mu_1)t+(\sigma_1^2+\sigma_2^2)\frac{t^2}{2}}$ → normaal verdeeld

Besluit: $X_1 + X_2 \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

$\mu_1+\mu_2$ kan je verwachten : $E(X_1+X_2) = E(X_1)+E(X_2) = \mu_1+\mu_2$

$\sigma_1^2+\sigma_2^2$ ook : $\text{var}(X_1+X_2) = \text{var}\,X_1 + \text{var}\,X_2 + 2\text{covar}(X_1, X_2)$

(: $2\text{covar}(X_1, X_2) = 0$: indien onafhankelijk)

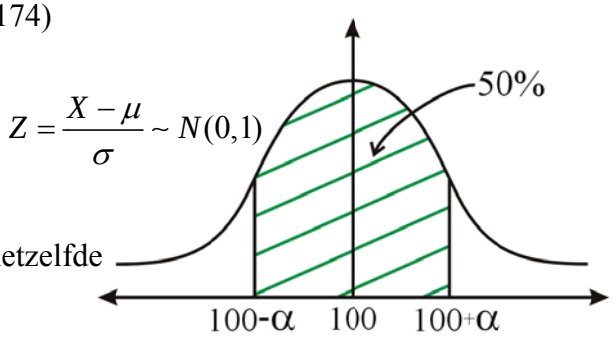
Oefeningen: (op Normale Verdelingen) (HB p174)

Oefening 1:

$X = IQ$ van 9-jarigen

Verdeling van IQ: $X \sim N(\mu = 100, \sigma = 11)$

$\sim N(\mu = 100, \sigma^2 = 121)$: hetzelfde



Vraag: Geef het 50% populatie interval rond de gemiddelde IQ

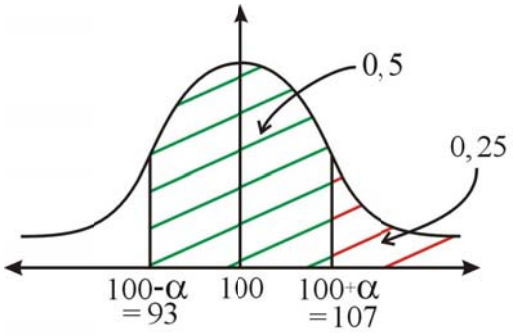
$P(100-\alpha \le X \le 100+\alpha) = 0,5$

Herschalen: -100 en delen door 11 dus: $\frac{-100}{11} = \frac{-\mu}{\sigma}$

$P(-\frac{\alpha}{\sigma} \le \frac{X-100}{11} \le \frac{\alpha}{\sigma}) = 0,5$

$P(-\frac{\alpha}{\sigma} \le Z \le \frac{\alpha}{\sigma}) = 0,5$

en $Z \sim N(0,1)$: omdat we herschaald hebben naar de standaard normaal functie



We weten dus de bovenstaartkans = 0,25 en we moeten nu de bijhorende Z waarde zoeken in de tabel: $Z_{0,25} = 0,67$ (tabel: 0,07, kolom:0,6 → 0,67)

$$\frac{\alpha}{\sigma} = 0,67$$
$$\alpha = 0,67\sigma$$
$$= (0,67)(11) = 7,37$$

$$P\left(-\frac{\alpha}{\sigma} \le Z \le \frac{\alpha}{\sigma}\right)$$
$$P\left(-\frac{7,37}{11} \le Z \le \frac{7,37}{11}\right)$$
$$P(-0,67 \le Z \le 0,67)$$
$$P(-0,67 \le \frac{X-100}{11} \le 0,67)$$

Antwoord: $P(93 \le X \le 107)$
→ de helft van de 9 jarigen heeft een IQ tussen 93 en 107.

Oefening 2:

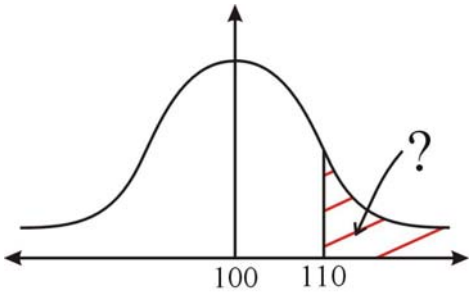
Vraag: Bruno heeft een IQ van 110, in welk percentage (topniveau) ligt hij?

$P(X \ge 110)$

Herschalen $X = \frac{x-\mu}{\sigma}$:

$$= P\left(\frac{X-100}{11} \ge \frac{110-100}{11}\right)$$
$$P(Z \ge \frac{10}{11} = 0,91) \approx 0.184$$

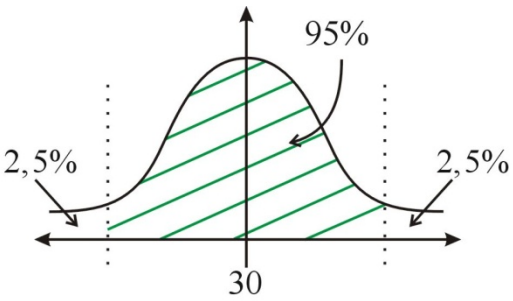
Antwoord: Bruno zit bij de top 18%



Oefening 3: (HB p185)

Een vertaal bureau vertaalt tekst in A4 formaat.
X = vertaal tijd per pagina in minuten

$X \sim N(30, \sigma = 6)$: 30min is de verwachte waarde
Kost = 50 EUR per uur



Vraag: Kost prognose met 95% zekerheid:
* voor tekst van 1 pagina
* voor tekst van 40 pagina's

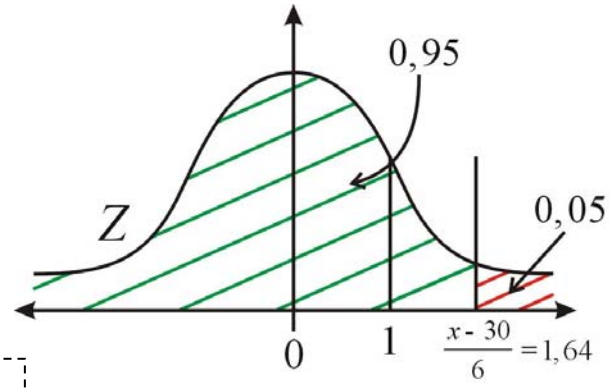
Deel 1 - Kost van 1 pagina: (met 95% zekerheid)
→ bereken de vertaal tijd en dan prijs

$$P(X \le x) = 0,95$$

$$P\left(\frac{X-30}{6} \le \frac{x-30}{6}\right) = 0,95$$

$$P\left(Z \le \frac{x-30}{6}\right) = 0,95$$

$$P(Z \le 1,64)$$



$Z_{0,05} = 1,64$
= bovenstaart dus OK

x zoeken:

$$\frac{X-30}{6} = 1,64$$

$$X = 30 + 6.(1,64) = \underline{39,84}$$

Met 95% zekerheid zal de tekst van 1 pagina in minder dan 39,8 minuten worden vertaald.
De kost is dus kleiner dan $\frac{40}{60}.50 = \underline{€33,30}$

Deel 2 - Kost van 40 pagina's: (met 95% zekerheid)

Foute methode : 1 pagina = €33,30 : 40 pagina's = 33,30 x 40 = €1332
→ dit zou het antwoord zijn voor 40 aparte pagina's door andere tolken
→ de spreiding klopt niet

Juiste methode : door spreiding moet de som aangepast worden voor 40 pagina's:
$V = X_1 + X_2 + ... + X_{40}$: totale vertaaltijd
$X_i \sim N(30, \sigma = 6)$
$V = X_1 + X_2 + ... + X_n \sim N(30\text{x}40, \sigma_V^2 = 36\text{x}40)$: belangrijke stap
: veronderstelt onafhankelijkheid
$V \sim N(1200, \sigma_V = 37,95)$

reduceren naar standaard normaal verdeling met de herschalings formule $\frac{x-\mu}{\sigma}$:

$$\frac{V-1200}{37,95} \sim N(0,1)$$

$$\frac{V-1200}{37,95} = 1,64 \text{ : via tabel}$$

$$V = 1200 + 37,95(1,64)$$

$$P(X \le x) = 0,95$$

$$P(\frac{X-1200}{37,95} \le \frac{x-1200}{37,95}) = 0,95$$

$$P(Z \le \frac{x-1200}{37,95}) = 0,95$$

$$Z_{0,05} = 1,64$$

Dus : $\frac{x-1200}{37,95} = 1,64$

$x = 1262$ minuten

→ met 95% zekerheid kunnen we zegen dat de vertaaltijd (V) kleiner is dan 1262 minuten

→ met 95% zekerheid is de kost $\le \frac{1262}{60}.50 = €1051,70$

---

* Normale verdeling

- definitie: $X \sim N(\mu, \sigma^2)$

- momentgenereerende functie: $M_X(t) = e^{\mu t + \sigma^2 \frac{t^2}{Z}}$

* Limietstelling

We merken op: zij (Ω, G, P) een kansruimte en X_1, X_2 stochastische variabelen:

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$
$$= \mu_1 + \mu_2$$

$$\text{var}(X_1 + X_2) = E\left((X_1 + X_2) - (\mu_1 + \mu_2)\right)^2$$
$$= E\left((X_1 - \mu_1) + (X_2 - \mu_2)\right)^2$$
$$= E\left((X_1 - \mu_1)^2 + 2(X_1 - \mu_1)(X_2 - \mu_2) + (X_2 - \mu_2)^2\right)$$
$$= E(X_1 - \mu_1)^2 + 2E(X_1 - \mu_1)(X_2 - \mu_2) + E(X_2 - \mu_2)^2$$
$$= \text{var } X_1 + 2\text{covar}(X_1, X_2) + \text{var } X_2 \quad (: \text{lineariteit van E})$$

! Indien X_1 en X_2 onafhankelijk : covariantie = 0

dan krijg je: $\text{var}(X_1 + X_2) = \text{var } X_1 + \text{var } X_2$ (: optelregel van de variantie)

Opmerkingen (HB p180)

* Zij $X_1, X_2, ..., X_n$ een rij van onafhankelijke identieke verdeelde (o.i.v) stochastische variabelen met gemiddelde gelijk aan μ $\left(= E(x)\right)$ en met variantie gelijk aan σ^2

Stel: $\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$ (: rekenkundig gemiddelde van een steekproef)

Opmerking a):

* Dan $E(\bar{X}_n) = \frac{1}{n} E(X_1 + X_2 + ... + X_n)$

$= \frac{1}{n}\left(E(X_1) + E(X_2) + ... + E(X_n)\right) = \underline{\mu}$

(: dit wordt afgeleid van $\bar{X}_n$ van hier boven)

→ alle stochasten hebben dezelfde verwachte waarde

Opmerking b):

* Dan $\text{var}(\bar{X}_n) = \text{var}\left(\frac{1}{n}(X_1 + X_2 + ... + X_n)\right)$

$= \frac{1}{n^2}\left(\text{var } X_1 + \text{var } X_2 + ... + \text{var } X_n\right) = \frac{1}{n^2} n\sigma^2 = \underline{\frac{\sigma^2}{n}}$

(: dit wordt afgeleid van $\tilde{s}^2 = \frac{1}{2}\left(\sum x_i - \bar{x}\right)^2$ van hier boven)

→ alle stochasten hebben dezelfde varianties

Opmerking c):

* bekijk de limiet voor n gaat naar oneindig :

$E(\bar{X}_n) \underset{n\to\infty}{\to} \mu$ (lees: de verwachte waarde convergeert naar μ als $n \to \infty$)

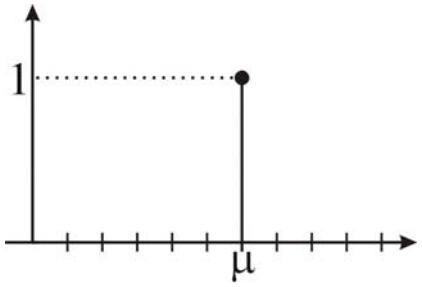
(: want het hangt niet af van n : Y is constant)

$\text{var}(\bar{X}_n) \underset{n\to\infty}{\to} 0$ (: σ^2 hangt wel af van n)

→ er is geen spreiding, (vb. een constante)

→ een stochast met verwachte waarde μ en variantie gelijk aan 0 noemen we ontaard
dit wil zeggen; alle kansmassa zit in μ

Voor een discrete stochast ziet een ontaarde stochast er zo uit:

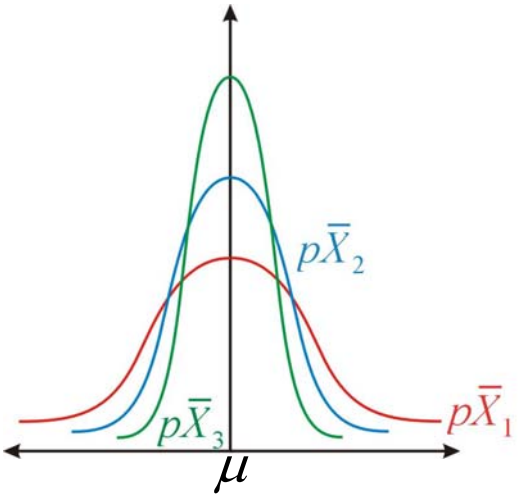


Opmerking d):

* Stel $X_1 + X_2 + ... + X_n$ onafhankelijk identieke verdeelde stochastische variabelen
en $X_i \sim N(\mu, \sigma^2)$

Dan $\bar{X}_n = \frac{X_1 + X_2 + ... + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$: formule **Centrale Limiet Stelling** : CLS

→ de steekproef gemiddelde van een steekproef uit een normale populatie $X \sim N(\mu, \sigma^2)$ is ook normaal verdeeld



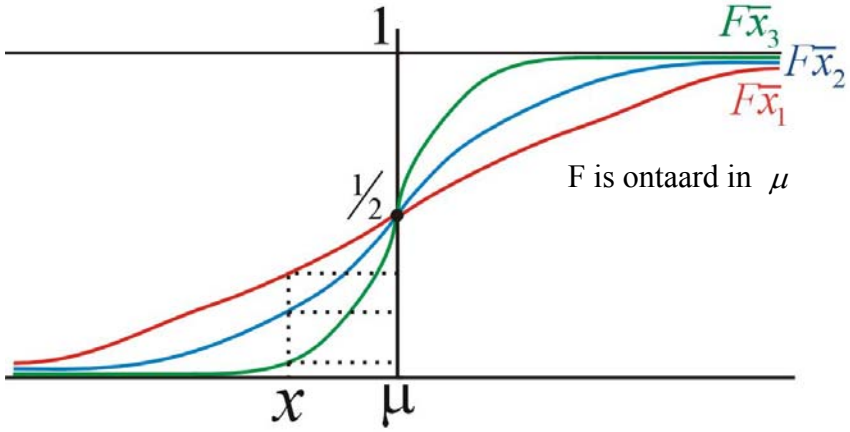
Probleem: moeilijk te tekenen, blad tekort want de grafiek gaat te hoog gaan.
Oplossing: cumulatieve verdeling gebruiken (gaat nooit boven 1)

$F(x) = P(X \leq x)$

$F(\mu) = P(X \leq \mu) = \frac{1}{2}$

: links van $\mu \to 0$

: rechts van $\mu \to 1$



Definitie: (formularium)

Zij $Y, Y_1, Y_2, ..., Y_n, ...$ een rij van stochastische variabelen

Dan convergeert Y_n "in verdeling" naar Y (denk aan de tekening)

als $F_{Yn}(y) \underset{n\to\infty}{\to} F_Y(y)$ voor elke y waar F_Y continu is

: dus niet in Y want in Y maakt het sprongen

We noteren $Y_n \overset{D}{\to} Y$ (: D = convergeert in verdeling)

Centrale Limiet Stelling: stelling zonder bewijs:

Zij $X_1, X_2, ..., X_n$ een rij van onafhankelijke identieke verdeelde stochastische variabelen (o.i.v st. v.) met gemiddelde gelijk aan μ en met variantie gelijk aan σ^2 $(< \infty)$: de spreiding moet eindig zijn.

Dan $\bar{X}_n \to Y$: steekproef gemiddelde convergeert in verdeling naar Y

met $Y \sim N(\mu, \frac{\sigma^2}{n})$

(CLS) : of $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{D}{\to} Z$ met $Z \sim N(0,1)$

Praktisch:

CLS: $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$ ($\approx$: is ongeveer verdeeld)

als $n \geq 30$ dan 2 cijfers na de comma

Oefening: (HB p189)

Een munt wordt 100 keer getost en genereert 60 keer 'munt'
Vraag: is deze munt eerlijk?

Oplossing: via de centrale limiet stelling

$X_i = 0$ indien kop

1 indien munt

$X_1 + X_2 + ... + X_{100}$ is onafhankelijke identiek verdeeld met variantie $< \infty$

Zij: $Y = X_1 + X_2 + ... + X_{100}$

Stel $X_i \sim b(1, p = 0,5)$: dit gebruikt de veronderstelling dat de munt eerlijk is

Vraag: bepaal $P(Y \geq 60)$

: indien de kans voldoende groot is dan kunnen we aannemen dat de munt eerlijk is.

Oplossing: $Y \sim b(100, p = 0,5)$: Bernouilli verdeling omdat er twee mogelijke uitkomsten zijn

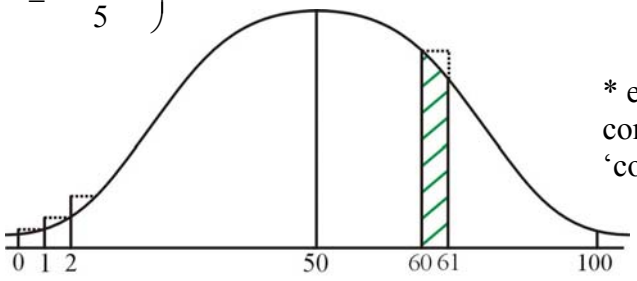
$\approx N(50, \sigma^2 = 100\text{x}0,5\text{x}0,5)$

$\approx N(50, 25)$

Steekproefversie:	Somversie:
$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$	$\frac{X - \mu}{\sigma}$

Centrale Limiet Stelling: $Z = \frac{Y - 50}{5} \approx N(0,1)$

$P\left(\frac{Y-50}{5} \geq \frac{60-50}{5}\right)$



* een discrete stochast moeten we continu gaan benaderen en dan een 'continuiteitscorrectie' doen

Continuiteitscorrectie: (omdat we van een discrete naar een continue stochast gaan)

$P(Y \geq 59,5)$

$P\left(\frac{Y-50}{5} \geq \frac{59,5-50}{5}\right) = P\left(Z \geq 1,9\right) = 0,0287$ (: Tabel)

Besluit : we kunnen twijfelen aan de eerlijkheid van de munt omdat de kans dat je 60 keer 'munt' tost zeer klein is (0,0287).

11. Speciale verdelingen (HB p194)

11.1 Hypergeometrische verdeling (HB p195)

Stel de variabele: X = aantal successen in een steekproef van grootte n

Bij trekking uit een populatie met:
N : grootte van de populatie
M : aantal successen in de populatie
$p = \frac{M}{N}$: populatieproportie aan successen

Wat is de verdeling van X?
1. Bij steekproeven met teruglegging, bij elke trekking blijft de succeskans p en dus heeft X een binominale verdeling:

$$X \sim b(n, p)$$

2. Bij steekproeven zonder teruglegging is bij opeenvolgende trekkingen de succeskans niet constant. X heeft de hypergeometrische verdeling met parameters N, M, n, genoteerd:

$$X \sim h(N, M, n)$$

Met de dichtheid:

$$p(k) = P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},\ k = 0,1,...,n$$

Oefening:
In een groep van 60 studenten zijn er 20 die goede volleybalspelers zijn. Kies willekeurig een team van 6 studenten. Wat is de kans dat u juist 3 goede spelers heeft?

X = aantal goede spelers op steekproef van 6 (steekproef zonder teruglegging)
Verdeling: $X \sim h(60, 20, 6)$
Kans op juist 3 goede spelers:

$$P(X = 3) = \frac{\binom{20}{2}\binom{40}{3}}{\binom{60}{6}} = \frac{(20 \text{ x } 19 \text{ x } 18) \text{ x } (40 \text{ x } 39 \text{ x } 38) \text{ x } (6 \text{ x } 5 \text{ x } 4 \text{ x } 3 \text{ x } 2)}{(3 \text{ x } 2) \text{ x } (3 \text{ x } 2) \text{ x } (60 \text{ x } 59 \text{ x } 57 \text{ x } 56 \text{ x } 55)} = \underline{0,225}$$

11.2 Gamma-verdeling (HB p197)
→ model voor levensduur of een wachttijd variabele (continu)

Oefening 1: $X \sim N(0,1)$: standaard normale verdeling

$$p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$Y = (X - 0)^2 = X^2$ (:0 is de verwachte waare)

We zijn geïnteresseerd in de verdeling van Y:
→ drie mogelijkheden : momentgenereerende functie, F, of p_X.
de juiste methode voor deze oefening:

$F_Y(y) = P(Y \le y)$: de cumulatieve verdeling

$= P(X^2 \le y)$

$= P\left(-\sqrt{y} \le X \le \sqrt{y}\right)$

$= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} . e^{-\frac{X^2}{2}}$

$F_Y(y) = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-\frac{X^2}{2}}$: symmetrie : maal 2 omdat beide kanten gelijk zijn

$p_Y(y) = F_Y'(y) = \frac{2}{\sqrt{2\pi}} e^{-\frac{y}{2}} \frac{1}{2} y^{-\frac{1}{2}}$

$X \sim N(0,1)$

$p_X^2(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} y^{-\frac{1}{2}}$: gamma-verdeling $X^2 \sim \Gamma\left(\frac{1}{2}, 2\right)$

→ $X \sim \Gamma(\alpha, \beta)$: gamma-verdeling

$p_X(x) = c^{tu} X^{\alpha-1} e^{-\frac{X}{\beta}}$ voor $X \ge 0$ $\left(: c^{tu} = \frac{1}{\Gamma(\alpha)\beta^\alpha} \right)$

Eigenschappen:

$\Gamma(1) = 1$

$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Oefening 2: gamma-verdeling

* $X \sim N(0,1)$

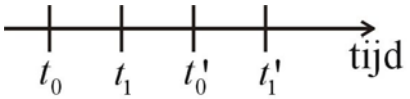
  $Y \sim gamma$

* $Y \sim Pois(\lambda)$: telstochast : aantal _ per _

Formularium: $p_K = \frac{\lambda^K}{k!} e^{-\lambda}$ voor k = 0,1,2...

Veronderstelling 1 : $Y(t_0, t_0+1) \sim Pois(\lambda)$: (t_0, t_0+1) :een tijdsinterval vb. we tellen over 1uur

Veronderstelling 2 : $Y(t_0, t_1)$ en $Y(t_0', t_1')$ zijn onafhankelijk : de tijdsperiodes zijn disjunct



$Y(t_0, t_1) \sim Pois((t_0, t_1)\lambda)$

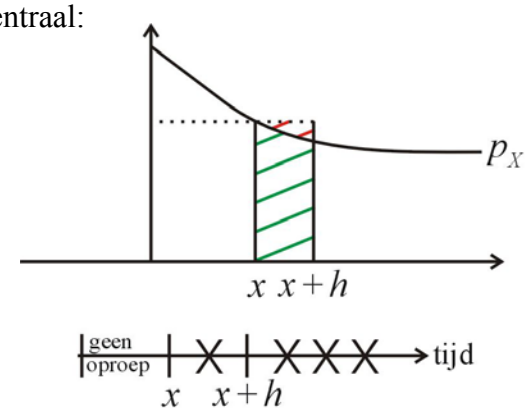
(t_0, t_1) = de tijd dat je daar zit

λ = verwachte waarde : afhankelijk van het volume

$\lambda = 100$: 100 oproepen per uur

X = wachttijd tot de <u>eerste</u> oproep in het telefoon centraal:

$$p_X(x).h = P\left(x \le X \le x+h\right)$$
$$= P\left(Y_{(0,X)} = 0 \text{ en } Y_{(x,x+h)} = 1\right)$$
$$= P\left(Y_{(0,X)} = 0 \text{ x } Y_{(x,x+h)} = 1\right)$$



Poisson omdat er tussen 0 en X geen oproep is
en tussen x en x+h er één oproep is.

$$= \frac{(x\lambda)^0}{0!}e^{-\lambda x}.\frac{(h\lambda)^1}{1!}e^{-\lambda h}$$
$$= e^{-\lambda x}\lambda h e^{-\lambda h}$$
$$p_x(x) = e^{-\lambda x}\lambda(1-\lambda h) \underline{= e^{-\lambda x}\lambda}$$

→ we bekomen dezelfde dichtheids functie als die van de exponentiële verdeling
→ de wachttijd tot eerste "oproep" is exponentieel verdeeld

$X \sim \Gamma(\alpha,\beta)$ (: in het formularium p4)

$$p_X(x) = \begin{cases} c^{tu}x^{\alpha-1}e^{-\frac{x}{\beta}} & \text{voor } x \ge 0 \\ 0 & \text{voor } x < 0 \end{cases}$$

<u>Indien</u> $X \sim N(0,1)$

Dan $Y = X^2 \sim \Gamma\left(\frac{1}{2},2\right)$

<u>Poisson proces:</u>

$Y_{(t_0,t_1)} \sim Pois\left((t_1-t_0)\lambda\right)$

t_0 : begin tijdstip

t_1 : eind tijdstip

Y_{I_1} en Y_{I_2} zijn onafhankelijk (disjunct)

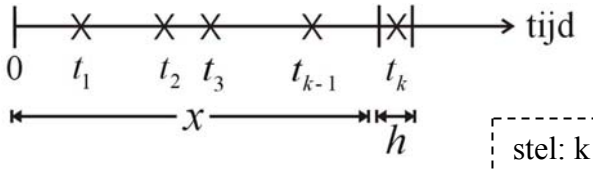
zodra de intervallen I_1 en I_2 disjunct zijn (hebben geen geheugen)
is de wachttijd tot de eerste oproep:

$X_1 \sim Exp(\lambda)$

$X_1 \sim \Gamma(1, 1/\lambda)$: zie vorige pagina

<u>Wachttijd tot de k-de oproep</u>: $\boxed{k > 1}$

X_k = wachttijd tot de k-de oproep:



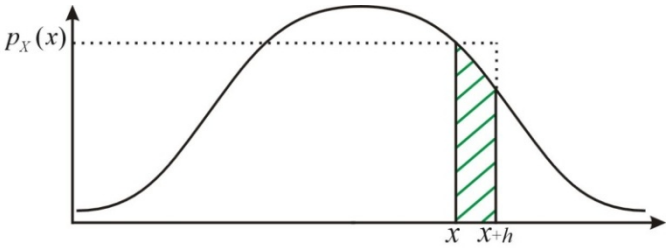
We bepalen de dichtheid door middel betekenis dichtheid:

stel: k = 1
$p_x = \lambda e^{-\lambda x}$: exponentiële

$$p_{X_k}(x).h = P\left(x \le X \le x+h\right)$$
$$= P\left(Y_{(0,x)} = k-1 \text{ en } Y_{(x,x+h)} = 1\right) \text{ : onafhankelijk}$$
$$= P\left(Y_{(0,x)} = k-1\right).P\left(Y_{(x,x+h)} = 1\right)$$
$$= \frac{(x\lambda)^{k-1}e^{-\lambda x}}{(k-1)!}.\frac{h\lambda e^{-h\lambda}}{1!}$$

$$p_{X_k}(x) = \frac{e^{-\lambda x}(x\lambda)^{k-1}\lambda}{(k-1)!}$$
$$= \frac{1}{(k-1)!}\lambda^k x^{k-1}e^{-\lambda x}$$

$$\underline{X_k \sim \Gamma\left(k, 1/\lambda\right)}$$



* de oppervlakte kunnen we interpreteren als een kans.
* als h klein is dan is het fout ook klein.
* h klein : $e^{-\lambda h} \approx 1$

Dus voor elke $k = 1, 2, ...$

$$X_k \sim \Gamma\left(k, 1/\lambda\right)$$
$$\boxed{\Gamma\left(1, 1/\lambda\right) = Exp(\lambda)}$$

Gamma-verdeling; bespreek deze functie:

$$X \sim \Gamma(\alpha, \beta)$$
$$p_X(x) = \begin{cases} c^{tu}x^{\alpha-1}e^{-\frac{x}{\beta}} & \text{voor } x \ge 0 \\ 0 & \text{voor } x < 0 \end{cases}$$

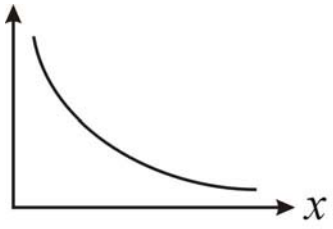
■ c<u>tu</u> hangt af van α en β en zorgt ervoor dat $\int_0^\infty p(x) = 1$ (: voorwaarde voor dichtheid)

(met andere woorden: $c^{tu} = \frac{1}{\int_0^\infty x^{\alpha-1}e^{-\frac{x}{\beta}}}$) : niet essentieel/kennen

■ bespreking van de functie $f : x \to x^{\alpha-1}e^{-\frac{x}{\beta}}$

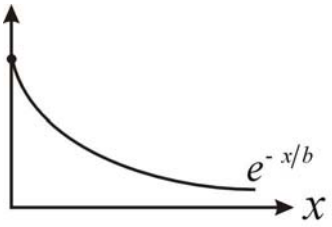
$$f'(x) = (\alpha-1)x^{\alpha-2}e^{-\frac{x}{\beta}} - \frac{1}{\beta}x^{\alpha-1}e^{-\frac{x}{\beta}}$$
$$= e^{-\frac{x}{\beta}}x^{\alpha-2}\left[\alpha - 1 - \frac{x}{\beta}\right]$$

* $\alpha - 1 < 0$

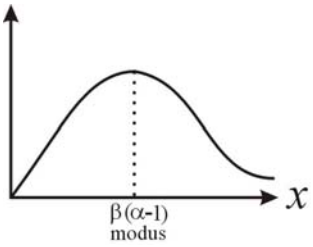


: asymptoot

* $\alpha = 0$



* $\alpha - 1 > 0$



■ $X \sim \Gamma(\alpha, \beta)$

- $E(X) = \alpha\beta = \mu$: verwachte waarde : ook in formularium
- $\text{var}(X) = \alpha\beta^2$: spreiding
- $M_X(t) = (1-\beta t)^{-\alpha}$: momentgenererende functie

■ 1ste speciaal geval : $X \sim Exp(\lambda)$: Exponentiële verdeling

: $\alpha = 1$, $\beta = \frac{1}{\lambda}$

■ 2e speciaal geval : $X \sim N(0,1)$: Normale verdeling

- $Y = X^2 \sim \Gamma(\frac{1}{2}, 2)$ (: hier is $\alpha = 1/2$ en $\beta = 2$)
- $E(Y) = 1$ (: $= \alpha\beta$)
- $\text{var}(Y) = 2$ (: $= \alpha\beta^2$)
- $M_Y(t) = (1-2t)^{-\frac{1}{2}}$

↓ Momentgenererende functie (zie eerder voor Poisson, Bernouilli, Normale verdeling)

Gevolg : $X_1, ..., X_n$ is onafhankelijk identiek verdeeld $N(0,1)$

$Y = X_1^2 + X_2^2 + ... + X_n^2$: momentgenererende functie van de som

onafhankelijk → product :

$$M_Y(t) = (1-2t)^{-\frac{1}{2}}.(1-2t)^{-\frac{1}{2}}.....(1-2t)^{-\frac{1}{2}}$$

$$= (1-2t)^{-\frac{n}{2}}$$

$$M_{X+Y}(t) = M_X(t).M_Y(t)$$

en bekom je dus na de som van de Normale verdelingen:

$Y \sim \Gamma(\frac{n}{2}, 2)$ of X_n^2 : dit is de Chi-kwadraat verdeling

→ dit is een belangrijke verdeling

$Y \sim \mathrm{X}_n^2$ (of $\Gamma(\frac{\mu}{2}, 2)$)

- $E(Y) = n \quad (:= \alpha\beta)$
- $\mathrm{var}(Y) = 2n \quad (:= \alpha\beta^2)$
- $M_Y(t) = (1-2t)^{-\frac{\mu}{2}}$

CLS : som voldoende groot kunnen we ons behelpen met normale verdeling
gamma : Poisson omzetten naar machtigen
$N(0,1) \to \mathrm{X}_n^2$

DEEL III
STATISTISCHE BESLUITVORMING

12. Verdeling van steekproefstatistieken (HB p205)

Machine in een productieproces vult dozen met 2kg waspoeder.
Vraag: is de machine goed afgesteld?

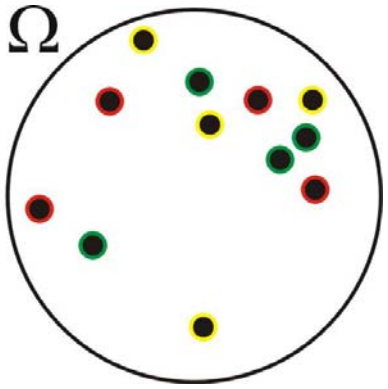
Ω = verzameling van alle dozen waspoeder gevuld door deze machine
$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$ = gewicht van waspoeder in doos ω (in gram)

Uit een steekproef van lengte 4 leren we:
2050, 1920, 1830, 2040 gram : de geobserveerde data
gemiddelde : $\bar{x} = 1960$ gram
standaard afwijking: $s = 90.8$ gram

Vraag: is het gemiddelde $\bar{x}$ echt te klein?
Vraag: is de machine juist afgesteld?

Hoe betrouwbaar is deze steekproef?
Intuitie: - hoe groter de steekproef, hoe betrouwbaarder.
- hoe kleiner de steekproefvariantie (spreiding), hoe betrouwbaarder.

Het antwoord moet rekening houden met de mogelijke fluctuaties in de data bij een herhaalde steekproef-trekkingen.



$\bar{X}$ $\bar{X}$ $\bar{X}$

$\bar{x}$ als stochastische variabele

Vraag: wat is de verdeling van $\bar{x}$?
$X_i : \Omega \to \mathbb{R}$ $i = 1, 2, ..., n$ is onafhankelijke identiek verdeeld (o.i.v)

$$\text{CLS} : \bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Verdeling van de populatie stochastische variabele X kan men uitdrukken als :
F_X : cumulatieve verdeling
$p(x)$: dichtheid
mgf M_X : momentgenereerende functie

Terminologie
Zij (Ω, G, P) een kansruimte
$X : \Omega \to \mathbb{R}$ een stochastische variabele met dichtheid p_X

■ Steekproef van lengte n uit X is een rij $X_1, X_2, ..., X_n$ (grote letters) van onafhankelijke identiek verdeelde stochastische variabelen met $p_{X_i} = p_x$
Het is een aselecte steekproef met teruglegging uit een " ∞ grote" populatie.

■ Geobserveerde steekproef van lengte n uit X is een rij van getallen (of dataset) $x_1, x_2, ..., x_n$ (kleine letters)

■ Een statistiek is een stochastische variabele gebaseerd op een steekproef.
vb. $\overline{x} = \dfrac{X_1, X_2, ..., X_n}{n}$
$S^2 = \sum (X_i - \overline{X})^2 / n - 1$: variantie met n-1 weging : in het formularium p1

■ Een geobserveerde statistiek is een reëel getal.
vb. $\overline{X}\big|_{obs} = \overline{x} = (x_1 + x_2 + ... + x_n)/n$
$S^2\big|_{obs} = s^2 = \sum (x_i - \overline{x})^2 / n - 1$

13. Parameters schatten (HB p209)

Schatten van populatieparameters
vb. populatie parameters: $\vartheta = \mu, \sigma^2, p$ (Bernoulli)

Gegeven : $X_1, X_2, ..., X_n$ een steekproef uit X.
Gevraagd - een "goede schatter" voor ϑ : puntschatter
- een "95% betrouwbaarheids interval" voor ϑ : intervalschatter
- een methode om schatters te construeren

Definitie:
■ Een puntschatt**er** $\hat{\vartheta}$ voor de populatieparameter ϑ is een stochastische variabele:
$\hat{\vartheta} = \hat{\vartheta}(X_1, X_2, ..., X_n)$ die gebruikt wordt om de parameter te schatten.

■ Een puntschatt**ing** is de waarde bij een geobserveerde steekproef
$\hat{\vartheta}_{obs} = \hat{\vartheta}(x_1, x_2, ..., x_n)$: observatie waarde of een getal: $\hat{\vartheta}_{obs}$

vb. X is een stochastische variabele
$\vartheta = \mu$: de populatie parameter is de gemiddelde van de populatie
: de schatting is goed als het getal niet ter ver ligt van μ
$X_1, X_2, ..., X_n$ is een steekproef uit X

We zoeken een schatter voor μ, hier zijn drie mogelijkheden:

1. $\hat{\vartheta} = X_1$ want $E(X_1) = \mu$

2. $\hat{\vartheta} = 2X_2 - X_n$ want $E(2X_2 - X_n) = \mu$ (: je gebruikt het laatste en 2de data punt)

3. $\hat{\vartheta} = \overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$ want $E(\overline{X}) = \mu$

→ al deze schatters zijn onvertekend (: unbiased, $E(X) = \mu$)

$\hat{\vartheta}$ is een onvertekende schatter voor ϑ indien:
$E(\hat{\vartheta}) = \vartheta$ (unbiased)
Het verschil $E(\hat{\vartheta}) - \vartheta$ noemen we "vertekening"

hoofdletters of griekse letters met een hoedje = stochastische veranderlijke

kleineletters = geobserveerde waardes

$\hat{\vartheta}_{n1} = X_1$ want $E(X_1) = \mu$
$\hat{\vartheta}_n = \overline{X}_n$ want $E(\overline{X}) = \mu$

Vraag: wat gebeurt er als n groter wordt? Welk effect heeft dit op de spreiding?
Een schatter is consistent als:
$P\left(\left|\hat{\vartheta}_n - \vartheta\right| > E\right) \underset{n\to\infty}{\to} 0$
→ wanneer de steekproef grote naar oneindig gaat, gaat de kans naar 0
De afstand tussen $\hat{\vartheta}_n - \vartheta$ is van belang.

Chebychev: $P\left(\left|x - \mu\right| \le k\sigma\right) \ge 1 - \frac{1}{k^2}$

$$P\left(\left|x - \mu\right| > k\sigma\right) \le \frac{1}{k^2}$$

$$P\left(\left|x - \mu\right| > \varepsilon\right) \le \frac{\sigma^2}{\varepsilon}$$

$$\left(\varepsilon = \frac{\varepsilon}{\sigma}.\sigma = k.\sigma\right)$$

We passen Chebychev toe voor $\hat{\vartheta}_n = \overline{X}_n$

CLS : $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$

$P\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \le \frac{\sigma^2}{\varepsilon.n} \underset{n\to\infty}{\to} 0$ (: neem willekeurig $\varepsilon > 0$)

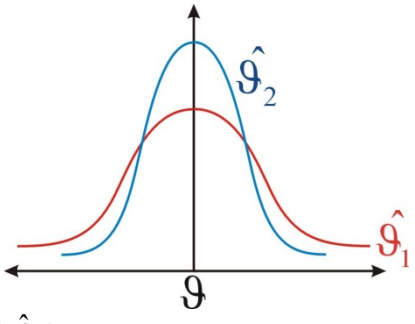
→ $\overline{X}_n$ is een consistente schatter (hier heb je enkel Chebychev voor nodig)

Extra: indien $E(\hat{\vartheta}_n) \underset{n\to\infty}{\to} \vartheta$

en $\text{var}(\hat{\vartheta}_n) \underset{n\to\infty}{\to} 0$ (: de spreiding moet naar nul gaan als n naar oneindig gaat)

dan is $\hat{\vartheta}_n$ consistent.

Stel : $\hat{\vartheta}_1$ en $\hat{\vartheta}_2$ zijn twee onvertekende schatters:



Vraag: welk kies je best? $\hat{\vartheta}_1$ of $\hat{\vartheta}_2$?

→ we verkiezen $\hat{\vartheta}_2$ vanwege de kleinere spreiding : naukeuriger

$\bar{X}_n$ (als een schatter voor μ) is van alle lineaire onvertekende schatters de meest efficiente : efficiëntie is gekoppeld aan spreiding: meest efficiente = de kleinste spreiding

Bewijs: stel $\hat{\vartheta} = \sum a_i X_i$ een lineaire schatter

$E(\hat{\vartheta}) = (\sum a_i)\mu$

$\sum a_i = 1$ ← onvertekend (want: $E(\hat{\vartheta}) = 1\mu$)

$\text{var}(\hat{\vartheta}) = \text{var}\left(\sum a_i X_i\right)$

$= \sum a_i^2 \text{ var } X_i$ (: optelregel voor een onafhankelijke stochast $\text{var}(ax) = a^2 \text{ var } x$)

omdat $\text{var } X_i = \sigma^2$:

$= \sigma^2 \sum a_i^2$

Oefening: minimaliseer $\text{var}(\hat{\vartheta}_n) = \sigma^2 \sum a_i^2$

onder de rand voorwaarden : $\sum a_i = 1$

Lagrange : $L = \sum a_i^2 - \lambda\left(\sum a_i - 1\right)$

$L'_{a_i} = 2a_i - \lambda = 0 \qquad i = 1, 2, ..., n$

$L'_{\lambda} = -\sum a_i + 1 = 0$

$a_i = \frac{\lambda}{2}$

$a_i = \frac{1}{n}$

$\hat{\vartheta}_n = \frac{X_1 + X_2 + ... + X_n}{n}$

$\bar{X}$ als schatter voor de populatie gemiddelde (μ) is : (p211)
* onvertekend
* consistent (spreiding=0)
* efficient (lineaire schatters)
→ BEST LINEAR UNBIASED ESTIMATOR : **BLUE**

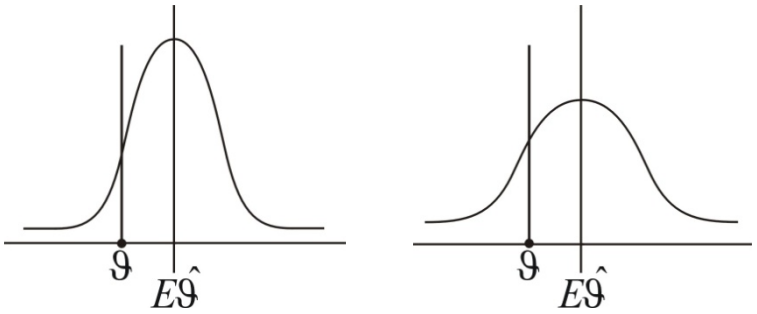
Let op: soms is een vertekende schatter beter dan een onvertekende schatter (zie cursus)

Vraag: Bestaat er altijd een onvertekende ($E(\hat{\vartheta}) = \vartheta$) schatter? NEEN
Oefening: $X \sim Exp(\lambda)$
$\vartheta = \lambda$: is vertekend

Indien vertekend : dan gebruiken we het "gemiddelde kwadratische fout":
: "mean squared error" of MSE

$$\begin{aligned} MSE(\hat{\vartheta}) &= E(\hat{\vartheta} - \vartheta)^2 \\ &= E(\hat{\vartheta}^2 - 2\vartheta\hat{\vartheta} + \vartheta^2) \\ &= E(\hat{\vartheta}^2) - 2\vartheta E(\hat{\vartheta}) + \vartheta^2 \\ &= E(\hat{\vartheta}^2) - E(\hat{\vartheta})^2 + (E\hat{\vartheta})^2 - 2\vartheta E(\hat{\vartheta}) + \vartheta^2 \\ &= \text{var}\,\hat{\vartheta} + (E\hat{\vartheta} - \vartheta)^2 \\ &= \text{variantie + vertekening}^2 \end{aligned}$$

(indien de vertekening = 0 dan kijken we enkel naar de variantie = de efficientie)



Dus: $\boxed{MSE(\hat{\vartheta}) = E(\hat{\vartheta} - \vartheta)^2 = \text{var}\,\hat{\vartheta} + (E\hat{\vartheta} - \vartheta)^2}$: niet in formularium

---

X is een stochastishe variabele

Populatie parameter	Schatter
$\mu = E(X)$	$\hat{\mu} = \bar{X} = \frac{X_1 + ... + X_n}{n}$
$p[X \sim b(1, p)]$	$\hat{p} = \bar{X}$
$\sigma^2 = \text{var}\,X = E(X - \mu)^2$	??

$\tilde{S}^2 = \dfrac{\sum (X_i - \bar{x})^2}{n}$: variantie met n weging (wordt minder gebruikt dan die met n-1 weging)

: is een onderschatter met fout $\dfrac{n-1}{n}$

$$E(\tilde{S}^2) = \frac{1}{n} E\left(\sum (X_i - \bar{x})^2\right)$$

We voegen μ toe (de som van μ is 0):

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

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

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

$$= \frac{1}{n}\left[n \operatorname{var} X_1 - 2n \operatorname{var} \bar{X} + n \operatorname{var} \bar{X}\right]$$

$$= \frac{1}{n}\left[n\sigma^2 - n\frac{\sigma^2}{n}\right]$$

$$\boxed{E(\tilde{S}^2) = \frac{n-1}{n}\sigma^2} \quad < \sigma^2$$

→ we stellen vast dat $\dfrac{n-1}{n}\sigma^2$ kleiner is dan σ^2

→ $\tilde{S}^2$ onderschat σ^2 : vertekening

<u>Besluit</u> : $\tilde{S}^2$ is vertekend, met neiging om σ^2 te onderschatten $\left(E(\tilde{S}^2) < \sigma^2\right)$

* $E(\tilde{S}^2) = \dfrac{n-1}{n}\sigma^2$

$$\frac{n-1}{n} E(\tilde{S}^2) = \sigma^2$$

$$E\left(\frac{n-1}{n}\tilde{S}^2\right) = \sigma^2$$

$$E(\tilde{S}^2) = \sigma^2$$

→ S^2 is een onvertekende schatter voor σ^2

$S^2 = \dfrac{\sum (X_i - \bar{X})^2}{n-1}$: $n-1$ weging

Conclusie :

$\tilde{S}^2$ = vertekend (neiging om te onderschatten)
= consistent ($\hat{\vartheta} \to \vartheta$ voor $n \to \infty$)
= asymptotisch normaal

S^2 = onvertekend
= consistent
= asymptotisch

13.4 Constructie van schatters (HB p221)

→ MW-schatter : de schatter van meeste waarschijnlijkheid
= ML estimator : maximum likelihood estimator

Basisveronderstelling: de verdeling van X is parametrisch

vb. $X \sim N(\mu, \sigma^2)$: continu stochast
$X \sim Pois(\lambda)$: discrete stochast
$X \sim \Gamma(\alpha, \beta)$: continu stochast

3 voorbeelden om de basisstructuur mee te geven:

$\bar{X}$ = onvertekend
= consistent
= asymptotisch normaal : BLUE

$\hat{p}$ = onvertekend
= consistent
= asymptotisch normaal

Voorbeeld 1: Bernoulli verdeling:

$X \sim b(1, p)$
: men wenst een populatie parameter ϑ te schatten op basis van een datasteekproef
schatter voor $\vartheta = p$
geobserveerde steekproef $\{1,1,1,0,1,0,0,1,0\}$

Vraag: wat is de kans om een dergelijke steekproef (van lengte g) te observeren?

Stap 1 : $L(1,1,1,0,1,0,0,1,0; p)$ (: L=likelihood : waarschijnlijkheid)

$p(k) = \binom{n}{k} p^k q^{n-k}$: dichtheid van Bernouilli (: in het formularium p3)

$= \binom{9}{5} p^5 (1-p)^4$ (: experiment wordt 9 keer herhaald met 5 keer success)

Welke waarde van p maakt L het grootst?

Stap 2 : Maximaliseer $\ln L = \ln \binom{9}{5} + 5 \ln p + 4 \ln(1-p)$

$$(\ln L)' = \frac{5}{p} - \frac{4}{1-p} = \frac{5-9p}{p(1-p)}$$

p	0	$p^* = 5/9$	1
$(\ln L)'$	+	0	-
$\ln L$	↗	max	↘

$\hat{p}\big|_{obs} = p^* = \frac{5}{9} \quad (=\bar{x})$

5/9 is de kans om dergelijke steekproef te observeren

*zie HB p222 voor de hele procedure.

Voorbeeld 2: Poisson verdeling:

$X \sim Pois(\lambda)$, $\vartheta = \lambda$, $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$

Vraag: wat is de kans om $\{x_1, x_2, ..., x_n\}$ te observeren?

Stap 1 : Likelihood : $L(x_1, x_2, ..., x_n; \lambda) = (e^{-\lambda})^n \frac{\lambda^{x_1+x_2+...+x_n}}{x_1!x_2!...x_n!}$

Stap 2 : $\ln L = -n\lambda + (x_1 + x_2 + ... + x_n)\ln\lambda - \ln(noemer)$

Stap 3 : Maximaliseer $\ln L$
λ

$$(\ln L)' = \frac{d}{d\lambda}\ln L = -n + (x_1 + x_2 + ... + x_n)\frac{1}{\lambda} - 0$$

λ	$\lambda^* = \frac{x_1 + x_2 + ... + x_n}{n}$	
$(\ln L)'$	+ 0	-
$\ln L$	↗ max	↘

$$\hat{\lambda}\big|_{obs} = \lambda^* = \frac{x_1 + x_2 + ... + x_n}{n} = \bar{X}\big|_{obs}$$

Voorbeeld 3: Exponentiële verdeling:

$X \sim Exp(\lambda)$, $\vartheta = \lambda$, $p_k = \lambda e^{-\lambda x} \quad (x \geq 0)$

Vraag: wat is de 'kans' om $\{x_1, x_2, ..., x_n\}$ te observeren?

Stap 1 : Likelihood : $L(x_1, x_2, ..., x_n; \lambda) = \lambda e^{-\lambda x_1}.\lambda e^{-\lambda x_2}...\lambda e^{-\lambda x_n}$

: niet als een kans maar als dichtheid

Stap 2 : $\ln L = n\ln\lambda - \lambda(x_1 + x_2 + ... + x_n)$

Stap 3 : Maximaliseer $\ln L$
λ

$$(\ln L)' = \frac{d}{d\lambda}\ln L = \frac{n}{\lambda} - (x_1 + x_2 + ... + x_n)$$

λ	$\frac{n}{(x_1 + x_2 + ... + x_n)}$	
$(\ln L)'$	+ 0	-
	↗ max	↘

$$\hat{\lambda}\big|_{obs} = \lambda^* = \frac{n}{x_1 + x_2 + ... + x_n} = \frac{1}{\bar{X}}\big|_{obs}$$

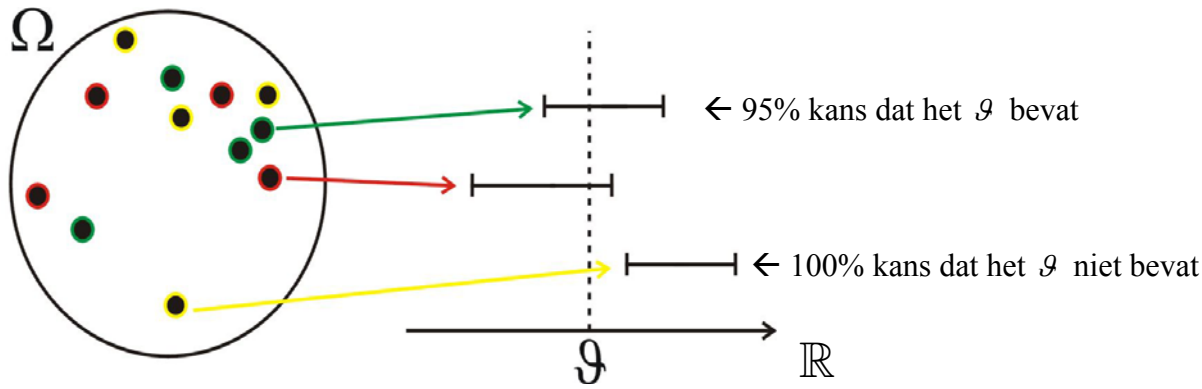
14. Betrouwbaarheidsintervallen (HB p227)
→ informatie toevoegen aan puntschattingen

Vraag: hoe groot is de kans dat de puntschatter waarden genereert die ver van ϑ liggen?
(: ϑ kunnen we niet zomaar observeren, wel via steekproef)

(Ω, G, P)
$X : \Omega \to \mathbb{R}$
ϑ is een populatieparameter
steekproef: $X_1, X_2, ..., X_n$
$0 < \alpha < 1$

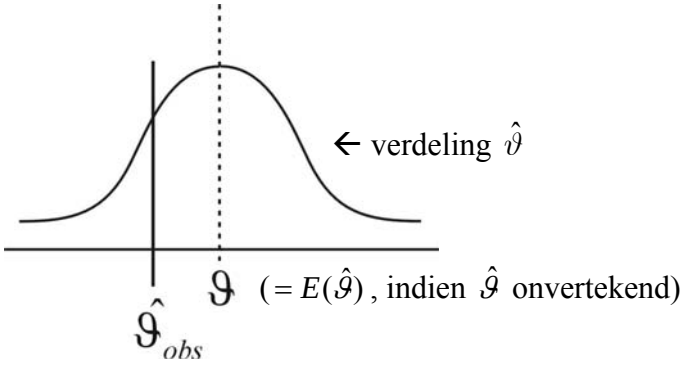
b.i = betrouwbaarheidsinterval
ϑ = parameter
$\hat{\vartheta}$ = stochastische variabele

Een $(1-\alpha).100\%$ betrouwbaarheidsinterval voor ϑ is een interval:
$[a_L = a_L(X_1, X_2, ..., X_n); a_R = a_R(X_1, ..., X_n)]$
zodat $P(a_L \leq \vartheta \leq a_R) = 1 - \alpha$
links rechts
* typisch gebruiken we een 95% betrouwbaarheids interval: $\alpha = 0,05$



ϑ is de waarde van de populatieparameter (die we niet observeren).
in α x 100% van de gevallen zal het b.i. $[a_L, a_R]$ de waarde ϑ NIET bevatten.

■ optellen van een betrouwbaarheids interval:
- puntschatter (st.v.)
- verdeling van de puntschatter



■ meestal is een betrouwbaarheids interval symmetrisch rond de puntschatting / puntschatter:

$\boxed{P\left(\left[\hat{\vartheta}-\varepsilon,\hat{\vartheta}+\varepsilon\right] \text{ bevat } \vartheta\right)=1-\alpha}$: 95% betrouwbaar : $\varepsilon = 1{,}96.\frac{0{,}4}{\sqrt{30}}$

Voorbeelden en oefeningen

Voorbeeld 1 : b.i. voor μ indien de stochast normaal verdeeld is en indien σ^2 gekend is (HB p227.4)

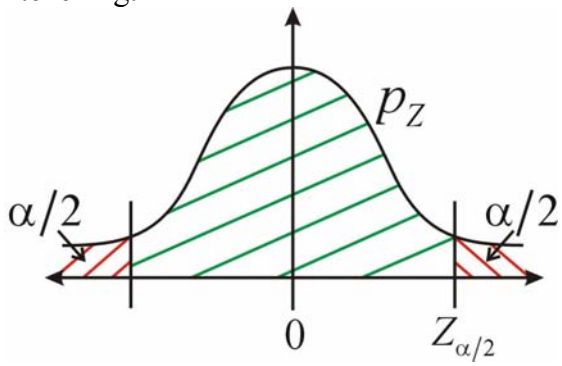
$X \sim N(\mu,\sigma^2)$ en $\vartheta = \mu$ (parameter) en σ^2 zijn gekend

■ schatter voor μ : $\bar{X}$: we vertrekken van een puntschatter

■ verdeling van $\bar{X}$: $\bar{X} \sim N(\mu,\frac{\sigma^2}{n})$

■ tabel: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$: we herschalen naar een standaard normale verdeling

■ tekening:



■ besluit: $P\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha$

Herformuleren: (oplossen naar μ):

$a_L = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad a_R = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

noteer : $\mu = \bar{X} + -z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

(: zelfde als: $1-\alpha = \hat{\vartheta} \pm \varepsilon$)

vb. een rekenvoorbeeld:
colaflesjes met inhoud 25cl
$X : \omega \to X(\omega) =$ inhoud van flesje ω in cl
$X \sim N(\mu,\sigma^2 = 0.16)$: σ^2 is in dit geval gekend
: je kan normaal niet zomaar veronderstellen dat je σ^2 kent.

We gebruiken een 95% betrouwbaarheids interval voor μ.
Data: $n = 30$, $\bar{x} = 24,8$: steekproef van 30 met gemiddelde 24,8

Uitkomst: $\alpha = 0,05$, $\alpha/2 = 0,025$

$z_{\alpha/2} = 1,96$ (: vanuit tabel)

$$\mu = \mu \pm \varepsilon = 24,8 \pm 1,96.\frac{0,4}{\sqrt{30}}$$

$$\mu \in [24,60; 24,96]$$

Formule : $\mu = \bar{X} \pm z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$

→ vergelijken met 25,00 cl
→ μ is voor 95% betrouwbaar dus 95 keer op 100 zal het flesje gevuld zijn tussen 24,60 cl en 24,96 cl.
→ we kunnen dus met 95% zekerheid zeggen dat de machine te weinig geeft.

* Opmerking:
Examenvraag: Stel een betrouwbaarheids interval op vertrek van puntschatter.
Dus niet enkel de formule invullen: alles opbouwen.

■ Symmetrische betrouwbaarheids interval voor μ : $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

comparatieve statica : wat gebeurt er als je parameters verandert?
$\alpha = 0,05$ $z_{\alpha/2} = 1,96$ 95% betrouwbaarheids interval
↓ een kleinere α
$\alpha = 0,01$ $z_{\alpha/2} = 2,58$ 99% betrouwbaarheids interval

■ halve intervalbreedte: $I = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

als n wordt vervangen door $4n$ dan wordt I vervangen door $\frac{1}{2}I$.

→ afnemende schaalopbrengsten

■ hoe groot moet "n" zijn zodat $I < \varepsilon$?

$$I = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \varepsilon$$

$$\frac{z_{\alpha/2}\sigma}{\varepsilon} < \sqrt{n}$$

■ het betrouwbaarheids interval voor μ: $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is bruikbaar zodra $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Voorwaarden:
CLS : n groot (: als n groot genoeg is dan is $\bar{X}$ normaal verdeeld)
σ^2 gekend

Voorbeeld 2 : b.i. voor μ indien de stochast normaal verdeeld is en indien σ^2 niet gekend is (HB p231)

$X \sim N(\mu,\sigma^2)$ en $\vartheta = \mu$ en σ^2 is niet gekend (dit is meer realistisch)

→ σ^2 is wel te schatten via de steekproef als $S^2 = \frac{(x_1 - \overline{x})^2}{n-1}$

■ schatter voor μ : $\overline{X}$

■ verdeling van $\overline{X}$: $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

■ tabel: $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$
LL RL

vergelijking: RL = gekend (tabel)
LL = $\overline{X}$ observeerbaar
n onder controle
μ en σ^2 zijn ongekend
→ 1 vergelijking in 2 onbekende

Vraag: σ^2 schatten door S^2 (: steekproef variantie met n-1 weging (dit is nodig om de onderschatting te vermijden)
→ σ wordt vervangen door S .
→ $\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim ?$: welke soort verdeling bekom je dan?
: $\overline{X}$, S en n ken je, dus de enige onbekende is μ .

Vraag: verdeling van $S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$ (: n-1 is nodig om de onderschatting te vermijden)
Als $Z_1, Z_2, ..., Z_n$ onafhankelijk identiek verdeeld $N(0,1)$
dan $\sum Z_i^2 \sim \mathrm{X}_n^2$: met n vrijheidsgraden
Dus $\sum \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \mathrm{X}_n^2$ (: formularium)

Hulpstelling (bewijs) : $+\overline{X} - \overline{X}$

$$\sum (X_i - \mu)^2 = \sum \left(X_i - \overline{X} + \overline{X} - \mu \right)^2$$
$$= \sum \left(X_i - \overline{X} \right)^2 + 2\left(\overline{X} - \mu \right) \sum \left(X_i - \overline{X} \right) + n\left(\overline{X} - \mu \right)^2$$
$$\sum (X_i - \mu)^2 = \sum \left(X_i - \overline{X} \right)^2 + n\left(\overline{X} - \mu \right)^2$$

$$\sum\left(\frac{X_i-\mu}{\sigma}\right)^2=\sum\left(\frac{X_i-\bar{X}}{\sigma}\right)^2+\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

$$\chi^2_n \qquad \chi^2_{n-1} \qquad \chi^2_1$$

Moment Genereerende Functie

X_i o.i.v. Normaal

Dan onafhankelijk : $\sum\left(X_i-\bar{X}\right)^2$ en $\bar{X}$

Linker Lid : $M(t)=(1-2t)^{-n/2}$

Rechter Lid : $M(t)=(1-2t)^{-1/2}$ enkel als ze onafhankelijk zijn

$X\sim N(\mu,\sigma^2)$

- $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$
- $\sum\left(\dfrac{X_i-\bar{X}}{\sigma}\right)^2=(n-1).\dfrac{1}{n-1}.\dfrac{(X_i-\bar{X})^2}{\sigma^2}=\dfrac{(n-1)S^2}{\sigma^2}\sim \chi^2_{n-1}$

Nu hebben we twee vergelijkingen in twee onbekenden (μ en σ^2):

$$\frac{\left(\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}\Big/n-1}}\sim t_{n-1} \qquad \left(\frac{N(0,1)}{\sqrt{\chi^2_{n-1}/n-1}}\right)$$

$\downarrow\ \sigma$ verdwijnt, na schrappen:

$T=\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\sim t_{n-1}$:T verdeling is ook symmetrisch maar heeft dikkere staarten dan de Normale

$Z=\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$

Definierende stelling:

<u>Zij:</u> $Z\sim N(0,1)$

$Y\sim \chi^2_n$ (: Y is de som van Z : $Z_1+Z_2+...+Z_n$)

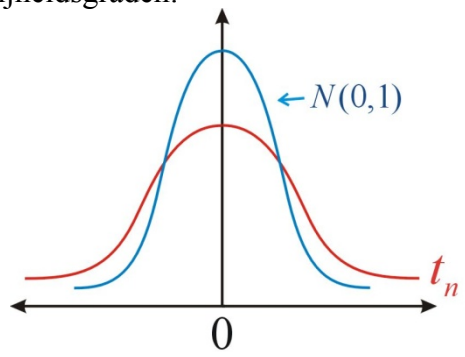
Z en Y zijn onafhankelijk

<u>Dan:</u> $T=\dfrac{Z}{\sqrt{Y/n}}=\dfrac{N(0,1)}{\chi^2_n/n-1}\sim t_n$

t-verdeling met n vrijheidsgraden:

$E(T) = 0$

$\text{var}(T) = \dfrac{n}{n-2}$



: als het aantal vrijheidsgraden groter wordt (naar oneindig) : valt t_∞ samen met $N(0,1)$

: we gebruiken de T verdeling ipv de normale verdeling bij kleine steekproeven

Besluit:

■ Indien σ^2 gekend is dan $N(0,1)$-verdeelde puntschatter.

: schatter $\bar{X}$

■ Indien σ^2 niet gekend, dan wordt σ^2 geschat door S^2, en een t_{n-1} verdeelde puntschatters

: schatter S^2

vb. Ω = verzameling van dozen waspoeder gevuld door een bepaalde machine

$X : \Omega \to \mathbb{R} : \omega \to X(\omega)$: inhoud van doos ω in gram

$X \sim N(\mu, \sigma^2)$

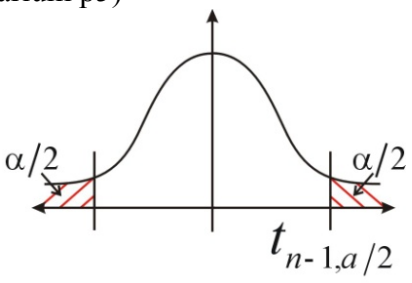
Vraag: zoek het 95% betrouwbaarheids interval voor μ

Data: $n = 8,\ \bar{x} = 2020,\ s^2 = 400$

Oplossing: $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$\sum\left(\dfrac{X_i - \bar{X}}{\sigma}\right)^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim X^2_{n-1}$ (: in het formularium p5)

$T = \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$



$P\left(-t_{n-1,\sigma/2} \le \dfrac{\bar{X} - \mu}{S/\sqrt{n}} \le t_{n-1,\sigma/2}\right) = 1 - \alpha$

betrouwbaarheidsinterval voor μ : $\mu = \bar{X} \pm \dfrac{S}{\sqrt{n}} . t_{n-1,\alpha/2}$: niet in formularium

$$\bar{X}\big|_{obs} \pm \frac{S\big|_{obs}}{\sqrt{n}} . t_{n-1,\alpha/2}$$

$$2020 \pm \frac{20}{\sqrt{8}} 2{,}365 = 2020 \pm 16{,}7$$

Conclusie:

→ met 95% zekerheid kunnen we zeggen dat μ in het interval $[2003,3;2036,7]$ ligt.

→ het machine "geeft" te veel.

Voorbeeld 3 :

b.i. voor $\mu_1 - \mu_2$ indien de stochasten normaal verdeeld zijn en indien σ_1^2 en σ_2^2 gekend zijn

$X_1 \sim N(\mu_1, \sigma_1^2)$
$X_2 \sim N(\mu_2, \sigma_2^2)$ zijn onafhankelijk

$\vartheta = \mu_1 - \mu_2$; σ_1 en σ_2 gekend dus $\sim N(0,1)$

■ schatter voor $\vartheta = \bar{X}_1 - \bar{X}_2$

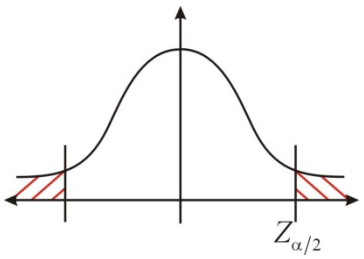
■ verdeling: $\bar{X}_i \sim N(\mu_i, \frac{\sigma_i^2}{n_i})$

$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

■ tabel : $Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

→ één vergelijking in één onbekende (μ)

■ tekening :



■ besluit : $P\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) = 1 - \alpha$

■ betrouwbaarheids interval voor $\mu_1 - \mu_2$: $\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

analoog $\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$

Voorbeeld 4 :

b.i voor $\mu_1 - \mu_2$ indien de stochasten normaal verdeeld zijn en indien $\sigma_1^2 = \sigma_2^2$ niet gekend zijn

$X_1 \sim N(\mu_1, \sigma_1^2)$
$X_2 \sim N(\mu_2, \sigma_2^2)$ zijn onafhankelijk

$\vartheta = \mu_1 - \mu_2$; $\sigma_1^2 = \sigma_2^2$ niet gekend

→ we hebben nu minder informatie dus gebruiken we de T verdeling (minder betrouwbaar)

uit voorbeeld 3 : $Z = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}}} = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\dfrac{\sigma}{\sqrt{n_1}} + \dfrac{\sigma}{\sqrt{n_2}}} \sim N(0,1)$

→ één vergelijking in <u>twee</u> onbekende (μ en σ)

$$\frac{(n_1-1)S_1^2}{\sigma^2} + \frac{(n_2-1)S_2^2}{\sigma^2} \sim \mathrm{X}^2_{n_1+n_2-2} \sim \mathrm{X}^2_{(n_1-1)+(n_2-1)}$$

$$E\left(\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\right) = (n_1+n_2-2)\sigma^2$$

$$T = \frac{\dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma^2}{n_1} + \dfrac{\sigma^2}{n_2}}}}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}} \sim t_{n_1+n_2-2} \quad : \sigma \text{ wegdelen}$$

■ betrouwbaarheids interval voor $\mu_1 - \mu_2$:

$$\bar{X}_1 - \bar{X}_2 \pm t_{n_1+n_2-2,\alpha/2} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} . \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

<u>Voorbeeld 5 :</u> GEEN LEERSTOF

b.i voor $\mu_1 - \mu_2$ indien de stochasten normaal verdeeld zijn en indien σ_1^2 <u>en</u> σ_2^2 <u>niet</u> gekend zijn

$X_1 \sim N(\mu_1, \sigma_1^2)$
$X_2 \sim N(\mu_2, \sigma_2^2)$ zijn onafhankelijk
$\vartheta = \mu_1 - \mu_2$; σ_1^2 en σ_2^2 zijn niet gekend

→ we hebben nu minder informatie dus gebruiken we de T verdeling (minder betrouwbaar)
→ één vergelijking met 3 onbekende

■ vergelijkingnen :

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1) \quad : \text{CLS}$$

$$Y = \frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \mathrm{X}^2_{n_1+n_2-2} \quad : \text{formularium}$$

de deling $\dfrac{Z}{\sqrt{Y/n_1+n_2-2}}$ werkt niet meer

: we kunnen σ_1^2 en σ_2^2 niet meer wegdelen uit de noemer zoals in de vorige oefening

Behrens-Fisher probleem:

→ we vervangen σ_1^2 en σ_2^2 door S^2

in Z de varianties σ_i^2 schatten door S_i^2

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_p \quad \text{(literatuur)}$$

$\frac{\bar{X}_i - \mu_i}{\sigma_i / \sqrt{n_i}} \sim N(0,1)$: S_i is een consistente schatter voor σ_i

$\frac{\bar{X}_i - \mu_i}{S_i / \sqrt{n_i}} \sim t_{n_i - 1} \approx N(0,1)$: voor n_i groot (CLS)

→ het enige verschil tussen de N en t_{n-1} verdelingen is dikkere staarten, maar als n groot is dan benadert de t_{n-1} ook de normale verdeling N (CLS).

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1)$$: benaderend betrouwbaarheids interval

Voorbeeld 6 :

$X \sim N(\mu, \sigma^2)$ (: nu hebben we maar één stochast)

$\vartheta = \sigma^2$ (: vorige oefening $\vartheta = \mu$)

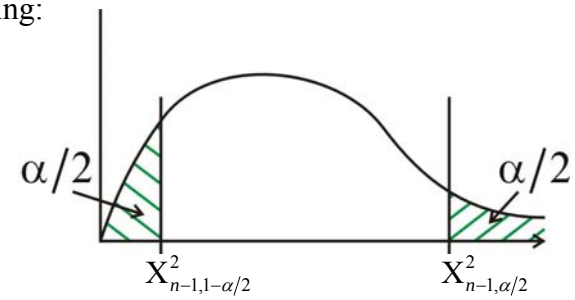
■ schatter : S^2

→ één vergelijking in één onbekende (σ^2)

■ vergelijkingen : $Y = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

→ het doel is om het rechter lid te kennen zodat we de tabel kunnen gebruiken

■ tekening:



■ tabel:

$$P\left(\chi^2_{n-1,1-\alpha/2} \le Y \le \chi^2_{n-1,\alpha/2}\right) = 1 - \alpha$$

■ betrouwbaarheids interval voor σ^2 :

$$\frac{X^2_{n-1,1-\alpha/2}}{(n-1)S^2} \le \frac{1}{\sigma^2} \le \frac{X^2_{n-1,\alpha/2}}{(n-1)S^2}$$

$$\frac{(n-1)S^2}{X^2_{n-1,1-\alpha/2}} \ge \sigma^2 \ge \frac{(n-1)S^2}{X^2_{n-1,\alpha/2}}$$: de ongelijkheids tekens veranderen van richting

$$\left[\frac{(n-1)S^2}{X^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{X^2_{n-1,1-\alpha/2}}\right]$$ met $1-\alpha$ zekerheid

<u>Rekenvoorbeeld:</u>

$n=10$, $\bar{X}\big|_{obs}=100$, $S\big|_{obs}=5,50$

betrouwbaarheids interval voor $\vartheta=\sigma^2$, $\alpha=0,05$, $\frac{\alpha}{2}=0,0025$

■ tabel: $X^2_{9;0,025}=19,02$

$X^2_{9;0,975}=2,70$

betrouwbaarheids interval: $\left[\frac{(n-1)S^2}{X^2_{n-1,\alpha/2}} \quad , \quad \frac{(n-1)S^2}{X^2_{n-1,1-\alpha/2}}\right]$

$$= \frac{9 \text{ x } (5,5)^2}{19,02} \qquad = \frac{9 \text{ x } (5,5)^2}{2,7}$$

$$= 14,31 \qquad = 100,8$$

<u>Voorbeeld 7 :</u> b.i. voor een proportie p (kans op succes in een Bernoulli experiment)

$X \sim b(1,p)$

$\vartheta = p$, $\hat{p} = \bar{X}$

We weten dus dat $\mu = n.p$ en $\sigma^2 = n.p.q$

■ vergelijking: $X_1, X_2, ..., X_n$ is een steekproef

$$X_1 + X_2 + ... + X_n \sim b(n,p)$$

$$\approx N(np, \sigma^2 = npq)$$

$$\bar{X} \approx N(p, \sigma^2 = \frac{pq}{n})$$

→ als n voldoende groot is kunnen we de normale verdeling gebruiken

→ CLS : som van onafhankelijke stochasten als n groot is dan gaat de benadering naar een normale verdeling $N(0,1)$

$\frac{\bar{X}_n - \mu_X}{\sigma/\sqrt{n}}$, n=1 dus valt weg: $\boxed{Z = \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)}$

→ één vergelijking in één onbekende

■ oplossen naar een betrouwbaarheids interval:

$$P\left(-z_{\alpha/2} \le \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \le z_{\alpha/2}\right) = 1-\alpha$$

$$\left(\frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}}\right)^2 \le z^2 \qquad (z = z_{\alpha/2})$$

■ ongelijkheid oplossen naar p : (n, $\bar{X}, z$ zijn observeerbaar : gekend)

$$\frac{n}{p(1-p)}(\bar{X}-p)^2 \le z^2$$
$$n(\bar{X}-p)^2 \le z^2 p(1-p)$$
$$n(\bar{X}^2 - 2\bar{X}p + p^2) \le z^2(p-p^2)$$
$$n\bar{X}^2 - 2n\bar{X}p + np^2 \le z^2(p-p^2)$$
$$n\bar{X}^2 - 2n\bar{X}p + np^2 \le z^2 p - Z^2 p^2$$
$$n\bar{X}^2 - 2n\bar{X}p + np^2 - z^2 p + z^2 p^2 \le 0$$
$$f(p) = (n+z^2)p^2 - (2n\bar{X} - z^2)p + n\bar{X}^2 \le 0$$ (de grafiek van f is parabool)

Nul punten zoeken p_1 en p_2 :

$$p_{1,2} = \frac{2n\bar{X} + z^2 \pm \sqrt{(2n\bar{X}+z^2)^2 - 4n(n+z^2)\bar{X}^2}}{2(n+z^2)}$$
$$= \frac{2n\bar{X} + z^2 \pm \sqrt{4n\bar{X}z^2 + z^4 - 4nz^2\bar{X}^2}}{2(n+z^2)}$$

$$\boxed{p_{1,2} = \frac{2n\bar{X} + z^2 \pm z\sqrt{z^2 + 4n\bar{X}^4 - 4n\bar{X}^2}}{2(n+z^2)}}$$

: betrouwbaarheids interval voor $p : [p_1, p_2]$

$$p_{1,2} = \frac{2n\bar{X} + z^2 \pm z\sqrt{z^2 + 4n\bar{X} - 4n\bar{X}^2}}{2(n+z^2)}$$

teller en noemer delen door $2n$:

$$p_{1,2} = \frac{\bar{X} + \frac{z^2}{2n} \pm z\sqrt{\frac{z^2}{4n^2} + \frac{\bar{X}(1-\bar{X})}{n}}}{1+\frac{z^2}{n}}$$

we schrappen $\frac{z^2}{2n}$, $\frac{z^2}{n}$ en $\frac{z^2}{4n^2}$ als de steekproef groot genoeg is (n groot)

$$\boxed{p_{1,2} \approx \bar{X} \pm z\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}}$$

benaderend betrouwbaarheids interval op basis van:

$\frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})/n}} \approx N(0,1)$: via CLS

$\frac{\bar{X}-p}{\sigma_{geschat}/\sqrt{n}} \approx N(0,1)$: indien n klein is

met $\sigma^2_{geschat} = \frac{\bar{X}(1-\bar{X})}{n} = p(1-p)$

<u>Rekenvoorbeeld:</u>

$n = 400 \qquad \hat{p}\big|_{obs} = \bar{X}\big|_{obs} = \frac{320}{400} = 0,8$

$\alpha = 0,01$

* CLS + kwadratische vergelijking : $[0,74;0,846]$

* CLS + σ^2 schatten: $[0,75;0,85]$

→ nauwkeuriger betrouwbaarheids interval (zeker voor "kleine" n)

halve intervalbreedte:

$$I = Z\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \qquad \bar{X}(1-\bar{X}) \le \frac{1}{4}$$

$$\le Z\sqrt{\frac{1}{4n}}$$

indien gewenst is dat $I \le \varepsilon$ dan kan men (conservatief) opleggen dat: $\frac{Z}{2\sqrt{n}} \le \varepsilon$

$$\boxed{\frac{Z}{2\sqrt{\varepsilon}} \le \sqrt{n}}$$

<u>Voorbeeld 8 :</u>

$X_1 \sim b(1, p_1)$
$X_2 \sim b(1, p_2)$ zijn onafhankelijk
$\vartheta = p_1 - p_2$

■ schatter : $\hat{\vartheta} = \bar{X}_1 - \bar{X}_2$

■ vergelijking : $\hat{\vartheta} = \bar{X}_1 + \bar{X}_2$

Stap 1: CLS : $\bar{X}_i \approx N(p_i, \frac{p_i(1-p_i)}{n_i})$

Stap 2: gebruik de onafhankelijkheid : $\bar{X}_1 - \bar{X}_2 \approx N\left(p_1 - p_2; \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$

■ benaderend b.i.

$$p_1 - p_2 = \bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{X}_1(1-\bar{X}_1)}{n_1} + \frac{\bar{X}_2(1-\bar{X}_2)}{n_2}}$$

Samenvatting : 3 soorten verdelingen

$$p = \bar{X} \pm z_{\alpha/2}\sqrt{\bar{X}\left(\frac{1-\bar{x}}{n}\right)}$$

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \mu = \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \ (\sigma^2 \text{ en } \mu \text{ niet gekend}) \qquad \mu = \bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

$$Y = \frac{(n-1)S^2}{\sigma^2} \sim X^2_{n-1} \ (\vartheta = \sigma^2 \text{ schatter} = S^2) \qquad \left[\frac{(n-1)S^2}{X^2_{n-1},\alpha/2}, \frac{(n-1)S^2}{X^2_{n-1},1-\alpha/2}\right]$$

15. Testen van hypothesen (HB p240)

* Roken verhoogt het risico op kanker.
* Het dragen van een gordel verlaagt het risico op ernstige verwondingen bij een ongeval.
* Het is opportuun voor een bedrijf om een nieuwe productielijn op te starten.
* ....

Vraag: Hoe "bewijzen" we dergelijke uitspraken?
- niet een mathematisch bewijs
- wel een betrouwbaarheidsnievau toekennen

vb. $\Omega =$ verzameling lampen uit Nieuwe productielijn
$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$: levensduur van lamp ω in uren
Stel $X \sim N(\mu_N, \sigma^2)$

Vraag: Moet de nulhypothese (H_0) $\mu_N \leq 1000$
- aanvaard worden?
- verworpen worden?

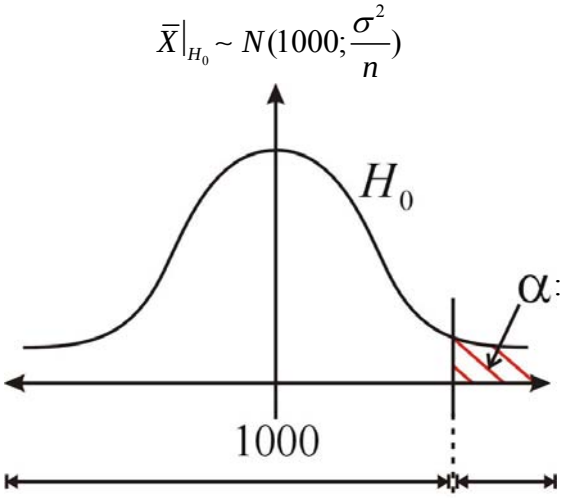
$H_0 : \mu_N \leq 1000$
$H_1 : \mu_N > 1000$
Aanvaard H_0 tenzij de geobserveerde data (labo) te slecht verklaard worden door H_0.
Verwerp H_0 indien de data te "ver" liggen van H_0

[Analoge oefening: munt die 60 keer K tost in 100 beurten. Is deze munt 'eerlijk'?]

$\mu_N \gtreqless 1000$?

Toetstochast : $\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n}$

Vraag: wat is de verdeling van deze toetsstochast indien H_0 waar "zou" zijn?



Notaties en definities

(Ω, G, P), $X : \Omega \to \mathbb{R}$

Hypothese: uitspraak over de verdeling van X

Voorbeelden: H: X is normaal verdeeld

$X \sim Exp(\lambda)$: enkelvoudig

$E(X) = \mu_0$: meervoudig

$E(X) \geq \mu_0$: meervoudig

■ parameter hypothese:

$X \sim b(1, p)$ $\quad H : p = 0,5$: enkelvoudig

$X \sim N(\mu, \sigma^2)$ $\quad H : \mu < \mu_0$: meervoudig

■ enkelvoudige hypothese legt de verdeling volledig vast:

$H : X \sim N(0,1)$

(* X normaal verdeelde kan enkelvoudig of meervoudig zijn.)

■ meervoudige hypothese

■ We willen H_1 bewijzen, we doen dat door H_0 te weerleggen.

→ Indien H_0 weerlegt wordt, is H_1 "bewezen"

Stap 1: de hypotheses definiëren:
H_0 : die hypothese die men wenst te weerleggen
: NUL-hypothese
H_1 : die hypothese die men wenst te bewijzen (door de andere te weerleggen)
: ALTERNATIEVE-hypothese

Stap 2: toetsstochast : $T = T(X_1, X_2, ..., X_n)$
Stap 3: verdeling van T onder de nulhypothese
Stap 4: kritisch gebied ← betrouwbaarheids niveau (bovenstaart kans) $(1-\alpha)$ x 100%
Stap 5: toets uitvoeren op basis van geobserveerde steekproef $\{x_1, x_2, ..., x_n\}$
Vraag: behoort $T(X_1, X_2, ..., X_n)$ tot het kritisch gebied?

Stap 6:
- Ja : verwerp H_0 : H_1 is dus "bewezen"
- Nee : verwerp H_0 niet: H_1 is dus niet "bewezen"

Voorbeelden en oefeningen

Voorbeeld 1:
Een machine snijt viltstiften. De voorgestelde lengte bedraagt 10cm.
Bij controle van de machine : n =16, $\overline{x} = 9,95$ cm
[$\sigma = 0,12$ cm : gekend (maar dit is onrealistisch)]

Vraag: moet deze machine bijgesteld worden?

Antwoord:
Stap 1: $\Omega =$ verzameling van stijften gesneden door die machine.
$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$: lengte van stift ω in cm
Veronderstel: X is normaal verdeeld

Stap 2:
H_0 : $\mu = E(X) = \mu_0 = 10$: de hypothese die men wilt weerleggen
H_1 : $\mu \neq \mu_0$: machine moet worden bijgesteld : de hypothese die men wilt bewijzen
$\alpha = 0,05$: betrouwbaarhieds interval; kies je zelf. Normaal kiezen we 95%.
Toetsstochast : $\overline{X}$

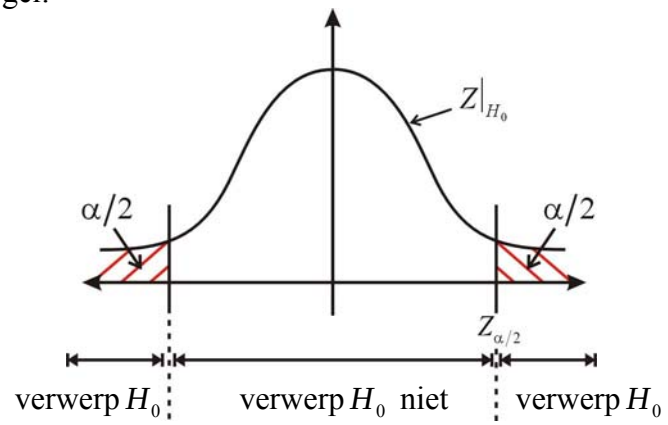
Stap 3: verdeling $\overline{X}$ onder H_0

$$\overline{X}\big|_{H_0} \sim N(\mu_0, \frac{\sigma^2}{n}) \sim N(10, \frac{(0,12)^2}{16})$$

transformatie van normale verdeling naar een standaard normale verdeling:

$$Z\big|_{H_0} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\Big|_{H_0} \sim N(0,1)$$

Stap 4: Beslissingsregel:



* Als H_0 verworpen wordt dan is H_1 bewezen.
* $Z_{\alpha/2}$ vind je in de tabel via de omgekeerde methode; je weet de kans en je moet weten welke plaats dat is op de verdeling.

Stap 5: Toets uitvoeren:
gegeven: $\bar{X} = 9,95,\ \mu_0 = 10,\ n = 16,\ \sigma = 0,12$
$\alpha = 0,05,\ \alpha/2 = 0,025,\ Z_{\alpha/2} = 1,96$

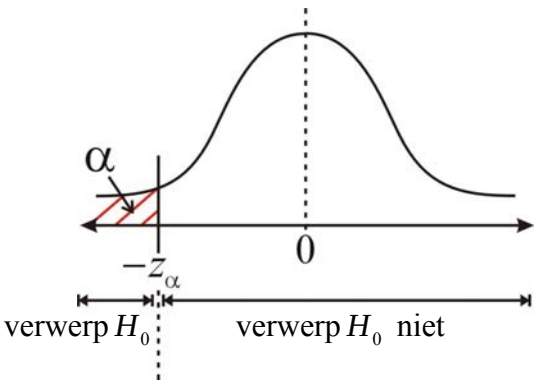
$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = Z_{obs} = \frac{9,95 - 10}{0.12/4} = -\frac{0,05}{0,03} = -1,66$$

$Z_{obs} = -1,66 \quad (> -1,96)$ dus het ligt niet in het kritisch gebied
→ H_0 kan niet verworpen worden.
→ de machine moet niet bijgesteld worden.

* Als $Z_{obs} < -1,96$ of $Z_{obs} > 1,96$ dan is het in het kritisch gebied en moet het machine bijgesteld worden, dit is niet het geval.

MAAR : H_0 : $\mu = \mu_0 = 10$
H_1 : $\mu < \mu_0$: we verwerpen nu alleen als het te klein is, niet als het te groot is
: in de vorige oefening gebruikten we $\mu \neq \mu_0$ dus zowel < als >
→ één zijdig verwerpingsgebied

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\Big|_{H_0} \sim N(0,1)$$



uitvoeren van deze toets:

$Z|_{obs} = -1,66 \quad -Z_\alpha = -1,64$

$Z|_{obs}$ ligt dus wel in het kritisch gebied en dus verwerpen we H_0.

→ deze machine moet wel bijgesteld worden.

Voorbeeld 2:

Iemand wil "bewijzen" dat het gewicht van 7-jarige vluchtelingen kinderen significant kleiner is dan 25kg.

$n = 10, \quad \bar{X}|_{obs} = 23,5, \quad S^2|_{obs} = 2,3$

Antwoord:

Stap 1: $\Omega =$ verzameling van 7-jarige vluchtelingenkinderen

$X : \Omega \to \mathbb{R}$

$\omega \to X(\omega)$: gewicht van vluchteling ω in kg

Stap 2: $H_0 : \ \mu = 25 = \mu_0 = E(x)$

$H_1 : \ \mu < 25$: kleiner dan, dus gebruik eenzijdig kritisch gebied

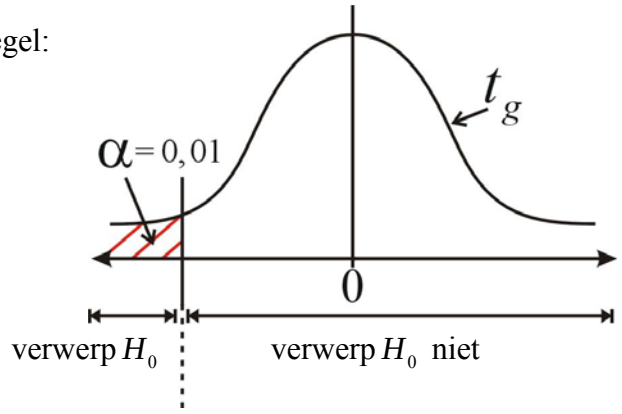
Stap 3: Toetsstochast : herschaald : $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$

Maar, σ^2 kennen we niet dus gebruiken we S^2 als een schatter

$T = \dfrac{\bar{X} - \mu_0}{S / \sqrt{n}}\Big|_{H_0} \sim t_{n-1} = t_g$: 1 vergelijking in 1 onbekende (μ_0)

→ omdat we σ^2 niet kennen gebruiken we de T verdeling ipv de standaard normale verdeling

Stap 4: Beslissingsregel:



Stap 5: Toets uitvoeren:

We gebruiken een 99% betrouwbaarheids interval

$-t_{9,0,01} = -2,821$: min teken omdat we in de tabel naar de boven staart kijken

: 9 vrijheidsgraden (n-1) met kans 0,01

$$T|_{obs} = \frac{23,5 - 25}{\sqrt{2,3} / \sqrt{10}} = -3,128$$

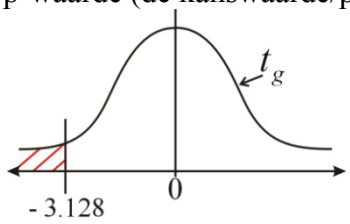
→ de waarde van $T|_{obs}$ ($-3,128$) ligt in het kritisch gebied

→ verwerp H_0

→ H_1 is dus bewezen (het gewicht van 7-jarige vluchtelingen kinderen is significant kleiner dan 25kg)

Vraag: Hoe groot is de kans om -3,128 te observeren indien $T|_{H_0} \sim t_g$?

p-waarde (de kanswaarde/probability):



wat is de kans dat je observeert in de rode zone (in het kritisch gebied):

$P(T \leq -3,128) = 0,006$ (via tabel) : oppervlakte onder de curve waar $T \leq -3,128$

= zeer onwaarschijnlijk.

Dus, "bewijs" voor hypothese H_1 op 99,4% betrouwbaarheids niveau.

Dus, als de waarde in het kritisch gebied ligt, kan je gaan kijken hoe ver hij in het rood staat.

De tabel van de T verdeling werkt omgekeerd van de tabel van de N verdeling :

vb. $Z_{0,025} = 1,96$: de algemene schrijfwijze

Je kent α (=0,025) dan moet je de waarde 0,025 gaan zoeken in al de rijen en kolommen van de tabel.

vb. $t_{9,0,01} = 2,821$

Je kent α (=0,01) dan moet je de waarde 0,01 gaan zoeken in de bovenste rij van de tabel en die kolom volgen tot op de rij van 9 vrijheidsgraden.

Voorbeeld 3:

Twee leermethodes vergelijken via de scores op een toets.

leermethode 1 : $n_1 = 15,\ \bar{X}_1|_{obs} = 65,3,\ S_1^2|_{obs} = 17,6$

leermethode 2 : $n_2 = 15,\ \bar{X}_2|_{obs} = 62,4,\ S_2^2|_{obs} = 15,2$

Vraag: Is methode 1 "beter" dan methode 2?

Antwoord:

Stap 1: Stel: $\bar{X}_i \sim N\left(\mu_1, \frac{\sigma_i^2}{n_i}\right)$

voor $i = 1,2$ onafhankelijk : $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

$\sigma_1^2 = \sigma_2^2$: omdat S_1^2 en S_2^2 dicht bij elkaar liggen, mag je dit veronderstellen

Stap 2: $H_0 : \mu_1 = \mu_2$: de twee methodes zijn even goed

$H_1 : \mu_1 > \mu_2$: methode 1 is beter dan methode 2

$\alpha = 0,05$: 95% betrouwbaarheids interval

Stap 3: $\left.\dfrac{\bar{X}_1 - \bar{X}_2}{\sigma.\sqrt{\dfrac{1}{n_1}+\dfrac{2}{n_2}}}\right|_{H_0} \sim N(0,1)$ (: we gebruiken σ omdat $\sigma_1^2 = \sigma_2^2 = \sigma$: gegeven)

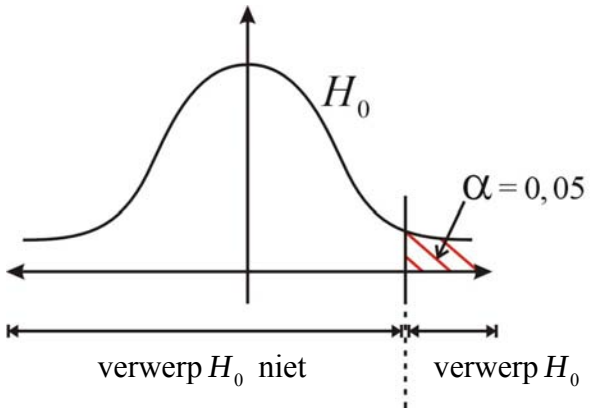
$$\left[Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}} \quad (: \mu_1 - \mu_2 = 0 \text{ wegens } H_0) \right]$$

We schatten σ^2 door: $S^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ (: formularium)

$$\frac{S^2}{\sigma^2} \sim X^2_{n_1+n_2-2}$$

$$T = \left.\frac{\bar{X}_1 - \bar{X}_2}{S.\sqrt{\dfrac{1}{n_1}+\dfrac{2}{n_2}}}\right|_{H_0} \sim t_{n_1+n_2-2}$$

Stap 4: Beslissingsregel:



Stap 5:

$T_{obs} = 1,96$

$t_{28,0.05} = 2,048$

Besluit: We verwerpen H_0 niet. Beide leermethodes zijn even goed.

* Was onze veronderstelling van $\sigma_1 = \sigma_2$ eigelijk wel juist? : GEEN LEERSTOF

$X_1 \sim N(\mu_1, \sigma_1^2)$ en $X_2 \sim N(\mu_2, \sigma_2^2)$ zijn onafhankelijk

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 > \sigma_2^2$

$$\frac{(n_i-1)S_i^2}{\sigma_i^2} \sim X^2_{n_i-1}$$

$$\left.\frac{(n_1-1)S_1^2/\sigma_1^2\,(n_1-1)}{(n_2-1)S_2^2/\sigma_2^2\,(n_2-1)}\right|_{H_0} \sim F(n_1-1, n_2-1)$$: Fisher : F-toets

Voorbeeld 4:

$X_1 \sim b(1, p_1)$
$X_2 \sim b(1, p_2)$ zijn onafhankelijk

Hypothese: $p_1 \gtrless p_2$

Twee vragen worden gesteld : A : Europa mag invoer uit USA beperken.
B : USA mag invoer uit Europa beperken.

Procent Ja-antwoorden:

Volgorde vraag	Ja op A	Ja op B	n
B dan A	24.2	14.0	178
A dan B	48.2	37.2	191

Vraag: Is de volgorde van de vragen belangrijk?

Stap 1:

$X_1 = \begin{matrix} 1 \leftarrow \text{ Ja op A in groep BA} \\ 0 \leftarrow \text{ Nee op A in groep BA} \end{matrix}$

$X_2 = \begin{matrix} 1 \leftarrow \text{ Ja op A in groep AB} \\ 0 \leftarrow \text{ Nee op A in groep AB} \end{matrix}$

Stap 2: toetsstochasten

$H_0 : p_1 = p_2 = p$

$H_0 : p_1 < p_2$

$\alpha = 0.01$: betrouwbaarheidsinterval van 99%

Stap 3:

$\bar{X}_1\big|_{H_0} \approx N(p, \sigma_1^2 = \frac{pq}{n_1})$ **(** : gegeven dat $p_1 = p_2 = p$) : CLS

$\bar{X}_2\big|_{H_0} \approx N(p, \sigma_2^2 = \frac{pq}{n_2})$

$\bar{X}_1 - \bar{X}_2\big|_{H_0} \approx N(0, \sigma_2^2 = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right))$: 0 wegens p-p

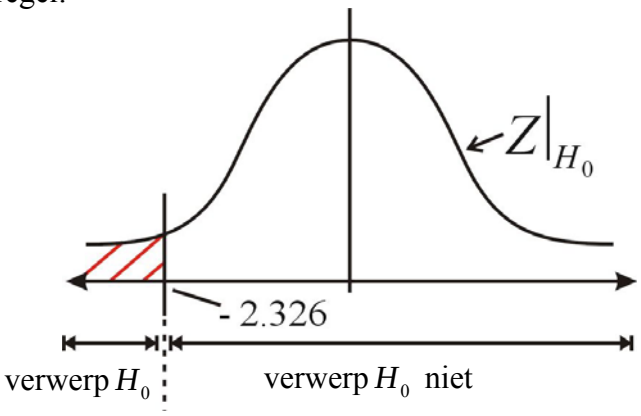
$$Z = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}\Big|_{H_0} \approx N(0,1)$$

p schatten door: $\hat{p}\big|_{obs} = \frac{24.2 \text{ x } 178 + 48.2 \text{ x } 19}{178 + 191}$

$= 0,365$

$\hat{q}\big|_{obs} = 0,635$ (: $\hat{q}\big|_{obs}$ is het compliment van $\hat{p}\big|_{obs}$, $= 1 - 0,365$)

Stap 4: Beslissingsregel:



Stap 5: Uitvoeren Toets :

$-Z_{0,01} = -2,326$

$$Z|_{obs} = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0,242 - 0,482}{\sqrt{0.365 \text{ x } 0,635 \text{ x } (\frac{1}{178}+\frac{1}{191})}} = -4,78$$

(: met $\bar{X}_1$ ja in groep BA en $\bar{X}_2$ ja in groep AB)

→ de waarde van $Z|_{obs}$ $(-4,78)$ ligt in het critisch gebied

→ dus H_0 verwerpen

* Hoe ver staat hij in het rood? (analoog als bij Voorbeeld 2):

$P(Z \le -4,78) \le 10^{-6}$

<u>Analoge Oefening als Voorbeeld 4:</u>

p = proportie tegenstanders tegen "stemrecht vreemdelingen"

$= \frac{320}{400} =$ <u>80%</u> (uit krant)

p = proportie tegenstanders tegen "stemrecht vreemdelingen" die hier lang genoeg legaal wonen

$= \frac{792}{2200} = 36\%$

→ de nuance is belangrijk

<u>Voorbeeld van het belang van foutenmarges:</u>

14 oktober 2004 : peiling kiesgedrag op Terzake:
VB : 24,3%
CD+V : 24,2%
Conclusie: VB de "grootste" partij in Vlaanderen
Maar: er was wel een foutemarge van $\pm 2,25\%$ dus conclusie is niet noodzakelijk juist.

15.3 Kwaliteiten van een test (HB p256)

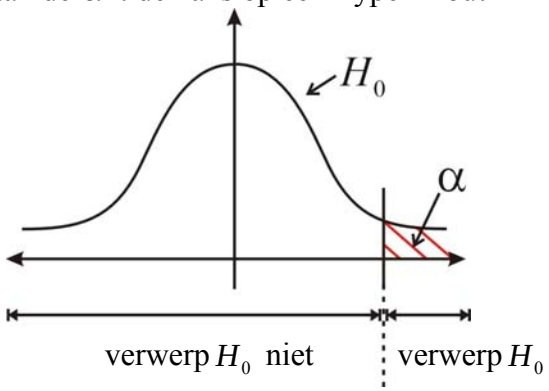
Besluit van een toets : verwerp H_0 / verwerp H_0 niet

! je kan nooit 100% zeker zijn dat het juiste besluit genomen wordt.

		werklijkheid H_0	werklijkheid H_1
besluit	H_0	OK $= 1-\alpha$	TYPE 2 FOUT $= \beta$
	H_1	TYPE 1 FOUT $= \alpha$	OK $= 1-\beta$

$P(\text{type } 1-\text{fout}) = P(H_0 \text{ verwerpen} \mid H_0 \text{ waar})$
$= \alpha$

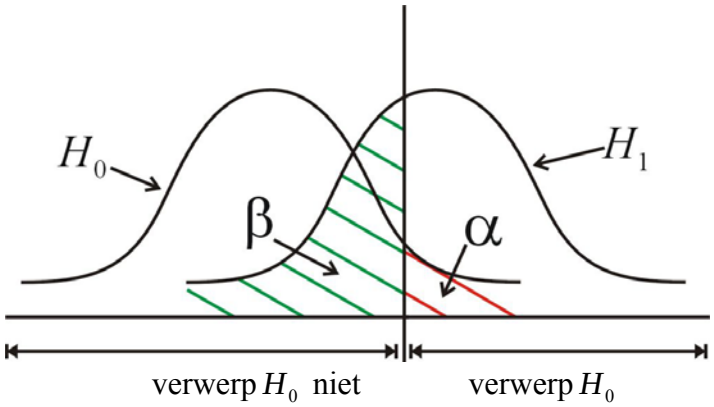
:een econoom verwerpt H_0 terwijl H_0 in de werklijkheid toch waar is. Hoe groot is die kans?
→ dat fout is precies gelijk aan de α : de kans op een Type 1-fout



! α hebben we onder controle : significantie van de toets

$P(\text{type } 2-\text{fout}) = P(H_0 \text{ aanvaarden} \mid H_0 \text{ niet waar})$
$= \beta$: de kans op een Type 2-fout

Illustratie:



Opmerking:

α verkleinen → β vergroten

α hebben we onder controle, β niet

→ Doel : H_0 verwerpen

Oefening: bereken β : de kans op een type 2-fout

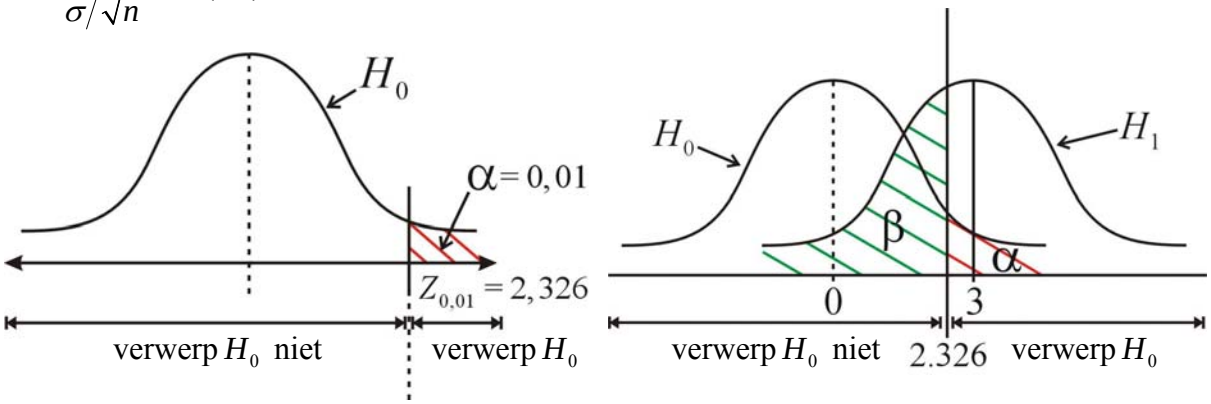
$X \sim N(\mu, \sigma^2 = 9)$, $n = 9$

$H_0 : \mu = 0$

$H_1 : \mu = 3$

$\alpha = 0,01$: 99% betrouwbaarheids interval

$$Z = \frac{\bar{X}}{\sigma/\sqrt{n}} \sim N(0,1)$$



$Z' = Z|_{H_1} \sim N(3,1)$: de toetsstochast van de alternatieve hypothese H_1

$\beta = P(\text{toetstochast} \leq 2,326 | H_1 \text{ is waar})$: kans H_0 aanvaarden gegeven dat H_0 niet waar is.

(bovenstaart kans berekenen: $P(Z \geq 0,674) = 0,25$ of 25%)

$= 0,25$: een 25% kans om een type 2-fout te maken

$= P(Z' - 3 \geq 0,674)$

Toetsen van hypothese en betrouwbaarheidintervallen (een oude examenvraag)

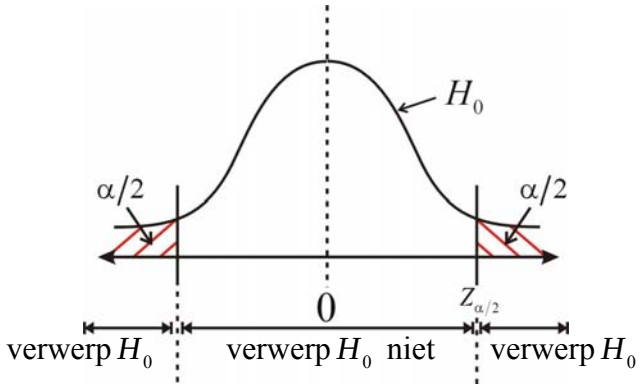
$X \sim N(\mu, \sigma^2)$ met σ^2 gekend

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

α

Toetsstochast: $Z|_{obs} = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$



Aanvaard H_0 indien:

$$-Z_{\alpha/2} \le Z\big|_{obs} \le Z_{\alpha/2}$$

$$-Z_{\alpha/2} \le \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \le Z_{\alpha/2}$$

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

[aanvaard H_0 indien μ_0 tot het $(1-\alpha)$ x 100% betrouwbaarheids interval behoort]
(hypotese toetsen = betrouwbaarheids interval)

15.5 Test op de onderliggende verdeling (HB p264)

$X : \Omega \to \mathbb{R}$ is een stochastische variabele
$\{x_1, x_2, ..., x_n\}$ is een geobserveerde steekproef
Vraag: gegeven deze data, heeft X een welbepaalde verdeling?

"goodness-of-fit" –toets, twee mogelijkheden:
* kwantieldiagram
* chi-kwadraat-toets (X^2)

a) kwantieldiagram (via voorbeeld)
→ toetsen dat de data inderdaad normaal verdeeld is

dataset van lengte 10 (n = 10) in ordening:

26	38	43	46	47	50	60	61	62	65
↑	↑					↑	↑	↑	↑
0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0,95

mediaan: $\frac{47+50}{2} = 48,5$

veronderstelling dat:
26 getrokken is uit het 0-0,10 kwantiel
38 getrokken is uit het 0,10-0,20 kwantiel
.....

Toets: H_0 : stochastische varibele $X \sim N(\mu, \sigma^2)$: een parametrische verdeling

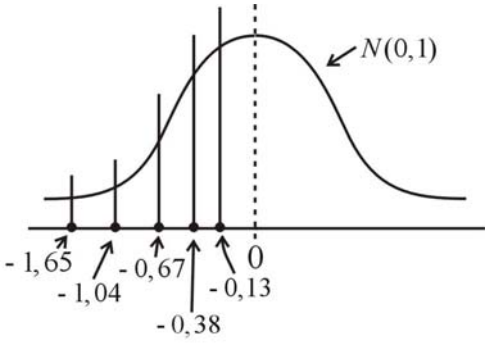
$$\mu \overset{geschat}{\leftarrow} \bar{x} = 49,8$$

$$\sigma^2 \leftarrow s^2 = (12,4)^2$$

We "berekenen" uit $N\left(49,8;(12,4)^2\right)$ het 0,05-kwantiel, 0,15-kwantiel, ...

$$P(X < x) = 0,05 \rightarrow P\left(Z \leq \frac{X - 49,8}{12,4}\right) = 0,05$$

$$Z = \frac{X - 49,8}{12,4} \sim N(0,1)$$



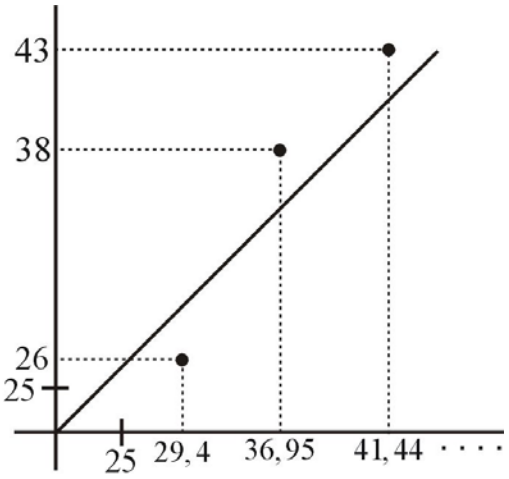
					49.80 ↓				
26	38	43	46	47	50	60	61	62	65
29.40	36.95	41.44	45.02	48.24	51.36	54.58	58.16	62.65	70.20

Hoe berekenen we de cijfers op de tweede rij?

We weten dat $Z = \frac{X - \mu}{\sigma} = \frac{X - 49,8}{12,4}$

In het 0,05 kwantiel : $Z_{0,05} = \frac{X - 49,8}{12,4}$:

- We zoeken $Z_{0,05}$ op in de tabel
- We vinden $Z_{0,05} = 1,64$
- We zoeken X; $\frac{X - 49,8}{12,4} = 1,64$
- $X = 29,40$



Bij veel data liggen de punten dicht bij elkaar.
H_0 ← als kwantiel dichtbij de diagonaal.

b) Pearson-goodness-of-fit X^2 (HB p264)

vb:
Vraag: zijn de verjaardagen van 48 studenten "uniform" verdeeld over de 4 seizoenen?

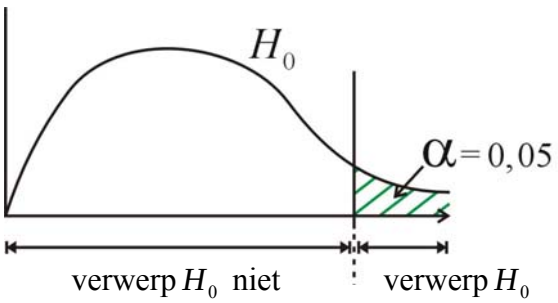
k =		O_k geobserveerde "cel" frequentie	e_k verwachte "cel" frequentie $_{H_0}$ ($H_0 : X \sim$ Uniform)
	1. lente	15	12
	2. zomer	14	12
	3. herfst	9	12
	4. winter	10	12
		+ 48	

$$X^2 = \sum_{k=1}^{4} \frac{(O_k - e_k)^2}{e_k} \sim X^2_{4-1}$$: 4 is het aantal sezoenen

↑

stelling zonder bewijs

Beslissingsregel:



→ we verwerpen alleen als de som te groot is

Uitvoeren van de toets

$X^2_{3,0.05} = 7,81$: dit is wat men zou verwachten als het uniform verdeeld zou zijn

$X^2\big|_{obs} = \frac{9+4+9+4}{12} = \frac{13}{6} = 2,17 < 7,81$: dit is kleiner dan $7,81$

→ we verwerpen H_0 niet.

Bruikbaarheid van deze X^2-toets

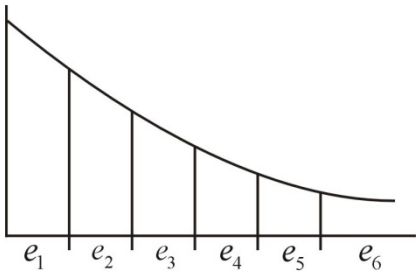
* $O_k\big|_{obs}$ en e_k zijn groter dan 5, anders moet je de cellen samenvoegen.

→ dit is toegelaten mits de verwachte cell frequentie nooit veel kleiner is dan 5.

* de e_k-tijes zijn niet te ver uit elkaar

* ook voor continue verdelingen

vb. $X \sim Exp(\lambda)$, het verloop van de exponentiele verdeling:



* de kansen zijn ongeveer even groot: $e_1 \approx e_2 \approx e_3 \approx .....$

H_0 : X heeft welbepaalde verdeling

$$X^2 = \sum_{k=1}^{\pi} \frac{(O_k - e_k)^2}{e_k} \sim X^2_{\pi-1}$$

(indien je parameters (λ) moet schatten (e_k) dan verlies je vrijheidsgraden)

15.6 Niet-parametrische tests (HB p268)

15.6.2 Wilcoxon teken-rangtest voor een mediaan (HB p271)

Probleem: toetsen van een betrouwbaarhieds interval die op de normale verdeling of CLS is gebaseerd

Vraag: Wat indien de condities niet zijn voldaan? Dus als het geen normale verdeling is? (vb. n klein, verdeling scheef....)
De "kost" is meestal verlies aan power: " $\beta \nearrow$ ": de kans op een type 2-fout wordt vergroot.

! niet parametrische toetsen
vb. de mediaan toets : locatie maat → locatie toets

Voorbeeld: de koers van zeven aandelen op twee tijdstippen t_1 en t_2 :

	$V = K_{t_2} - K_{t_1}$
1	-100
2	+200
3	-50
4	+20
5	+50
6	+200
7	+100

Vraag : is de beurs "vooruitgegaan" ?
→ is de mediaan > 0 ?

Toetsstochasten: (mediaan = $V = K_{t_2} - K_{t_1}$)

H_0 : mediaan V = 0

H_1 : mediaan V > 0 : we gebruiken een één zijdig verwerpingsgebied

* orden de absolute waardes van de V's ($|V_k|$) voor $k = 1, 2, ..., 7$ maar onthou/noteer het teken:

+	-	+	-	+	+	+
20	50	50	100	100	200	200
1	2	3	4	5	6	7
	2,5	2,5	4,5	4,5		

* rangnummers corrigeren voor knopen
→we hebben twee maal 50 dus beide krijgen rank 2,5 in de plaats van 2 en 3.

Stochast 1 : ranksom van de positieve deviaties op m_O :

$T^+ = 1 + 2,5 + 4,5 + 6 + 7 = 21$

Stochast 2 : ranksom van de negatieve deviaties op m_O :

$T^- = 2,5 + 4,5 = 7$

■ $T^+ + T^- = 1 + 2 + \ldots + k = 28 = \dfrac{k(k+1)}{2}$

■ Verwachte waarde = $E(T^+ | H_0) = \dfrac{k(k+1)}{4} = \mu$

■ $\text{var}(T^+ | H_0) = \dfrac{k(k+1)(2k+1)}{24} = \sigma^2$

Stelling zonder bewijs:

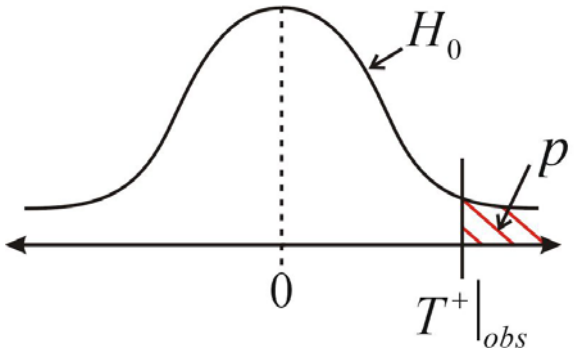
als n groot is, dan: $T^+|_{H_1} = N\left(\dfrac{k(k+1)}{4}, \dfrac{k(k+1)(2k+1)}{24}\right)$ (: CLS $n > 20$)

p-waarde $= P\left(T^+ > T^+|_{obs}\right)$

($T^+|_{obs}$ is een discrete stochast)

$$= P\left(Z > \frac{T^+|_{obs} - 0,5 - k(k+1)/4}{\sqrt{k(k+1)(2k+1)/24}}\right)$$

* als n klein is, dan tabellen voor $T^+|_{obs}$



DEEL IV
RELATIE-ONDERZOEK

16. Tests voor onafhankelijkheid in een kruistabel (HB p279)

$$(X,Y):\Omega \to \mathbb{R}^2$$
$$\omega \to \big(X(\omega),Y(\omega)\big)$$

In dit deel gaan we toetsen of :
* X en Y gecorreleerd zijn
* X en Y onafhankelijk zijn
* en verband tussen X en Y bestudeeren (voorbereiding econometrie)

Inferentie voor correlatie (wederzijdse afhankelijkheid)

Dataset $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$

Pearson correlatiecoefficient:
Voor een steekproef :

$$\pi = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\left[\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2\right]^{1/2}}$$

$-1 \le \pi \le 1$

Voor een populatie:

$$\rho_{X,Y} = \frac{E(XY)-E(X)E(Y)}{\sigma_X \sigma_Y} = \frac{\text{cov}(x,y)}{\sigma_X \sigma_Y}$$

Vraag: $\rho_{X,Y} \overset{schatter}{\leftarrow} \pi_{X,Y} = R_{X,Y}\big|_{obs}$?
: is de correlatiecoefficient van een steekproef een goede schatter voor die van de populatie?

Indien X en Y bivariaat normaal verdeeld zijn:

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

$$f(x,y) = \frac{1}{2\pi\sqrt{\det\Sigma}} e^{-\frac{1}{2}z'\Sigma^{-1}z} \text{ met } z = \begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}$$

: de twee dimentionele dichtheidsfunctie
: dit is een uitbreiding van de dichtheidsfunctie in één dimentie

$$X \sim N(\mu,\sigma^2) \ , \ p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

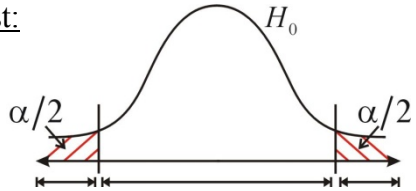
Indien: $\rho_{X,Y} = 0$: niet gecorrelleerd : dus onder H_0

Dan: $\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{n-2}$: niet in formularium, stelling zonder bewijs

Correlatietest:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

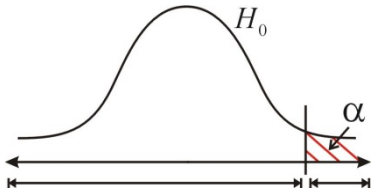


→ tweezijdig

verwerp H_0 verwerp H_0 niet verwerp H_0

$H_0 : \rho = 0$

$H_1 : \rho > 0$



→ éénzijdig

verwerp H_0 niet verwerp H_0

Contingentie Analyse:

Data set:

Inkomen / Regio	0-30	30-60	60-90	90+	som
Noord	58	91	35	14	198
Zuid	47	38	13	4	102
som	105	129	48	18	300

Vraag: zijn "regio" en "inkomen" onafhankelijk? : is er correlatie tussen de twee?

H_0 : regio en inkomen zijn onafhankelijk

H_1 : niet H_0 : regio en inkomen zijn afhankelijk

Hoe zou de tabel eruit zien indien H_0 waar is? :

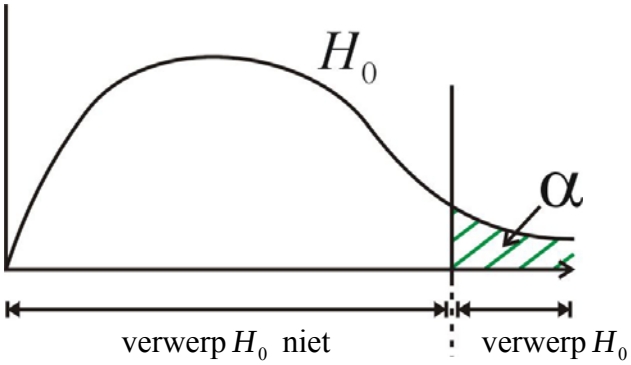
Inkomen / Regio	0-30	30-60	60-90	90+	som
Noord	$\frac{198}{300}$ x $105 = 69,3$	$\frac{198}{300}$ x $129 = 85,2$	$\frac{198}{300}$ x $48 = 31,7$	11,8	198
Zuid	105-69,3=35,7	129-85,2=43,8	48-31,7=16,3	6,2	102
som	105	129	48	18	300

Vraag: in hoever zijn Tabel $_{H_0}$ (de tabel indien H_0 waar is) en Tabel $_{obs}$ (de geobserveerde tabel) "statistisch" verschillend?

* Pearson X^2 goodness of fit:

$X^2 = \sum_{i,j} \frac{(O_{ij} - e_{ij})^2}{e_{ij}} \sim X^2_{(4-1)(2-1)=3}$: in het formularium p6

Beslissingsregel:



Toets uitvoeren:

$X^2_{3;0.025} = 9,35$: 3 vrijhiedsgraden, 25% betrouwbaarheids interval

$X^2\big|_{obs} = 9,2$: rekenwerk, analoog aan de oefening van de verjaardagen

→ H_0 niet verwerpen $X^2_{3;0.025} > X^2\big|_{obs}$

17. Regressie (HB p282) : Schatten van relaties

17.1 Het regressie probleem (HB p282)

geobserveerde data : bivariaat:

$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

vb. $(x_1, y_1) = (\text{inkomen, spaarvolume})$ van individu i

Vraag: is er een "model" achter deze data?

: de data verklaren door een model (een verband, het regressiemodel) dat goed past op de data

$y = f(x)$

$y = a + bx$: lineair model

$y = ae^{bx}$: exponentieel model

$y = a + b\ln(1+x)$: logaritmisch model

....

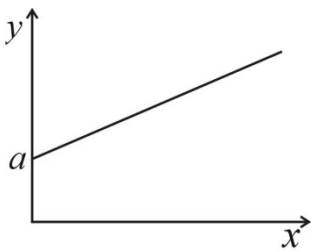
1. fit: goodness-of-fit

→ geef de lineaire regressie van de spaarsom als functie van het inkomen en geef maten voor de kwaliteit van de fit.

2. voorspellingen maken op basis van dit model

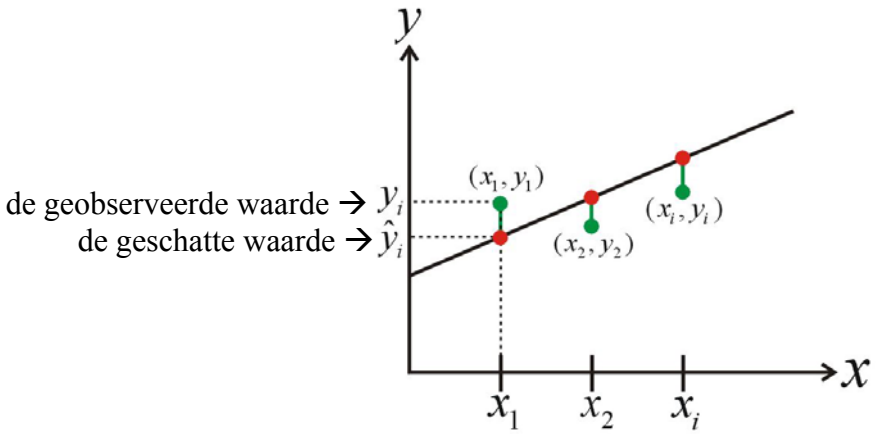
vb. een lineair eerste graads model:

$y = a + bx$



Er zijn twee types van regressie die we gebruiken:
1. Deterministische regressie: oplossen via de kleinste kwadraten methode
: de data zijn precies
2. Stochastische regressie: men beschouwt de data Y, als stochastische grootheden onderling onderheving aan toevallige (stochastische) fluctuaties.
y_i zijn observaties van de stochastische variabele Y_i:
$Y = a + bx + \mu$: hier is de μ een stochast.

17.2 Deterministische lineaire regressie - de kleinste kwadratenrechte (HB p285)



■ x_i
- y_i : de geobserveerde waarde
- $\hat{y}_i = \hat{a} + \hat{b}x_i$: de geschatte waarde via de regressie rechte : $y = \hat{a} + \hat{b}x$

■ residu: $u_i = y_i - \hat{y}_i$
: de groene lijn : het verschil tussen de geobserveerde waarde en de geschatte waarde.

■ residuele som : $V_{res} = \sum_{i=1}^{n} u_1^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$: in het formularium p6

■ Kleinste kwadraten (KK) methode:
De geschatte regressie rechte is de kleinste kwadraten rechte, dat is onder alle mogelijke rechten deze waarvan de som van de kwadratische fouten $\sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2$ minimaal wordt.

Minimaliseer de residuele som:

$\underset{\hat{a},\hat{b}}{\text{minimaliseer}}\ V_{res} = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2$

voorwaarden van 1ste orde:

$D_1 V_{res}(\hat{a},\hat{b}) = -2\sum (y_i - \hat{a} - \hat{b}x_i) = 0$: partiele afgeliede naar $\hat{a}$ (1)

$D_2 V_{res}(\hat{a},\hat{b}) = -2\sum (y_i - \hat{a} - \hat{b}x_i)x_i = 0$: partiele afgeliede naar $\hat{b}$ (2)

(1') $\sum y_i - n\hat{a} - \hat{b}\sum x_i = 0$

$n\overline{y} - n\hat{a} - n\hat{b}\overline{x} = 0$

1ste vergelijking: $\underline{\hat{a} = \overline{y} - \hat{b}\overline{x}}$: punt $(\overline{x}, \overline{y})$ ligt op de regressie rechte

(2') $\sum x_i y_i - n\hat{a}\overline{x} - \hat{b}\sum x_i^2 = 0$

$\sum x_i y_i - n(\overline{y} - \hat{b}\overline{x})\overline{x} - \hat{b}\sum x_i^2 = 0$

2e vergelijking: $\hat{b} = \dfrac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x}^2}$: makkelijker voor berekening

$\underline{\hat{b} = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}}$: hetzelfde maar makkelijker voor interpretatie

→ sterke positieve correlatie en grote y-spreiding ten opzichte van de x-spreiding moet leiden tot een steile helling naar boven $(\hat{b} > 0)$

* de voorwaarden 2de orde (Hessiaan veriferen): OK!

■ Conclusie:

$\hat{b} = \dfrac{V_{xy}}{V_{xx}} = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$: in het formularium p6

$\hat{a} = \overline{y} - \hat{b}\overline{x}$

Merk op : $y = \hat{a} - \hat{b}\overline{x}$

* $(\overline{x}, \overline{y})$ behoort tot de regressierechte.

* $\hat{b} = \dfrac{\text{covariantie } (x, y)}{\text{variantie } x} \quad \left[= \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \right]$

als spreiding (variantie) in x-waarden groot → $\hat{b}$ klein
(logisch, staat in de noemer)

vb.

inkomen x	sparen y
8000	600
11000	1200
9000	1000
6000	300
6000	700
6000	500

$\hat{b} = 0{,}144$

$\hat{a} = -395{,}5$

$$\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} = a\begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix} + b\begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}$$: alleen a en b zijn onbekend, de rest zijn observaties

$$\begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix}\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix}\begin{pmatrix} 1 & 1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}$$

vermenigvuldigen:

$$\begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}$$

! twee normaalvergelijkingen

merk op : $\det = n\sum x_i^2 - \left(\sum x_i\right)^2$

$= n\sum \left(x_i - \overline{x}\right)^2$

$= V_{xx}.n$

> 0

$$\vec{y} = X\begin{pmatrix} a \\ b \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 1 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$$

$$X^T \vec{y} = X^T X\begin{pmatrix} a \\ b \end{pmatrix}$$

$$\begin{pmatrix} a \\ b \end{pmatrix} = \left(X^T X\right)^{-1} X^T \vec{y}$$: de KK methode

---

<u>Maten voor de kwaliteit van de fit van de regressie rechte</u> (HB p287)

<u>Verklarings- of determiniatie coefficient</u>

$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$

totale afwijking = deel verklaard door regressie + deel niet verklaard (residu)

$$\sum (y_i - \overline{y})^2 = \sum \left((\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)\right)^2$$

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

$$\boxed{V_{tot} = V_{reg} + V_{res}}$$

totale variatie = verklaarde variatie + niet verklaarde variatie
(Let op: variatie is niet variantie)

$$\sum(\hat{y}_i - \overline{y}).(y_i - \hat{y}_i) = 0 \quad ?$$

$$L.L. = \sum(\hat{a} + \hat{b}x_i - \hat{a} - \hat{b}\overline{x}).(y_i - \hat{a} - \hat{b}x_i)$$

$$= \hat{b}\sum(x_i - \overline{x}).(y_i - \hat{a} - \hat{b}x_i) \qquad = 0 \ (2^{\text{de}})$$

$$= \hat{b}\sum x_i.(y_i - \hat{a} - \hat{b}x_i) - \hat{b}\overline{x}\sum(y_i - \hat{a} - \hat{b}x_i) \qquad = 0 \ (1^{\text{ste}})$$

$$0 \le \frac{V_{reg}}{V_{tot}} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = r^2 \le 1$$

r^2 : de determinatiecoefficient

$r^2 = 1$ als $\hat{y}_i = y_i \quad \forall_i$: y_i is een punt op de regressie rechte

Let op : determinatiecoefficent r^2 $\neq$ correlatiecoefficient r

r^2 is de proportie van de totale variatie die verklaard wordt door het model.

Opmerking: een lineaire regressie is slechts zinvol van correlatie coefficient $r \ge 0,70$.

In dat geval is de $r^2 \ge \frac{1}{2} = 50\%$, dit wil zeggen dat pas dan de lineaire regressie minstens de helft van de variabiliteit op y zal verklaren.

De residuele som is geschikt voor vergelijkingen van modellen

→ een model $y = f(x)$ is te verkiezen boven een model $y = g(x)$ indien in het eerste de residuele som lager is.

Vraag: gebruiken we best $x = g(y)$ of $y = f(x)$?

Antwoord: Determinatiecoefficient : $r^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}$

$$= \frac{\sum(\hat{a} + \hat{b}x_i - \hat{a} - \hat{b}\overline{x})^2}{\sum(y_i - \overline{y})^2}$$

$$= \hat{b}^2 \frac{\sum(x_i - \overline{x})^2}{\sum(y_i - \overline{y})^2}$$

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

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

→ r^2 is symmetrisch in $\vec{x}$ en $\vec{y}$

→ de keuze tussen $y = f(x)$ en $x = g(y)$ heeft geen effect op de determinatiecoefficient

[indien $r^2 \ge \frac{1}{2}$: de regressie verklaart de helft (of meer) van de variabiliteit in $\vec{y}$]

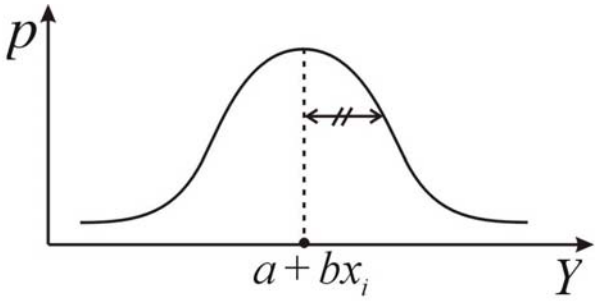
17.3 Stochastisch model (HB p289)

Model: $Y_i = a + bx_i + U_i$: slecht model; moeten voowaarden op de storingswaarden plaatsen de stochastische variabele; U_i is de storingsterm (moet bij voorkeur klein zijn)

- $E(U_i) = E(U_i U_j) = 0$: gemiddelde is gelijk aan nul
- $E(U_i^2) = \sigma^2$ (voor alle i)
- $E(U_i X_i) = E(U_i x_i) = 0$

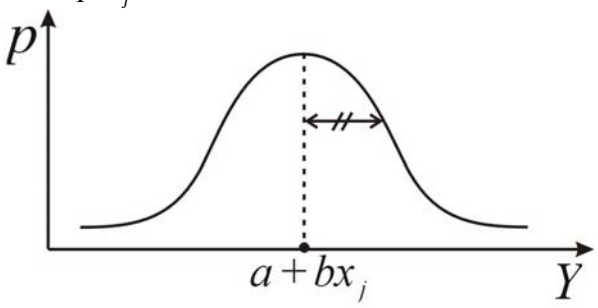
verdeling van Y conditioneel op x_i:

- $E(Y|x_i) = a + bx_i$
- $\sigma^2_{Y/x_i} = \sigma^2$

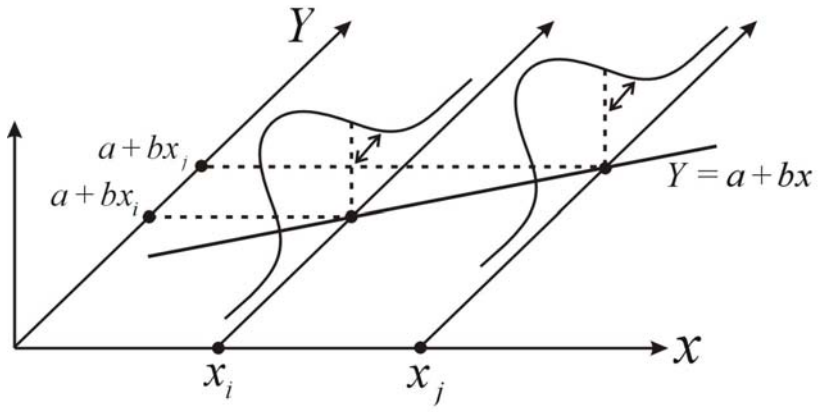


verdeling van Y conditioneel op x_j

- $E(Y|x_j) = a + bx_j$
- $\sigma^2_{Y/x_j} = \sigma^2$



Model: $Y_i = a + bx_i + U_i$ grafisch:



KK: ■ $\hat{b} = \dfrac{\sum (x_i - \overline{x})(Y_i - \overline{Y})}{\sum (x_i - \overline{x})^2}$

- $\hat{a} = \overline{Y} - \hat{b}\overline{x}$

  : Y_i en $\overline{Y}$ zijn stochasten

  : het verschil met de deterministische methode zijn de hoofdletters.

* de U_i zijn st.v. → Y_i st.v. → $\hat{a}, \hat{b}$ st.v.

* $\hat{a}, \hat{b}$ schatters voor a, b

Vraag: welke eigenschappen (vertekening, consistentie,.....)

$E(\hat{a}),\ E(\hat{b})$

$\text{var}(\hat{a}),\ \text{var}(\hat{b})$

stochastiek in $\hat{a}$, $\hat{b}$ "blootleggen"; $\hat{a}$, $\hat{b}$ schrijven in termen van U_i:

$$\hat{b} = \frac{\sum (x_i - \overline{x})(Y_i - \overline{Y})}{\sum (x_i - \overline{x})^2}$$
$$= \sum (x_i - \overline{x}) Y_i - \sum (x_i - \overline{x}) \overline{Y}$$
$$= \sum (x_i - \overline{x})(a + b x_i + U_i)$$
$$= b \sum (x_i - \overline{x})(x_i - \overline{x}) + \sum (x_i - \overline{x}) U_i \quad \text{: de term die wordt toegevoegd wordt = 0}$$

$$\boxed{\hat{b} = b + \frac{\sum (x_i - \overline{x}) U_i}{\sum (x_i - \overline{x})^2}} \quad \text{: in het formularium p6}$$

* indien $E(U_i) = 0$, dan $E(\hat{b}) = b$ onvertekend.

$$E(\hat{b} - b)^2 = E\left(\frac{\sum (x_i - \overline{x}) U_i}{\sum (x_i - \overline{x})^2} \right)^2$$
$$= \frac{1}{\left(\sum (x_i - \overline{x})^2 \right)^2} . E\left(\sum_i (x_i - \overline{x}) U_i \right)^2$$
$$= \frac{1}{\left(\sum (x_i - \overline{x})^2 \right)^2} . E\left(\sum_{ij} (x_i - \overline{x})(x_j - \overline{x}) U_i U_j \right)$$
$$= \frac{1}{\left(\sum (x_i - \overline{x})^2 \right)^2} . E\left(\sum_i (x_i - \overline{x})^2 U_i^2 \right)$$
$$= \frac{1}{\left(\sum (x_i - \overline{x})^2 \right)^2} . \left(\sum_i (x_i - \overline{x})^2 E(U_i^2) \right)$$

$$\boxed{\text{var}\,\hat{b} = \frac{1}{\sum (x_i - \overline{x})^2} . \sigma^2} \quad \text{: analoog voor } \hat{a} \text{ : in het formularium p6}$$

$$\left(Z_1 + Z_2 + ... + Z_n \right)^2$$
$$= \left(Z_1 + Z_2 + ... + Z_n \right)\left(Z_1 + Z_2 + ... + Z_n \right)$$
$$= Z_1 Z_1 + Z_1 Z_2 + ... + Z_1 Z_n$$
$$Z_2 Z_1 + Z_2 Z_2 + ... + Z_2 Z_n$$
$$.....$$
$$Z_n Z_1 + Z_n Z_2 + ... + Z_n Z_n$$
$$= \sum_{ij} Z_i Z_j$$

* grote spreiding in X_i → kleine variantie $\hat{b}$

Vraag: heeft $\hat{b}$ de kleinste variantie onder de lineaire schatters zonder vertekening?

$$\hat{\hat{b}} = \sum g_i Y_i = \sum g_i (a + bx_i + U_i) = \sum g_i (a + bx_i) + \sum g_i U_i$$

■ $E(\hat{\hat{b}}) = \sum g_i (a + bx_i) = b$: onvertekend

: dat kan enkel maar als $\sum g_i x_i = 1$ en $\sum g_i = 0$.

■ $\text{var}(\hat{\hat{b}}) = \sum E(g_i^2 U_i^2)$

$= \left(\sum g_i^2\right)\sigma^2$

*covarianties = 0
* $E(U_i^2) = \sigma^2$

* Minimaliseer $\sum_i g_i^2$, twee randvoorwaarden:

Lagrange : $L\left(\vec{g}, \lambda, \mu\right) = \sum_i g_i^2 - \lambda(\sum g_i x_i - 1) - \mu \sum g_i$

$$\frac{\partial L}{\partial g_i} = 2g_i - \lambda x_i - \mu = 0$$

$$2g_i = \lambda x_i + \mu \quad \overset{\sum_i}{\rightarrow} \quad 0 = \lambda \sum x_i + n\mu$$

$$\mu^* = -\lambda^* \bar{x}$$

vermenigvuldigen met x_i :

$$2g_i x_i = \lambda^* x_i^2 - \lambda^* \bar{x} x_i$$

sommeer : $\lambda^* (\sum x_i^2 - \bar{x} \sum x_i) = 2$

$$\lambda^* (\sum (x_i - \bar{x})^2) = 2$$

$$\lambda^* = \frac{2}{\sum (x_i - \bar{x})^2}$$

$$\mu^* = \frac{2\bar{x}}{\sum (x_i - \bar{x})^2}$$

$$2g_i = \frac{2(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

$$g_i^* = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$

Bijgevolg : $\hat{\hat{b}}^* = \hat{b}$ is BLUE : Best Linear Unbiased Estimator

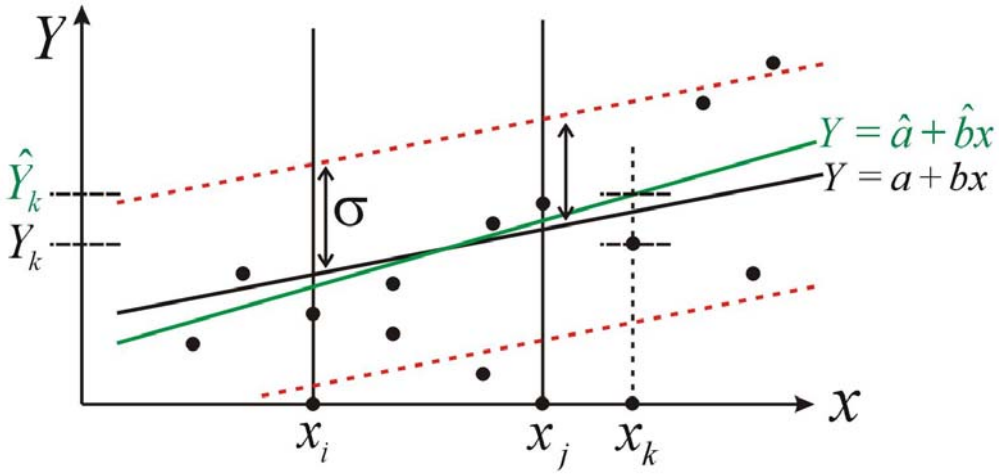
Gauss-Markov theorema:

KK-schatters $\hat{a}$ en $\hat{b}$ zijn BLUE (bewijsje enkel voor $\hat{b}$)
!! geen 'specifieke' veronderstellingen omtrent de 'verdeling' van U_i

Model: $Y_i = a + bx_i + U_i$ $\quad a \leftarrow \hat{a}$

$\sigma^2 = E(U_i^2)$ $\quad b \leftarrow \hat{b}$

σ is het standaardfout (spreiding van de storingsterm U) van de regressie



$Y = a + bx$: "ware" model

$Y = \hat{a} + \hat{b}x$: "geschatte" model op basis van de punten

<u>Vraag:</u> is dit een goede schatter voor σ^2 ?

$\sigma^2 \leftarrow \dfrac{\sum(\hat{Y}_i - Y)^2}{n-2}$: in het formularium p1

: verlies van 2 vrijheidsgraden (2 normaal vergelijkingen.)

: als men n-1 zou gebruiken dan is er een overschatting

<u>Voorspellingen:</u>

$x_0 \notin \{x_1, x_2, ..., x_n\}$

Y_0 ware waarde : $a + bx_0 + U_0$

voorspelling via regressie rechte: $\hat{Y}_0 = \hat{a} + \hat{b}x_0$

$E(\hat{Y}_0) = E(\hat{a} + \hat{b}x_0) = a + bx_0 = E(Y_0)$

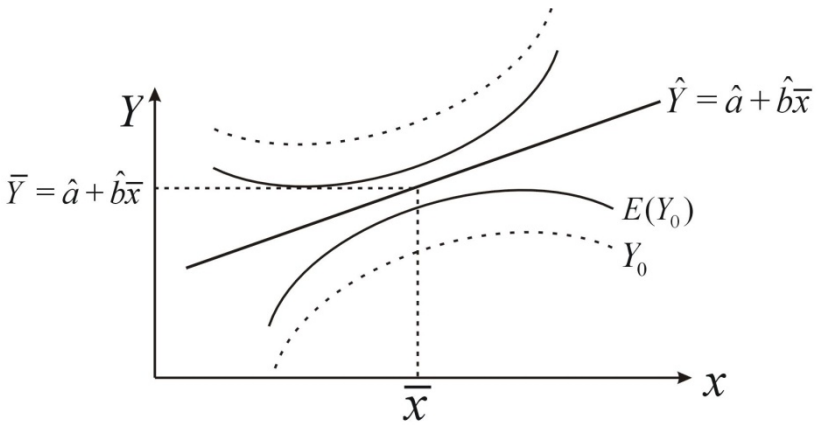
de schatter $\hat{Y}_0$ is: onvertekend voor $E(Y_0)$

vertekend voor Y

$$MSE = (\hat{Y}_0 - E(Y_0))^2 = E(\hat{a} + \hat{b}x_0 - a - bx_0)^2$$
$$= E(\hat{a} - a + x_0(\hat{b} - b))^2$$
$$\vdots$$
$$= \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2} \right\}$$
$$= f\left[\sigma^2, n, \sum(x_i - \bar{x})^2, (x_0 - \bar{x})^2\right]$$
$$+ \quad - \qquad - \qquad +$$

$$E(\hat{Y}_0 - Y_0)^2 = E(\hat{Y}_0 - E(Y_0) - U_0)^2$$
$$= E(\hat{Y}_0 - E(Y_0))^2 + E(U_0^2)$$
$$= \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\} + \sigma^2$$
$$= \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right\}$$

1 + : omdat het voor specifieke waarde is voor X_0 met als voor gemiddelde = 0



<u>Betrouwbaarheids interval en hypothesetoetsen voor a en b</u>

Model : $Y_i = a + bx_i + U_i$ met de 3 standaard condities

$U_i \sim N(0, \sigma^2)$: onafhankelijk identisch verdeeld

$$\hat{b} = b + \frac{\sum (x_i - \bar{x}) U_i}{\sum (x_i - \bar{x})^2}$$

$$\hat{b} \sim N(b, \sigma_{\hat{b}}^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

σ^2 kennen we niet dus wordt geschat door $\frac{\sum \left(\hat{Y}_i - Y_i \right)^2}{n-2}$

- $\frac{\hat{b} - b}{\sqrt{\sigma^2 / \sum (x_i - \bar{x})^2}} \sim N(0,1)$

- $\frac{\hat{b} - b}{\sqrt{\frac{\sum (\hat{Y}_i - Y_i)^2}{(n-2)\sum (x_i - \bar{x})^2}}} \sim t_{n-2}$

vb. inkomen-spaaroefening:

$Y = -373.4 + 0.142x$
$R^2 = 0.786$

betrouwbaarheids interval voor b : $0.142 \pm t_{4,\alpha/2}.0.037 > 0$

$[\alpha = 0.02 : 3.747]$

: 98% betrouwbaarheid en 3.747 is de t waarde

Extra : Ranksom toets

■ $T^t = \sum_1^n kX_k$

met X_k : 1 indien k-de verschil >0
0 indien k-de verschil <0

■ $X_k \sim b(1, \frac{1}{2})$

$E(X_k) = \frac{1}{2}$, $\text{var}(X_k) = \frac{1}{4}$

■ $E(T^t) = E(\sum kX_k)$
$= \sum kE(X_k)$
$= \frac{1}{2}\sum k = \boxed{\frac{1}{2}.\frac{n(n+1)}{2}}$

■ $\text{var}(T^t) = \text{var}(\sum kX_k)$
$= \sum K^2 \text{ var } X_k$
$= \frac{1}{4}\sum K^2 = \boxed{\frac{1}{4}.\frac{n(n+1)(2n+1)}{6}}$

OEFENZITTINGEN

Oefenzitting 1 – Beschrijvende Statistiek

1. De volgende dataset heeft betrekking tot de webstek 'Werkgroep Econometrie (CES)' en geeft het aantal 'bezoekers per dag'. De data werden genoteerd in koppels: de eerste coordinaat geeft de dag weer, de tweede het aantal bezoekers. Het eerste (laatste) koppel heeft betrekking tot 11 september (7 oktober) 2001.

(11.09, 32) (12.09, 8) (13.09, 14) (14.09, 11) (15.09, 4)
(16.09, 3) (17.09, 5) (18.09, 11) (19.09, 8) (20.09, 11)
(21.09, 11) (22.09, 1) (23.09, 1) (24.09, 35) (25.09, 26)
(26.09, 18) (27.09, 23) (28.09, 27) (29.09, 5) (30.09, 10)
(01.10, 46) (02.10, 31) (03.10, 46) (04.10, 39) (05.10, 29)
(06.10, 10) (07.10, 9).

1.A. De volgende vragen hebben enkel betrekking tot de tweede coordinaat van deze data.
• Geef de range van deze data,
• Bereken het gemiddelde, de mediaan, en de modus,
• Bereken de variantie. Verifieer het verband tussen de variantie, de som der kwadraten, en het gemiddelde,
• Verifieer de stelling van Tchebychev met k = 1, 2, 3,
• Groepeer de data (in 6 tot 8 klassen, je mag hier zelf een keuze maken), teken de corresponderende histogram en cumulatieve polygoon. Geef uiteraard ook de voorbereidende tabel.

De georderde dataset (van klein naar groot) :
1, 1, 3, 4, 5, 5, 8, 8, 9, 10, 10, 11, 11, 11, 11, 14, 18, 23, 26, 27, 29, 31, 32, 35, 39, 46, 46

■ Laagste observatie = 1 , Hoogste observatie = 46
Range = 46 – 1 = 45

■ Rekenkundig gemiddelde = $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{474}{27} = 17,56$

■ Mediaan : de 14de observatie: $Me = x_{\langle 14 \rangle} = 11$

■ Modus : de waarde die het meest voorkomt : $Mo = 11$

■ Variantie : $\tilde{s}^2 = \frac{\sum (x_i - \overline{x})}{n} = \frac{4906,67}{27} = 181,73$
(Excel berekent de 'steekproef' variantie (met n-1 weging = 26-weging)

■ Verband variantie, som der kwadraten, gemiddelde :
$n\tilde{s}^2 = \sum (x_i - \overline{x})^2$
$= \sum (x_i^2 - 2\overline{x}x_i + \overline{x}^2) = (\sum x_i^2) - 2n\overline{x}^2 + n\overline{x}^2 = (\sum x_i^2) - n\overline{x}^2$
Om te verifieeren kiezen we een waarde voor n. Als n = 27, dan wordt:
$n\tilde{s}^2 = (27)(181,73) = 4906,71$ en $\sum x_i^2 = 13228$ en $n\overline{x}^2 = 8325,54$
De stelling : $n\tilde{s}^2 = (\sum x_i^2) - n\overline{x}^2$ → $4906,71 = ! = 13228 - 8325,54$

■ Verifieer de stelling van Tchebychev met k = 1, 2, 3;

Stelling: het interval $\left[\overline{x}-k\tilde{s},\overline{x}+k\tilde{s}\right]$ bevat ten minste $(1-\frac{1}{k^2})$ x 100% van de data

<u>k = 1</u>: het interval $\left[(17,56)-(1)\left(\sqrt{181,73}\right),(17,56)+(1)\left(\sqrt{181,73}\right)\right]$

$=\left[(17,56)-(13,48),(17,56)+(13,48)\right]$

$=\left[4,08\ ,\ 31,04\right]$

Tchebychev: dit bevat ten minste $(1-\frac{1}{1^2})$ x 100% van de data = 0%

In werkelijkheid: 18 van de 27 waardes liggen in dit interval = $\frac{18}{27}=66,67\%$

<u>k = 2</u>: het interval $\left[(17,56)-(2)\left(\sqrt{181,73}\right),(17,56)+(2)\left(\sqrt{181,73}\right)\right]$

$=\left[(17,56)-(26,96),(17,56)+(26,96)\right]$

$=\left[-9,4\ ,\ 44,52\right]$

Tchebychev: dit bevat ten minste $(1-\frac{1}{2^2})$ x 100% van de data = 75%

In werkelijkheid: 25 van de 27 waardes liggen in dit interval = $\frac{25}{27}=92,59\%$

<u>k = 3</u>: het interval $\left[(17,56)-(3)\left(\sqrt{181,73}\right),(17,56)+(3)\left(\sqrt{181,73}\right)\right]$

$=\left[(17,56)-(40,44),(17,56)+(40,44)\right]$

$=\left[-22,88\ ,\ 58,00\right]$

Tchebychev: dit bevat ten minste $(1-\frac{1}{3^2})$ x 100% van de data = 88,89%

In werkelijkheid: 27 van de 27 waardes liggen in dit interval = $\frac{27}{27}=100\%$

Tabel:

interval	Tchebychev	in werkelijkheid
$\overline{x}\pm 1\tilde{s}$	0%	18/27 = 66,67%
$\overline{x}\pm 2\tilde{s}$	75%	25/27 = 92,59%
$\overline{x}\pm 3\tilde{s}$	88,89%	27/27 = 100%

De dataset bevat meer elementen tussen $\overline{x}-k\tilde{s}$ en $\overline{x}+k\tilde{s}$ dan de ondergrens via Tchebychev (de waarden die in de laatste kolom liggen zijn, zoals het hoort, boven de waarden in de tweede kolom).

■ Voorbereidende tabel met 7 klassen:

klasse	grenzen	midden	frequentie	cumulatieve frequentie	relatieve cumulatieve frequentie
1	1-7	4	6	6	6/27
2	8-14	11	10	16	16/27
3	15-21	18	1	17	17/27
4	22-28	25	3	20	20/27
5	29-35	32	4	24	24/27
6	36-42	39	1	25	25/27
7	43-49	46	2	27	27/27

Teken de corresponderende histogram en cumulatieve polygoon.

1.B. De volgende vragen vertrekken van de histogram uit de vorige opgave.
• Bereken aan de hand van de gegroepeerde gegevens het gemiddelde (vergelijk met het resultaat uit 1.A),
• Bereken aan de hand van de cumulatieve polygoon de mediaan (vergelijk met het resultaat uit 1.A),
• Bereken aan de hand van de gegroepeerde gegevens de modus (moeilijk !!),
• Geef een werkwijze om aan de hand van gegroepeerde gegevens de variantie te berekenen.

■ Ga ervan uit dat elk element binnen een klasse evenveel voorkomt, zodat per klasse het gemiddelde overeenkomt met het klassemidden. Vermenigvuldig voor elke klasse het klassemidden met haar relatieve frequentie. Het gemiddelde van het geheel wordt 'geschat' via de som van deze termen:

$$\overline{x}_{\text{gegroepeerde data}} = \frac{(4 \text{ x } 6) + (11 \text{ x } 10) + (18 \text{ x } 1) + (25 \text{ x } 3) + (32 \text{ x } 4) + (39 \text{ x } 1) + (46 \text{ x } 2)}{27} = \frac{486}{27} = 18$$

Omdat we ons baseren op de gegroepeerde data gaat een deel van de informatie verloren en wijkt deze 'schatting' af van het 'ware gemiddelde' (17, 56). De fout is echter beperkt.

■ De mediaan berekenen we via lineaire interpollatie.
De mediaan is de $\frac{27+1}{2}$de observatie = de 14^{de} observatie. We gebruiken de cumulatieve polygoon en merken dat de 14^{de} observatie tussen de cumulatieve frequenties 6 en 16 ligt.
De mediaan ligt dus in deze classen; tussen de grenzen 7 en 14.
Er zijn 10 observaties tussen 6 en 16 en de 14^{de} observatie ligt 8 observaties weg van 6.
Omdat we lineairiteit veronderstellen kunnen we nu de mediaan vinden:

$$Me_{\text{gegroepeerde data}} = 7 + \left(\frac{8}{10} \text{ x } 7 \right) = 12,6$$

■ Modus aan de hand van gegroepeerde data. De modus geeft die observatie met de hoogste frequentie. Indien de klassebreedten dezelfde zijn, dan wordt de modus in die klasse gelegd met de hoogste frequentie. Het klassemidden kan dan de rol van modus vervullen.

Deze werkwijze houdt echter geen rekening met de frequenties van de aanliggende klassen. Een voorbeeld ter illustratie (continue gegroepeerde data):

Dataset 1:
klasse 1 (10-20) freq = 4, klasse 2 (20-30) freq = 20, klasse 3 (30-40) freq = 4.

Dataset 2:
klasse 1 (10-20) freq = 4, klasse 2 (20-30) freq = 20, klasse 3 (30-40) freq = 17.

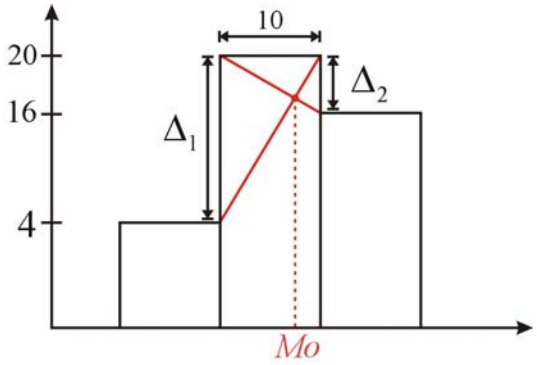
Voor beide datasets is klasse 2 (met klassemidden 25) de modale klasse.

Nochtans zijn de twee datasets essentieel verschillend. In tegenstelling tot de eerste dataset, suggereert de tweede dataset dat hogere waarden frequenter voorkomen (zie klasse 3). De volgende bepaling van de modus houdt rekening niet alleen met de modale klasse (de klasse met de hoogste frequentie) maar ook met de aanliggende klassen.

Voor dataset 2: De modale klasse steekt met $\Delta_1 = 16$ observaties uit boven haar links aanliggende klasse, en met $\Delta_2 = 3$ observaties boven de rechts aanliggende klasse.

De lengte van de klassebreedte is overal gelijk aan 10. Vandaar:

$$Mo = 20 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \text{ x } 10 = 28,4$$



■ Variantie : $n\tilde{s}^2 = \left(\sum x_i^2\right) - n\overline{x}^2$

Werkwijze voor variantie bij gegroepeerde data :

$n\tilde{s}^2 = \left(\sum m_i^2 F_i\right) - n\overline{x}^2$

waarbij m_i het midden van de i-de klasse is, F_i de absolute frequentie, en waarbij $\overline{x}$ 'geschat' wordt zoals hoger aangegeven.

2. Een fietser legt een licht stijgend traject af aan een constante snelheid van 20 km/uur. Daarna keert hij via dezelfde weg terug, hij doet dit aan de constante snelheid van 30 km/uur.
Toon aan de gemiddelde snelheid van deze fietser over het gehele traject precies gelijk is aan 24 km/uur.

Het traject van afstand 'l' : 20 km/uur (l) 30 km/uur (l)

Tijd in eerste richting : $t_1 = \frac{\text{afstand}}{\text{snelheid}} = \frac{l}{20}$ uur

Tijd in tweede richting : $t_2 = \frac{\text{afstand}}{\text{snelheid}} = \frac{l}{30}$ uur

Totale tijd om $2l$ af te leggen = $t_1 + t_2$

Gemiddelde snelheid = $\frac{\text{totale afstand}}{\text{totale tijd}} = \frac{2l}{t_1 + t_2} = \frac{2l}{\left(\frac{l}{20} + \frac{l}{30}\right)}$

De lengte is van geen belang want het wordt weg gedeeld:

$$= \frac{2}{\left(\frac{1}{20} + \frac{1}{30}\right)} \quad (: \text{dit is gelijk aan het harmonisch gemiddelde : H})$$

$$= 24$$

3. Bekijk de boxplot van de examenresultaten (een copie werd uitgedeeld). Tracht een mogelijke verklaring te vinden voor:
• het feit dat in de box het plus-teken bijna overal beneden de horizontale streep ligt,
• de vreemde plot bij de vakken PE35 en S295.
Zou het hier om een eerste of om een tweede zittijd gaan ? Waarom ?
Reconstrueer (bij benadering) op basis van de boxplot de histogram (of de verdeling) van de punten bij het vak D268.

■ plus-teken < horizontale streep = gemiddelde < mediaan
→ omdat er een hogere kans is op een zeer lage score (0) dan een zeer hoge score (20). Het gemiddelde wordt sterk beinvloed door deze uitliggers, de mediaan niet.

■ Examen PE35 werdt afgelegd door twee studenten: één met een 10 en de andere met een 11. Examen S295 werdt afgelegd door maar één student die een 11 behaalde.

■ Ik denk tweede zit, omdat de mediaan bijna altijd ≥ 10 is. Dus de slaag percentage ligt boven de 50% wat gewoonlijk niet zo is in eerste zit.

■ Verdeling van de punten in D268:

Score klasse	Percentage van de studenten
0 – 4	10%
4 – 6	15%
6 – 9	25%
9 – 11	25%
11 – 12	15%
12 - 13	10%

Teken een histogram.

4. Gegevende de volgende tijdreeks:

(01.01.1996, 100)
(01.01.1997, 150)
(01.01.1998, 236)
(01.01.1999, 529)
(01.01.2000, 1200).

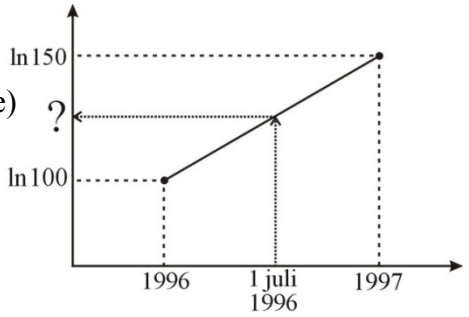
• Stel deze reeks voor in het vlak waarbij de verticale as een logaritmische schaal heeft,
• Gegeven deze grafiek, wat kun je zeggen omtrent de evolutie in deze cijfers,
• Interpolleer de waarde op 1 juli 1996.

.

De data:

Datum	Waarde	Log waarde
01.01.1996	100	ln100
01.01.1997	150	ln150
01.01.1998	236	ln236
01.01.1999	529	ln529
01.01.2000	1200	ln1200

■ Wat is de waarde op 1 juli 1996?
Dit ligt half weg tussen 1996 en 1997 dus ook half weg tussen ln100 en ln150:
(we veronderstellen een constant groeiritme in die periode)



$$= \frac{\ln 100 + \ln 150}{2}$$

$$= \frac{\ln(100 \text{ x } 150)}{2} = \ln \sqrt{100 \text{ x } 150}$$

Het BBP op 1 juli 1996 = <u>122,47</u>

(! dit is niet gelijk aan $\frac{150+100}{2}$)

5. Meerkeuzevragen. Geef telkens een korte verantwoording.

5.A. Indien de observaties $S_x : x_1, x_2, ...x_n$ vervangen worden door $S_y : y_1 = \alpha x_1 + \beta,\ y_2 = \alpha x_2 + \beta, ...,\ y_n = \alpha x_n + \beta$ met α verschillend van 0, wat kan je dan zeggen over het gemiddelde van S_y (genoteerd met $\overline{y}$) en het gemiddelde van S_x (genoteerd met $\overline{x}$):

A • $\overline{y} = \overline{x}$
B • $\overline{y} = \alpha \overline{x}$
C • $\overline{y} = \overline{x} + \beta$
D • $\overline{y} = \alpha \overline{x} + \beta$

Omdat het gemiddelde een centrum maat is; zal de optelling ($+\beta$) en de vermenigvuldiging (maal α) allebei effect hebben op het gemiddelde.
<u>Antwoord:</u> **D** • $\overline{y} = \alpha \overline{x} + \beta$

5.B. Het inkomen tussen 2 personen is als volgt verdeeld : 0.2 mio euro, 0.3 mio euro.
Welke van de volgende inkomensverdelingen (met 4 personen) heeft een identieke Lorenzcurve?

A • 0.2 mio euro, 0.25 mio euro, 0.25 mio euro, 0.3 mio euro,
B • 0.1 mio euro, 0.125 mio euro, 0.125 mio euro, 0.15 mio euro,
C • 0.1 mio euro, 0.1 mio euro, 0.15 mio euro, 0.15 mio euro,
D • de Lorenzcurve met 4 personen zal steeds verschillen van de curve met 2 personen.

Twee personen : 0.2 mio euro, 0.3 mio euro
totaal = 0,5 mio euro
de onderste 50% van de bevolking heeft 40% van het inkomen ($0,2/0,5$)
de bovenste 50% van de bevolking heeft 60% van het inkomen ($0,3/0,5$)

A: $\Sigma = 1,0$: onderste 50% = 45% : bovenste 50% = 55%
B: $\Sigma = 0,5$: onderste 50% = 45% : bovenste 50% = 55%
C: $\Sigma = 0,5$: onderste 50% = 40% : bovenste 50% = 60%
<u>Antwoord:</u> **C** • 0.1 mio euro, 0.1 mio euro, 0.15 mio euro, 0.15 mio euro,

5.C. Hoeveel datasets van lengte 4 genereren de volgende statistieken: gemiddelde = 0, variantie = 1, skewness = 0?
A • geen enkele,
B • precies 1,
C • precies 2,
D • oneindig veel.

Een dataset van lengte vier : $\{a,b,c,d\}$
je hebt drie voorwaarden:
1. gemiddelde = 0 : $\frac{a+b+c+d}{4}=0 \rightarrow a+b+c+d=0$
2. variantie =1 : $\frac{a^2+b^2+c^2+d^2}{3}=1 \rightarrow a^2+b^2+c^2+d^2=3$
3. skewness = 0 : $a^3+b^3+c^3+d^3=0 \rightarrow a^3+b^3+c^3+d^3=0$

Stel : $d=-a$ en $c=-b$
$\rightarrow 2(a^2+b^2)=3$
en $a^2+b^2=\frac{3}{2}$
Antwoord: **D** • oneindig veel.

5.D. Dertien studenten leggen een examen af.
Van twaalf studenten zijn de scores (op 30pt) gekend:
11 18 12 21 16 19 6 19 18 20 24 14.
Het resultaat van de 13de student zal geen enkele invloed hebben op:
A • het gemiddelde,
B • het eerste kwartiel,
C • de mediaan,
D • het derde kwartiel.

De dataset ordenen : 6 11 12 14 16 18 18 19 19 20 21 24
Het gemiddelde zal zeker invloed hebben. vb. als de 13^{de} student een 0 haalt.
De kwartielen zullen ook beinvoled worden.
De mediaan: met 12 studenten is de mediaan het $6{,}5^{de}$ getal = $\frac{18+18}{2}=18$
met 13 studenten is de mediaan het 7^{de} getal :
als de 13^{de} student <18 behaalt, dan is de nieuwe mediaan: $=18$
als de 13^{de} student >18 behaalt, dan is de nieuwe mediaan: $=18$
als de 13^{de} student $=18$ behaalt, dan is de nieuwe mediaan: $=18$
Antwoord: **C** • de mediaan

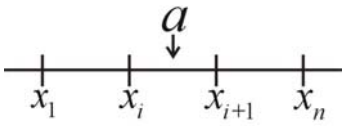
6. Beschouw een dataset $\{x_1,x_2,...,x_n\}$ van lengte n.
Beschouw de afbeelding $a \rightarrow \sum_{i=1}^{n}|x_i-a|$
Toon aan dat deze afbeelding een minimum bereikt in de mediaan van de dataset.

De afbeelding $f: a \to \sum_{j=1}^{n} |x_j - a|$ bereikt een minimum in a= de mediaan van $\{x_1,...,x_n\}$.

We bekijken het stijgen en dalen van f.

Probleem: voor $a = x_j$ is f niet afleidbaar.

Stel $x_i < a < x_{i+1}$



$f(a) = (a - x_1) + ... + (a - x_i) + (x_{i+1} - a) + ... + (x_n - a)$

: al de getallen die je optelt zijn positief

$f'(a) = i - (n - i) = 2i - n$

$f'(a) < 0$ indien $2i - n < 0$

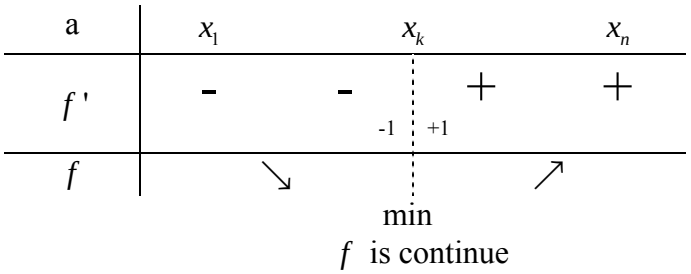
$f'(a) > 0$ indien $2i - n > 0$

* $n = 2k - 1$, oneven

$2(k-1) - n = 2k - 2 - n$

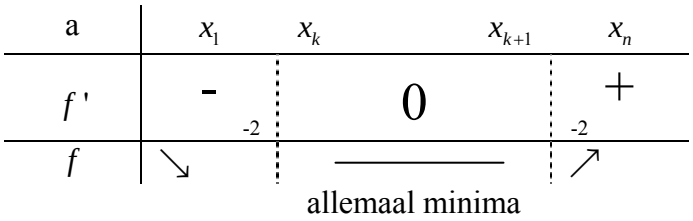
$= -1 < 0$

$2k - n = 1 > 0$



* $n = 2k$, even

mediaan $= \dfrac{xk + xk + 1}{2}$



Tip: het is gemakkelijker als je n vervangt door een concreet getal, vb. n=3

7. In 1999 pakte een bank uit met de volgende reclame (de bedragen zijn nog in BEF):

De kracht van regelmatig sparen!

Iedere belegger wenst een zo hoog mogelijk rendement te halen uit zijn belegging. Als U zich afvraagt hoe U dit op een eenvoudige, maar efficiente, wijze kunt realiseren, hebt U er alle belang bij deze tekst aandachtig te lezen.

U haalt een hoog rendement indien U steeds investeert wanneer de koersen laag zijn en niet wanneer de koersen hoog staan. Alleen, dit is theorie. In de praktijk is deze eenvoudige regel niet toepasbaar. Niemand kan immers met zekerheid de toekomst voorspellen. En dit koffiedik kijken is nu juist wat U nodig hebt om te kunnen investeren op het laagste moment.

Er bestaat nochtans een eenvoudige manier om uw doelstelling zoveel mogelijk te benaderen: door regelmatig een vast bedrag te investeren, is uw gemiddelde koers van aankoop steeds lager dan de gemiddelde koers op de markt. En bent U in staat om de beurs te kloppen. Te simpel om waar te zijn ? Toch niet. We verduidelijken met een (fictief) voorbeeld.

De heer Thomaes belegt door middel van een doorlopende opdracht maandelijks 10 000 BEF (=250 euro) in Open Life. Afhankelijk van de koers, telkens geldig op het moment van aankoop, verwerft hij meer of minder aandelen met dat vast bedrag:

Bedrag	**Koers**	**Aantal**
10 000	1,00	10 000
10 000	0,85	11 765
10 000	0,95	10 526
10 000	1,05	9 524
10 000	1,02	9 804
10 000	1,20	8 333
10 000	1,26	7 937
10 000	1,00	10 000
80 000		77 889

Het is duidelijk dat dhr. Thomaes steeds meer eenheden verwerft als de koers lager is en minder wanneer de koers hoger is. De gemiddelde koers over bovenstaande periode is 1,041. De gemiddelde koers, waartegen dhr. Thomaes aangekocht heeft is lager! Deze bedraagt 1,027 (=80 000/77 889).

Ondanks het feit dat dhr. Thomaes geen glazen bol bezit, is hij toch in zijn opzet geslaagd. En dit door op regelmatige tijdstippen een vast bedrag te beleggen.

Met een kleine regelmatige spaarinspanning bouwt U bovendien stelselmatig een kaptiaal op. En dit op een manier die U vaak niet merkt. Vele kleintjes maken ook in dit geval samen ééen groot. U bent overtuigd van de kracht van regelmatig sparen ?
De verdere tekst van deze reclame is voor de oefening niet relevant.

Opdrachten

• Heeft deze reclame U kunnen overtuigen ?
• Bereken het rekenkundig gemiddelde van de koersen (=1,041), bereken ook het harmonisch gemiddelde.
• Voor een willekeurige dataset van positieve getallen is het harmonisch gemiddelde altijd kleiner dan (of gelijk aan) het rekenkundig gemiddelde. Toon deze eigenschap aan voor een dataset van lengte twee.
• Dit onderdeel maakt gebruik van verwachte waarde.
Veronderstel dat dhr. Thomaes geen glazen bol bezit (bijgevolg enkel de eerste koers kent) en de verwachte waarde van zijn belegging maximaliseert. Welke strategie is dan de beste: regelmatig beleggen met kleine bedragen of één keer een groot bedrag inzetten.

■ rekenkundig gemiddelde van de koersen :

$$\overline{x} = \frac{1,00 + 0,85 + ... + 1,26 + 1,00}{8} = 1,041$$

harmonisch gemiddelde van de koersen :

$$H = \frac{n}{\left(\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}\right)} = \frac{8}{\left(\frac{1}{1,00} + \frac{1}{0,85} + ... + \frac{1}{1,00}\right)} = 1,027$$

■ Bewijs dat $H \leq \overline{x}$ voor de dataset $\{x, y\}$:

$$H = \frac{2}{\left(\frac{1}{x}+\frac{1}{y}\right)} = \frac{2}{\left(\frac{x+y}{xy}\right)} = \frac{2xy}{x+y} \overset{?}{\leq} \frac{x+y}{2} = \overline{x}$$

$$4xy \leq (x+y)^2$$
$$0 \leq (x+y)^2 - 4xy$$
$$0 \leq (x+y)^2 \text{ : dit klopt}$$

■ Doel van de klant = de verwachte waarde van de portefuille maximeren op het einde.

<u>* 1 keer een groot bedrag beleggen:</u> bedrag van 4B op tijdstip t: (t = 1, 2, 3, 4)

t :			koers :	aantal :
	1	↓	k_1	$4B/k_1$
	2		k_2	$4B/k_2$
	3		k_3	$4B/k_3$
	4		k_4	$4B/k_4$

Verwachte waarde van de portefuille indien Thomaes <u>lukraak</u> één tijdstip kiest:

$$E = \frac{1}{4}\left(\frac{4B}{k_1}+\frac{4B}{k_2}+\frac{4B}{k_3}+\frac{4B}{k_4}\right) \text{ x } k_4$$

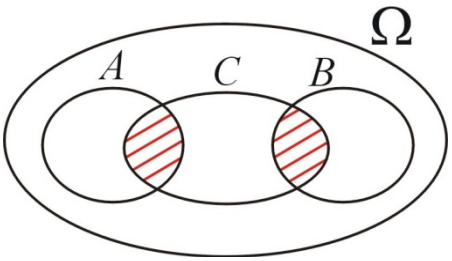
<u>* 4 keer een bedrag B beleggen:</u>

$$V = \left(\frac{B}{k_1}+\frac{B}{k_2}+\frac{B}{k_3}+\frac{B}{k_4}\right) \text{ x } k_4$$

<u>Merk op:</u> $E = V$: dus het argument klopt niet

Oefenzitting 2 – Kansruimten, telproblemen, Bayes,...

1. Beschouw een kansruimte (Ω, G, P). De gebeurtenissen A en B zijn disjunct. Toon aan: $\forall C \in G : P\big((A \cup B) \cap C\big) = P(A \cap C) + P(B \cap C)$



Omdat de gebeurtenissen A en B disjunct zijn, zijn ook de gebeurtenissen $A \cap C$ en $B \cap C$ disjunct. We gebruiken de optelregel ($P(A \cup B) = P(A) + P(B)$):

$$\begin{aligned} P\big((A \cup B) \cap C\big) &= P\big((A \cap C) \cup (B \cap C)\big) \\ &= P(A \cap C) + P(B \cap C) \end{aligned}$$

2. Beschouw een kansruimte (Ω, G, P). De gebeurtenissen A en B zijn onafhankelijk. Toon aan dat ook de gebeurtenissen A en $\overline{B} = \Omega - B$ onafhankelijk zijn.

Gegeven dat A en B onafhankelijk zijn, bewijs dat A en $\overline{B}$ onafhankelijk zijn.

[vb. indien hoog IQ en Meisje onafhankelijk zijn ook hoog IQ en Jongen onafhankelijk]

We willen bewijzen dat: $P(A \cap \overline{B}) = P(A).P(\overline{B})$

$A \cap \overline{B} = A \setminus (A \cap B)$

$\downarrow$

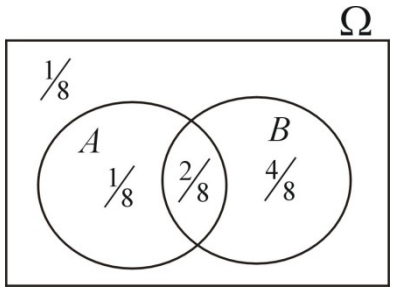
$$\begin{aligned} P(A \cap \overline{B}) &= P(A) - P(A \cap B) && \text{: via de somregel } A \cap B \not\in A \\ &= P(A) - P(A).P(B) && \text{: A en B zijn onafhankelijk} \\ &= P(A).\big(1 - P(B)\big) && \text{: P(A) buiten de hakjes brengen} \\ &= P(A).P(\overline{B}) && \text{: complementaire gebeurtenis} \end{aligned}$$

3. Beschouw een kansruimte (Ω, G, P). Van gebeurtenissen A en B weten we dat $P(A \cup B) = 7/8$, $P(A \cap B) = 2/8$ en dat $P(\overline{A}) = 5/8$. Notatie: $\overline{A} = \Omega - A$.
Bepaal: $P(A)$, $P(B)$, $P(A \cap \overline{B})$, $P(\overline{A} \cup \overline{B})$, $P(\overline{A} \cup B)$, en $P\big(A \cup (A \cap \overline{B})\big)$

- $P(A) = 1 - P(\overline{A}) = 1 - \dfrac{5}{8} = \dfrac{3}{8}$
- $P(A \cap B) = P(A) + P(B) - P(A \cup B)$: we weten niet of A en B onafhankelijk zijn

  $P(B) = P(A \cap B) + P(A \cup B) - P(A) = \dfrac{2}{8} + \dfrac{7}{8} - \dfrac{3}{8} = \dfrac{6}{8}$

Met deze informatie kan je al een grafiek tekenen:
Deze is zeer nuttig om de rest te berekenen:



■ $P(A \cap \overline{B}) = P(A) - P(A \cap B) = \frac{3}{8} - \frac{2}{8} = \frac{1}{8}$

■ $P(\overline{A} \cup \overline{B}) = P(\overline{A}) + P(\overline{B}) - P(\overline{A} \cap \overline{B}) = \frac{5}{8} + \frac{2}{8} - \frac{1}{8} = \frac{6}{8}$

■ $P(\overline{A} \cup B) = P(\overline{A}) + P(B) - P(\overline{A} \cap B) = \frac{5}{8} + \frac{6}{8} - \frac{4}{8} = \frac{7}{8}$

■ $P\left(A \cup (A \cap \overline{B})\right) = P(A) + P(A \cap \overline{B}) - P\left(A \cap (A \cap \overline{B})\right) = \frac{3}{8} + \frac{1}{8} - \frac{1}{8} = \frac{3}{8}$

4. Beschouw een kansruimte (Ω, G, P). De gebeurtenissen A en B zijn onafhankelijk en voldoen aan $P(A) = 0,3$ en $P(B) = 0,5$.
Bepaal $P(A \cap B)$ en $P(A \cup B)$.

■ $P(A \cap B) = P(A).P(B)$ (: A en B zijn onafhankelijk)
$= (0,3)(0,5) = 0,15$

■ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (: optelregel, in het formularium)
$= 0,3 + 0,5 - 0,15 = 0,65$

5. Beschouw een kansruimte (Ω, G, P) en de gebeurtenissen A, B, en C. Geef de definitie voor 'de drie geberutenissen A, B, en C zijn onafhankelijk'.
Veronderstel dat de gebeurtenissen A, B, en C paarsgewijs onafhankelijk zijn, dat $P(A) = P(B) = P(C) = 1/2$ en dat $P(A \cap B \cap C) = 1/4$.
Toon aan dat de drie gebeurtenissen A, B, en C <u>niet</u> onafhankelijk zijn.

Er zijn vier voorwaarden voor 'de drie geberutenissen A, B, en C zijn onafhankelijk':
1. $P(A \cap B) = P(A)$ x $P(B)$
2. $P(A \cap C) = P(A)$ x $P(C)$
3. $P(B \cap C) = P(B)$ x $P(C)$
4. $P(A \cap B \cap C) = P(A)$ x $P(B)$ x $P(C)$

■ We gebruiken de formule voor de voorwaardelijke kans;
als A, B, en C onafhankelijk zijn dan is $P(A|B|C) = P(A)$.

$P(A|B|C) = \frac{P(A \cap B \cap C) =}{P(B).P(C)} = \frac{1/4}{(1/2).(1/2)} = 1 \quad \neq P(A) = 1/2$

→ A, B, en C zijn afhankelijk.

6. Twee staatslieden ontmoeten elkaar in Cairo. De ene spreekt enkel Arabisch, de andere spreekt Engels, Frans, en Duits. Van de beschikbare tolken die Arabisch spreken zijn er:
• 80% die Engels spreken,
• 50% die Frans spreken,
• 30% die Duits spreken.
Zij Ω de verzameling van de beschikbare tolken. Veronderstel dat de gebeurtenissen 'Engels spreken', 'Frans spreken', en 'Duits spreken' onafhankelijk zijn. Hoe groot is de kans dat met behulp van een lukraak getrokken tolk de twee staatslieden met elkaar kunnen overleggen?

$P(E) = 0,8$: 'kans dat de tolk Engels spreekt'

$P(\overline{E}) = 0,2$: 'kans dat de tolk geen Engels spreekt'

$P(F) = 0,5$: 'kans dat de tolk Frans spreekt'

$P(\overline{F}) = 0,5$: 'kans dat de tolk geen Frans spreekt'

$P(D) = 0,3$: 'kans dat de tolk Duits spreekt'

$P(\overline{D}) = 0,7$: 'kans dat de tolk geen Duits spreekt'

In plaats van al de combinaties te berekenen waar dat ze wel kunnen overleggen (dit zijn er veel om dat de ene drie talen spreekt) is het gemakkelijker om de kans te berekenen dat ze niet kunnen overleggen:

$P(\text{ze kunnen niet overleggen}) = P(\overline{E}).P(\overline{F}).P(\overline{D}) = (0,2)(0,5)(0,7) = 0,07$

Kans dat ze wel kunnen overleggen = 1-0,07 = 0,93.

De kans dat met behulp van een lukraak getrokken tolk de twee staatslieden met elkaar kunnen overleggen is 93%.

7. Een eerlijke dobbelsteen wordt tweemaal geworpen. Indien gegeven is dat de gegooide getallen verschillend zijn, bepaal de kans dat:
• de som zes is,
• er minstens één 1 gegooid werd,
• de som strikt kleiner is dan vijf.

Omdat we weten dat de gegooide getallen verschillend zijn, beperken we de uitkomstenverzameling (Ω) tot de 30 koppels met verschillende coordinaten:

$\Omega = \{(1,2),(1,3),...,(1,6),(2,1),(2,3),...,(6,5)\}$.

Elk element wordt getrokken met een kans gelijk aan $1/36$ (uniforme verdeling).

■ Er zijn 4 koppels in Ω met de som gelijk aan 6: (1,5),(2,4),(4,2),(5,1)
De kans dat de some zes is = $4/30$.

■ Er zijn 10 koppels in Ω een cordinaat gelijk aan 1: (1,2),(1,3),...,(1,6),(2,1),...,(6,1)
De kans dat er minstens een 1 gegooid werd = $10/30 = 1/3$.

■ Er zijn 4 koppels in Ω met de som strikt kleiner is dan vijf: (1,2),(1,3),(3,1),(2,1)
De kans dat de som strikt kleiner is dan vijf = $4/30$.

8. Een klant wordt gevraagd om vijf verschillende biersoorten $a,b,c,d,$ en e te rangschikken volgens zijn voorkeur. Veronderstel dat deze voorkeursordening (genoteerd met '>') strikt is, m.a.w. er treden geen indifferenties op.
Hoeveel verschillende rangschikkingen zijn er mogelijk? Wat is de kans dat hij

a > b > c > d > e

als voorkeur heeft?
Wat is de kans dat zijn top drie bestaat uit de verzameling $\{a,b,c\}$?

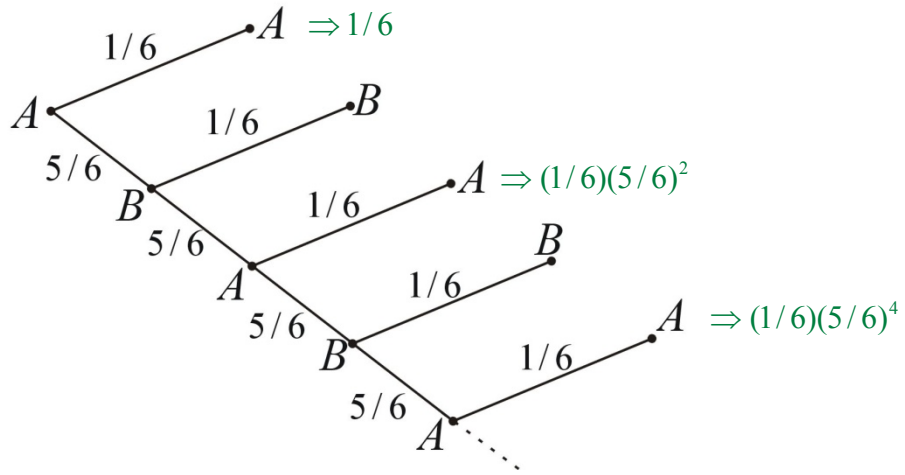
■ Vijf voorwerpen kunnen op 5! = 120 verschillende wijzen geordend worden.

■ Hiervan zijn er 3! x 2! = 12 waarbij het trio $\{a,b,c\}$ bovenaan staat en het duo $\{d,e\}$ onderaan.

De kans dat de top drie samenvalt met de verzameling $\{a,b,c\}$ is gelijk aan:

$$\frac{3!\ 2!}{5!} = \frac{12}{120} = 10\%$$

9. Twee politici A en B houden een debat. Om te bepalen wie eerst zal spreken, maakt men gebruik van een eerlijke dobbelsteen. Om beurten mogen de politici de dobbelsteen gooien. Wie het eerst een 6 werpt, mag het eerst spreken. Indien A het eerst mag gooien, wat is de kans dat hij het eerst mag spreken?



1ste Methode:

$$P(\text{politicus A mag eerst spreken}) = \frac{1}{6} + \frac{1}{6}\left(\frac{5}{6}\right)^2 + \frac{1}{6}\left(\frac{5}{6}\right)^4 + \ldots.$$

$$= \frac{1}{6}\cdot\left(1 + \left(\frac{5}{6}\right)^2 + \left(\frac{5}{6}\right)^4 + \ldots\right)$$

$$= \frac{1}{6}\cdot\frac{1}{1-(25/36)} = \frac{6}{11}$$

$$1 + c + c^2 + c^3 + \ldots = \frac{1}{1-c} \qquad -1 < c < 1$$

2e Methode:

P=kans om een zes te gooien x 1 + kans om (indien geen zes) terug aan de beurt te komen x P

$$P = \frac{1}{6} \text{ x } 1 + \frac{25}{36} \text{ x P}$$

$$\frac{11}{36}P = \frac{1}{6} \text{ of } P = \frac{6}{11}$$

3de Methode:

P = kans dat politicus A begint
Q = kans dat politicus B begint

P+Q=1

$Q = \frac{5}{6}P$: om dat Q eerst nog aan zet moet komen

$$P + \frac{5}{6}P = 1 \text{ of } P = \frac{6}{11}$$

10. • Guido heeft twee kinderen. Wat is de kans dat Guido twee dochters heeft, indien je weet dat het oudste kind een meisje is?
Je mag veronderstellen dat bij de geboorte de kans op een jongen gelijk is aan de kans op een meisje.
• Guido heeft twee kinderen. Wat is de kans dat Guido twee dochters heeft, indien je weet dat een van de kinderen een meisje is?
• Guido heeft (nog steeds) twee kinderen. In een winkel ontmoeten we Guido samen met een meisje, dat hij aan ons voorstelt als zijn dochter. Wat is de kans dat Guido twee dochters heeft?
Je mag veronderstellen dat indien Guido een zoon en een dochter heeft, beide kinderen evenveel kans hebben om met hem te gaan winkelen.

■ Je weet dat het oudste kind een meisje is dus zijn er maar twee mogelijkheden:

$$\Omega' = \{(M,J),(M,M)\}$$

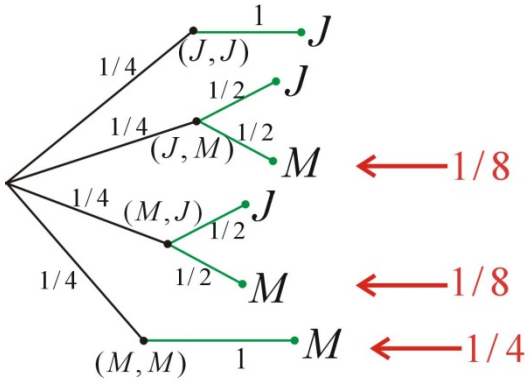
$$P(\{(M,M)\}) = \frac{1}{2}$$

■ Je weet dat Guido een dochter heeft, er zijn nu drie mogelijkheden:

$$\Omega' = \{(M,J),(J,M),(M,M)\}$$

$$P(\{(M,M)\}) = \frac{1}{3}$$

■ We tekenen een kansboom:



$$P(\{(M,M)\}) = \frac{(1/4)}{(1/8)+(1/8)+(1/4)} = \frac{1}{2}$$

11. Je bent winnaar van een quiz. Op een tafel liggen drie omslagen waarvan je er één mag kiezen. De hoofdprijs (een cheque van 1 miljoen euro) is verborgen in één van deze omslagen, de andere twee zijn leeg. Nadat je een omslag gekozen hebt, opent de spelleider één van de twee overgebleven omslagen en toont dat deze leeg is. Vervolgens vraagt hij of je uw omslag wil ruilen voor de andere (nog ongeopende) omslag. Ga je op dit aanbod in?

Definitie: 1 = de enveloppe bevat de prijs
2, 3 = de andere enveloppes

We hebben twee stochastische varibelen:
• A = de eerst gekozen enveloppe
$A \sim U(1,2,3)$

• B = de tweede gekozen enveloppe

stel: B = A : kans p (: enveloppe bewaren)
B ≠ A : kans 1-p (: de enveloppe ruilen)

Vraag: $P(B=1)$
Wet van de totale kans: (een kansboom)
$P(B=1) = P(B=1|A=1).P(A=1) + P(B=1|A\neq 1).P(A\neq 1)$
: de kans dat de tweede gekozen enveloppe de prijs bevat is de som van de twee mogelijke winnende situaties: ofwel zit je in het begin met de juiste enveloppe en bewaar je hem ofwel zit je in het begin met de foute enveloppe en ruil je voor de juiste.

$$P(B=1) = p.\frac{1}{3} + (1-p).\frac{2}{3}$$
$$= \frac{2-p}{3}$$

→ $P(B=1)$ bereikt een maximum voor p = 0 dus heb je er baat bij om te wisselen.

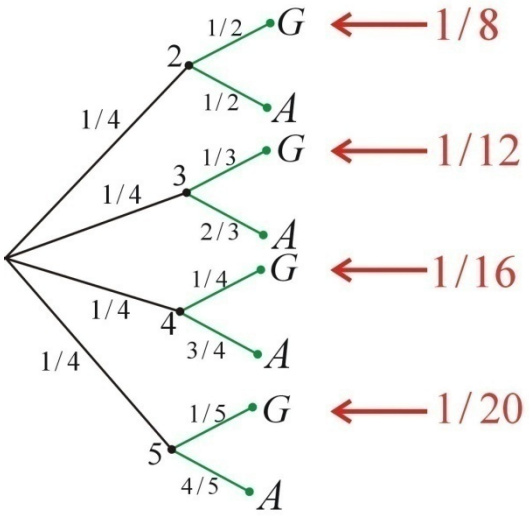
12. Een commissie van 5 personen wordt willekeurig gekozen uit 12 echtparen.
• Hoeveel verschillende samenstellingen van de commissie zijn mogelijk?
• In de veronderstelling dat 2 echtgenoten niet beiden tot de commissie mogen behoren, hoeveel samenstellingen zijn er dan nog mogelijk?

■ Je trekt 5 personen uit een set van 24. Dit kan op $\binom{24}{5} = 42504$ verschillende manieren.

■ Voor de eerste persoon heb je 24 mogelijkheden. Voor de tweede nog 22 mogelijkheden; de partner van de eerst getrokkene wordt immers uitgesloten. Voor de derde (vierde, vijfde) persoon zijn er 20 (18, 16) mogelijkheden. In totaal zijn er bijgevolg 24×22×20×18×16 mogelijkheden om vijf personen te trekken. Dit getal moet echter nog gedeeld worden door 5!. Immers, de volgorde waarin de trekkingen gebeuren, is niet belangrijk (het gaat om die commissie van 5 personen, en niet om de volgorde waarin die getrokken werd). Dit geeft (1/120) × 24 × 22 × 20 × 18 × 16 = 25 344 verschillende commissies van 5 personen waarin geen echtparen zetelen.

Alternatieve route. We trekken eerst 5 echtparen uit de groep van 12. Vervolgens trekken we één persoon uit elk van de gekozen echtparen. Dit geeft $\binom{12}{5}$ x $2^5 = 25344$ mogelijkheden.

13. Veronderstel dat het aantal herten in een reservaat 2, 3, 4, of 5 is met gelijke kansen voor elk van deze aantallen.
• Lukraak wordt één van de herten gevangen, gemerkt en weer losgelaten. De volgende week wordt er opnieuw lukraak een hert gevangen. Het blijkt het gemerkte hert te zijn. Bepaal de kans dat het reservaat slechts 2 herten herbergt?
• Lukraak worden 2 herten gevangen, gemerkt en weer losgelaten. De volgende week worden opnieuw 2 herten gevangen. Beiden blijken gemerkt te zijn. Wat is nu de kans dat er slechts 2 herten in het reservaat zijn?

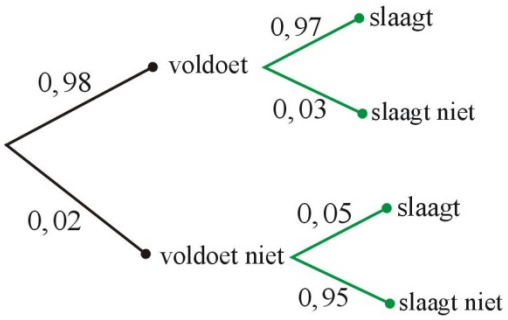


G = gemerkt
A = ander

$$P(\text{2 herten} \mid \text{G}) = \frac{(1/8)}{(1/8)+(1/12)+(1/16)+(1/20)} = \frac{(1/2)}{(1/2)+(1/3)+(1/4)+(1/5)} = \frac{(30/60)}{(77/60)} = \frac{30}{77}$$

* Voor het tweede deel moet je de getallen herschalen met de bijkomende informatie.

14. Alle producten die van een productieband rollen worden onderworpen aan een test. Van alle geproduceerde producten voldoen 98% aan de specificaties. Bij het testen is er een kans van 3% om een goed product te verwerpen en een kans van 5% om een slecht product te aanvaarden.
Wat is de kans dat een product dat slaagt in de test, voldoet aan de specificaties?



We gebruiken Bayes:

$$P(\text{voldoet}|\text{slaagt}) = \frac{P(\text{slaagt}|\text{voldoet}).P(\text{voldoet})}{P(\text{slaagt})} = \frac{(0,97).(0,98)}{(0,97).(0,98)+(0,02).(0,05)} = \frac{0,9506}{0,9516} = 0,99$$

15. De resultaten van een week werk in twee werkplaatsen zijn als volgt:

	Aantal defecte	Aantal goede producten
Werkplaats a	151	563
Werkplaats b	85	307

Zou je stellen dat de waarschijnlijkheid van een defect product onafhankelijk is van de werkplaats? Argumenteer.

In het totaal werden er 1106 producten vervaardigd, waarvan 236 defecte. Dit is een verhouding van 21.3% defecte producten per honderd.
In werkplaats a is deze verhouding gelijk aan 151/714 = 21,1%.
In werkplaats b is deze verhouding gelijk aan 85/392 = 21,7%.
Deze drie verhoudingen liggen allemaal voldoende dicht bij elkaar om de hypothese van onafhankelijkheid te ondersteunen.
Later (hypothese-toetsen) komen we terug op deze oefening.

16. Simpson's paradox. In Belgie werd voor een welbepaalde kwaal en nieuw geneesmiddel N getest en afgewogen tegenover een oud geneesmiddel O. De volgende tabel vat de resultaten samen, gesplitst naar regio's (Vlaanderen, Wallonie).

	Nieuw	pos	neg	Oud	pos	neg
Vlaanderen	144	54	90	36	12	24
Wallonie	36	18	18	144	66	78

Dus, in Vlaanderen werden 144 zieken behandeld met het nieuwe geneesmiddel waarvan 54 met gunstig gevolg. In Wallonie werden 144 zieken behandeld met het oude geneesmiddel waarvan 66 met gunstig gevolg.
Welk geneesmiddel scoort procentueel het beste in Vlaanderen, in Wallonie, in Belgie als geheel? Vind je het antwoord intuitief aanvaardbaar of niet? Leg uit.

In Vlaanderen scoort N hoger: 54/144 = 0.375 tegenover 12/36 = 0.333.
In Wallonië scoort N ook hoger: 18/36 = 0.500 tegenover 66/144 = 0.458.
Maar, in België scoort O hoger: 78/180 = 0.433 tegenover 72/180 = 0.400.
Tegen de intuïtie, vandaar paradox.

17. De vraag of er buitenaards leven is, is en blijft intrigerend. In 1950 formuleerde Enrico Fermi de volgende stelling, nu bekend als 'de paradox van Fermi'. Indien er 'aliens' bestaan, dan moeten er ook aliens bestaan met een meer gevorderde cultuur en kennis als de onze. Het universum is immers zo oud (14 miljard jaar) en zo groot (de Melkweg telt miljarden sterren). Deze aliens zouden, gegeven hun intelligentie, de Aarde al lang gekoloniseerd hebben. Het feit dat we nog altijd geen aliens gezien hebben, bewijst dat er geen aliens zijn.

In 1961 presenteerde Frank Drake een eenvoudig model om het aantal buitenaardse beschavingen te bepalen. Frank Drake was werkzaam als astronoom in Harvard, hij had geen weet van de paradox van Fermi. We beschrijven een aantal, volgens Drake, relevante parameters:

• Het aantal beschavingen in ons melkwegstelsel is evenredig met het aantal sterren in onze Melkweg. Volgens schattingen zijn er in onze Melkweg meer dan 100 miljard sterren.

• Het aantal beschavingen in het volledige universum is evenredig met het aantal sterren in het universum. 100 miljard melkwegen in het universum is wellicht een onderschatting (op basis van waarnemingen met de Hubble Space Telescope).

• Ongeveer de helft van de sterren zou planeten hebben. Volgens Michel Mayor, de ontdekker van de eerste planeet buiten ons Zonnestelsel, ligt die fractie veel hoger, zelfs dicht bij de 100%.

• Niet elke planeet is geschikt om leven te herbergen. De afstand tot de ster is hier cruciaal, denk maar aan ons Zonnestelsel. De fractie van planeten binnen de 'juiste' afstand tot de centrale ster wordt gezet op 11,11% (of 1/9).

• Tenslotte moet er nog nagedacht worden over die fractie van de geschikte planeten waar DNA zich ontwikkelt en overleeft. Zonder uit te weiden over de complexiteit van dit proces, wordt een fractie van 1 biljoenste als zijnde zeer bescheiden (wellicht onderschatting) beschouwd.

Opdrachten:

• Bepaal op basis van deze veronderstellingen de kans dat er leven is in het universum. Herinner dat 1 miljard gelijk is aan 10^9, en 1 biljoen aan 10^{12}. Om met dergelijke grote getallen te rekenen, maak je best gebruik van de logaritme (en van de benadering $\ln(1+x) \approx x$ voor x dicht bij 0).

• Hoe wijzigt deze kans op leven indien ons Zonnestelsel (of onze Melkweg) buiten beschouwing wordt gelaten ?

Bron: AD Aczel, 1998, `Probability 1', Harcourt, San Diego.

p = de kans op leven in het universum

1-p = kans op geen leven in het universum

$= q^{\text{aantal sterren}} = q^{(100\text{x}10^9)(100\text{x}10^9)}$

met q gelijk aan de kans dat X = 0

$$q = \left(1 - \frac{1}{2}.\frac{1}{9}.\frac{1}{10^{12}}\right)$$

Dus: $1 - p = \left(1 - \frac{1}{2}.\frac{1}{9}.\frac{1}{10^{12}}\right)^{10^{22}}$

We gebruiken het eigenschap van de logaritme:

$$\ln(1-p) = 10^{22}.\ln\left(1 - \frac{1}{2}.\frac{1}{9}.\frac{1}{10^{12}}\right)$$

$$\approx -\frac{10^{22}}{(20)(10^{12})} = -\frac{10^{10}}{20}$$

$(1-p) = e^{-\frac{10^{10}}{20}} = 0,000.....$

Dus: $\underline{p = 1}$

18. John Allen Paulos is een verwoed nieuwskijker. Zappend tussen het NBC- en ABC-news van 20.00 uur stelt hij vast dat deze twee zenders dikwijls hetzelfde onderwerp op hetzelfde tijstip behandelen. Werken deze zenders samen, of speelt het toeval hier een rol?

Om wat te modelleren, maken we de volgende veronderstellingen. In een nieuwsuitzending zijn er een viertal items belangrijk genoeg om uitgewerkt te worden. Beide zenders nemen deze hoofdpunten op. We veronderstellen ook dat elk hoofdpunt ongeveer even veel tijd krijgt toebedeeld.

• Bepaal de kans dat er minstens één van de hoofdpunten op dezelfde plaats gerangschikt staat.

■ Teloefening – permutaties: Hoeveel zonder ongewijzigde plaatsen? (met vaste plaatsen)

Oefening voor 2 items :

Nieuws 1	Nieuws 2
•1	•**1** •2
•2	•**2** •1

→ 1 van de mogelijke 2 situaties overlappen

→ kans op overlapping = $\frac{1}{2}$

Oefening voor 3 items :

Nieuws 1	Nieuws 2					
•1	**•1**	**•1**	•2	•2	•3	•3
•2	**•2**	•3	•1	•3	•1	**•2**
•3	**•3**	•2	**•3**	•1	•2	•1

→ 4 van de 6 situaties overlappen

→ kans op overlapping = $\frac{4}{6}$

$P\left(\begin{array}{c}\text{beide zenders tonen}\\ \text{simultaan hetzelfde item}\end{array}\right)$ → geval 1 = 1/2

→ geval 2 = 4/6

Oefening voor 4 items : $\{1,2,3,4\}$

Precies 1 vast punt $=\binom{4}{1}.2=8$: er zijn 8 situaties uit de 24 waar precies 1 item overlapt

Precies 2 vaste punten $=\binom{4}{2}=6$

(als je er 3 hebt, dan heb je automatisch 4)

Precies 4 vaste punten $=1=1$

$\underline{\Sigma=15}$

kans $=\frac{15}{24}=0,63$

Met andere woorden, indien de zenders de volgorde van de vier hoofdpunten lukraak trekken, dan zal in 63% van de gevallen ten minste één hoofdpunt tegelijkertijd worden uitgezonden.

19. Zij (Ω, F, P) een kansruimte. We zeggen dat de gebeurtenis F een negatieve invloed heeft op de gebeurtenis E indien:

$$P(E|F) < P(E)$$

Opdrachten:

• Voor elke gebeurtenis E, heeft de complementaire gebeurtenis van E een negatieve invloed op E. Toon aan.

• Indien F een negatieve invloed heeft op E, dan heeft E een negatieve invloed op F. Toon aan.

• De relatie 'heeft een negatieve invloed op' is niet transitief. Toon aan.

■ Toon aan dat: $P(E|\bar{E}) < P(E)$.

We gebruiken de formule van de voorwaardelijke kans:

$\frac{P(E \cap \bar{E})}{P(\bar{E})} < P(E)$

$0 < P(E)$: bewezen

Deze uitkomst is logisch; wat is de kans dat het een jongen is gegeven dat het een meisje is? = 0

■ Toon aan dat: indien $P(E|F) < P(E)$ dan $P(F|E) < P(F)$.
We gebruiken opnieuw de formule van de voorwaardelijke kans:

$$\frac{P(F \cap E)}{P(E)} \overset{?}{<} P(F)$$

$$P(F \cap E) \overset{?}{<} P(F).P(E)$$

$$\frac{P(F \cap E)}{P(F)} \overset{?}{<} P(E)$$: dit is gegeven : $P(E|F) < P(E)$

■ E negatief op $\bar{E}$, $\bar{E}$ negatief op $\bar{E} = E$
maar $P(E|E) = 1 > P(E)$
Dus; E heeft een positief effect op E

20. De regel van Bayes dient om 'initiële beliefs' bij te sturen wanneer er nieuwe (relevante) informatie toekomt. De formulering is als volgt: zij A een gebeurtenis en laat de gebeurtenissen $B_1,...,B_n$ een partitie vormen van de populatie Ω, dan:

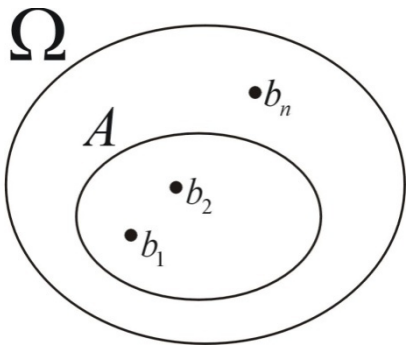
$$P(B_i|A) = \frac{P(A|B_i) \text{ x } P(B_i)}{\sum_k P(A|B_k) \text{ x } P(B_k)}$$

Een typisch voorbeeld is de dokter die op basis van een aantal symptomen en/of testen de ziekte van een welbepaalde patiënt probeert te achterhalen. We beschouwen een discrete uitkomstenverzameling $\Omega = \{b_1, b_2, ..., b_n\}$ hierbij wordt b_i geïnterpreteerd als 'die patiënt heeft ziekte b_i', met $i = 1,...,n$. Merk op dat:

$$\Omega = \{b_1\} \cup \{b_2\} \cup ... \cup \{b_n\}$$
$$B_1 \quad B_2 \quad B_n$$

In eerste instantie interpreteren we de gebeurtenis A als 'de symptomen die de patiënt vertoont sluiten een aantal van de ziekten uit', (Ω wordt gereduceerd tot A).
• **Toon aan** dat in de bijgestuurde verdeling een ziekte b_i ofwel geëlimineerd wordt, ofwel een grotere kans krijgt toebedeeld.

$\Omega = \{b_1, b_2, ..., b_n\}$ is discreet
$p_i = P(b_i) > 0$



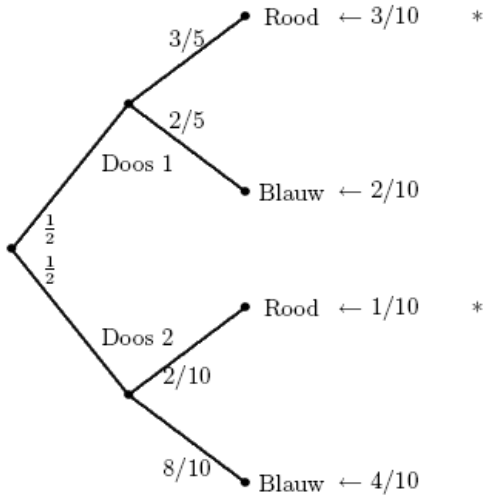
$$q_i = P(\{b_i\}|A) = \frac{P(\{b_i\} \cap A)}{P(A)} = 0 \text{ indien } b_i \notin A$$
$$= \frac{P_i}{P(A)} \text{ indien } b_i \in A$$

merk op $q_i = 0$ of $q_i \geq p_i > 0$

Oefenzitting 3 – Kansrekenen, dichtheidsfuncties, verwachting, …

1. Doos 1 bevat drie rode en twee blauwe knikkers, doos 2 bevat twee rode en acht blauwe knikkers. Verder, is er een eerlijk muntstuk.
Het muntstuk wordt getost. Indien kruis, dan wordt er lukraak een knikker uit doos 1 getrokken. Indien munt, dan wordt er lukraak een knikker uit doos 2 getrokken.
• Bepaal de kans dat een rode knikker getrokken wordt,
• Bepaal de kans dat 'kruis' getost werd, gegeven dat de getrokken knikker rood is.

■ We tekenen een kansboom. Startpunt is het tossen van de munt. Daar vertrekken twee takken, beiden met kans gelijk aan 0.5. Daarna wordt er uit de gepaste doos een knikker getrokken. De vier eindpunten van de kansboom zijn vergezeld van hun kansen. Ter controle: deze kansen tellen op tot 1.



De kans op een rode knikker is gelijk aan 3/10 + 1/10 = 4/10 = 0.4.

■ Gegeven dat een rode knikker getrokken werd, wat is de kans dat deze afkomstig is uit Doos 1.
We reduceren de kansboom tot die eindpunten met een ster.
(De kansen 3/10 en 1/10 worden dan herschaald tot 3/4 en 1/4.)

$$P(\text{Doos 1} \mid \text{Rood}) = \frac{P(\text{Doos 1} \cap \text{Rood})}{\text{P(Rood)}} = \frac{3/10}{3/10+1/10} = \frac{3}{4}$$

2. Twee schakers A en B spelen een tornooi. De regels zijn als volgt. Er worden maximaal tien partijen gespeeld. De eerste die een partij wint, wint ook het tornooi. Indien er 10 keer gelijkspel (remise) gespeeld wordt, dan is ook de eindscore remise. Bij elke partij heeft speler A een kans van 40% om te winnen, speler B een kans van 30% om te winnen, en is er 30% kans op een gelijkspel.
• Geef de dichtheid van het aantal partijen die gespeeld zullen worden verifieer uiteraard dat het om een dichtheid gaat,
• Bepaal de kans dat speler A het tornooi wint.

■ Ω = verzameling van tornooien (volgens die regels)
$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$ = aantal partijen in tornooi ω

mogelijke uitkomsten : $\{1,2,...,10\}$

$p_k = P(X=k) = 0,3^{k-1}$ x $0,7$ met $k = 1,...,9$

kans A wint : 0,4
kans B wint : 0,3
kans geen winnaar : 0,3

$X \sim Geo(p)$
$p(k) = q^{k-1}p$

$p_{10} = P(X=10) = 0,3^{10-1}$

Is dit een dichtheid?

1. $p_k \geq 0$: OK
2. $p_1 + ... + p_{10} = 0,7(1 + 0,3 + ... + 0,3^8) + 0,3^9$
$= 0,7\frac{1-0,3^9}{1-0,3} + 0,3^9$
=1 : OK

■ p = kans dat A wint
$= 0,4 + (0,3x0,4) + ... + (0,3^9 x 0,4)$
$= 0,4(1 + 0,3 + ... + 0,3^9)$
$= 0,4\left(\frac{1-0,3^{10}}{1-0,3}\right)$
$= \frac{0,4}{0,7}(1-0,3^{10}) \cong \frac{4}{7}$
$= 0,5714$

3. Stefan Banach is een verwoed pijproker. Hij beschikt dan ook permanent over twee doosjes lucifers, één in elk van zijn twee broekzakken. Telkens hij zijn pijp opsteekt, kiest hij lukraak een van zijn broekzakken om er een lucifer te zoeken. Op een keer merkt hij dat het (gekozen) doosje leeg is. Bepaal de kans dat het andere doosje ook leeg is. Veronderstel dat de twee luciferdoosjes initieel evenveel lucifers bevatten. Bereken deze kans indien elk doosje initieel tien lucifers bevat (uitkomst: 17.7%).

■ Er zijn 2 mogelijke uitkomsten dus gebruiken we Bernouilli
(links lucifer pakken of rechts lucifer pakken)

$X_i =$ 0 als Banach links kiest
1 als Banach rechts kiest

$X_i \sim b(1,\frac{1}{2})$: 1 herhaling en de kans is lukraak dus 0,5

$p = P($2 doosjes TERGELIJKERTIJD leeg$)$
$= P($*n* keer rechts en *n* keer links$)$

n = het aantal lucifers in ieder doosje

$X_1 + ... + X_{2n} \sim b(2n, \frac{1}{2})$: nu 2n herhalingen voor dat beide doosjes leeg zijn

Bernouilli dichtheid: $p(k) = \binom{n}{k} p^k q^{n-k} = \binom{2n}{n} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^{2n-n}$

$$p = P(X_1 + ... + X_{2n} = n) = \binom{2n}{n} \left(\frac{1}{2}\right)^{2n}$$

■ Voor doosjes met 10 lucifers: (n=10):

$$p = P(X_1 + ... + X_{20} = 10) = \binom{20}{10} \left(\frac{1}{2}\right)^{20} = 0{,}177$$

4. De promotor van een sportgebeurtenis kan moeilijk beslissen of hij al dan niet een verzekering tegen regen zal nemen. Indien het niet regent zal hij 2000 euro winnen, tegenover 200 euro indien het regent. De verzekering betaalt 2 000 euro indien het regent, maar de premie bedraagt 500 euro.
Bepaal de kritische waarde voor p (= de kans op mooi weer), boven dewelke een verzekering niet langer interessant is.

Beslissen onder onzekerheid. De volgende tabel geeft de nettowinsten weer. De verzekering kost 500 euro, maar betaalt indien het regent een premie van 2 000 euro.

	p mooi	1-p regen
Verzekering	2000 - 500	200 – 500 + 2000
Niet verzekering	2000	200

$$E(\text{winst} \mid \text{verzekering}) = 1500p + 1700(1-p)$$
$$= 1700 - 200p$$

$$E(\text{winst} \mid \text{geen verzekering}) = 2000p + 200(1-p)$$
$$= 200 + 1800p$$

Kritische waarde van p waar $E(\text{winst} \mid \text{verzekering}) = E(\text{winst} \mid \text{geen verzekering})$

$1700 - 200p = 200 + 1800p$

$2000p = 1500$

$p^* = \frac{15}{20} = \frac{3}{4}$

Antwoord: Indien $p > p^*$ dan niet verzekeren.
Indien $p = p^*$ dan ben je indifferent.
Indien $p < p^*$ dan wel verzekeren.

5. Een aannemer wil een bod doen op een project. Hij schat zijn materiaalkosten op 15 000 USD en zijn arbeidskosten op 600 USD per dag. Verder denkt hij dat het aantal benodigde dagen de volgende kansverdeling volgt:

Benodigde dagen	10	11	12	13	14
Kans	0,1	0,3	0,3	0,2	0,1

Bereken zijn verwachte kosten en de variantie hierop.

De ondernemer heeft een 'zekere' kost van 15 000 + 6 000 USD (de eerste term verwijst naar de materiaalkost, de twee term geeft weer dat er minimaal 10 werkdagen nodig zijn). De 'onzekere' kost heeft betrekking tot de extra werkdagen en varieert van 0 tot 4 × 600.
Zij X de stochast die het aantal werkdagen boven de 10 weergeeft (X neemt waarden aan uit {0, 1, 2, 3, 4}). De stochast Y (=totale kosten) wordt gegeven door:

$$Y = 21\,000 + 600 \times X,$$

de eerste term is de 'zekere' kost, de tweede term is 600 keer het aantal extra dagen.
Om de verwachte kost te berekenen, doen we beroep op de lineariteit van de verwachtingsoperator:

$$E[Y] = 21\,000 + 600E[X].$$

De verwachte waarde van X bedraagt

$$E[X] = 0.1 \times 0 + 0.3 \times 1 + 0.3 \times 2 + 0.2 \times 3 + 0.1 \times 4 = 1.90.$$

Bijgevolg: $E[Y] = 21\,000 + 1.9 \times 600 = 22\,140$ USD.

De variantie van de stochast Y berekenen we als volgt:

$$\text{var}(Y) = \text{var}(\alpha X + \beta) = \alpha^2 \,\text{var}(X) = 600^2 \,\text{var}(X)$$

We bepalen het tweede ruwe moment van X:

$$E[X^2] = (0.1 \times 0^2) + (0.3 \times 1^2) + (0.3 \times 2^2) + (0.2 \times 3^2) + (0.1 \times 4^2) = 4.90$$

(hoofdrekenen!). De variantie van X is dan gelijk aan $4.9 - (1.9)_2 = 1.29$.
Bijgevolg, $\text{var}(Y) = 600^2 \times 1.29 = 464\,400 = 681.47^2$

Besluit: de stochast Y (=totale kosten) heeft een verwachte waarde gelijk aan 22 140 USD met een standaardafwijking (spreiding) van 681.47. Deze twee cijfers geven een 'eerste' indruk omtrent de stochast Y.

6. Geef een verantwoording voor het verbetersysteem van multiple-choice examens : Indien er 4 antwoordmogelijkheden zijn per vraag, dan wordt 1 punt toegekend per goed antwoord en 1/3 punt afgetrokken per fout antwoord.
Stel dat er 5 mogelijke antwoorden zijn en dat 1 punt toegekend wordt per goed antwoord. Hoeveel moet er nu afgetrokken worden per fout antwoord?

■ Veronderstel dat een goed antwoord 1 punt toebedeeld krijgt en een fout antwoord f punten. Indien de student lukraak gokt, dan is zijn verwachte score per vraag gelijk aan:

$$E_4 = \frac{1}{4} \text{ x } 1 + \frac{3}{4} \text{ x } f$$

(het subscript 4 verwijst naar de 4 antwoordmogelijkheden). We wensen dat de verwachte score van een lukrake gok gelijk is aan nul. Bijgevolg moet f gelijk zijn aan $-1/3$.

■ In geval van n antwoordmogelijkheden (waarvan precies 1 correct) wordt de vorige uitdrukking:

$$E_n = \frac{1}{n} \text{ x } 1 + \frac{n-1}{n} \text{ x } f$$

De verwachte score is gelijk aan nul indien $f = -1/(n-1)$

Bij 5 mogelijke antwoorden moet er dus $-1/4$ worden afgetrokken.

7. Op een tafel liggen twee omslagen. Beide omslagen bevatten een bedrag in euro. Het bedrag in één van de omslagen is precies het dubbele van het bedrag in de andere. Twee spelers zitten aan tafel. Eén van hen kiest lukraak een van de omslagen, de andere omslag is voor de tegenspeler. Vooraleer de omslagen geopend worden, vraagt de spelleider aan de tegenspeler (die we Eddy noemen) of hij de omslagen wil verwisselen. Eddy redeneert als volgt. Noteer met b het bedrag in mijn omslag. Indien ik de omslagen verwissel, dan mag ik een bedrag verwachten gelijk aan
Immers, met kans 1/2 bevat de andere omslag de helft van het bedrag in de mijne, en met kans 1/2 bevat de andere omslag het dubbele van het bedrag in de mijne. Vermits 1.25 b groter is dan b, moet ik ruilen. Eddy verwisselt de omslagen.
Waar zit de fout in Eddy's redenering ?

Het ware beter geweest indien Eddy de bedragen B en 2B had genoemd, in plaats van 'zijn' bedrag als referentie te gebruiken (de bedragen 2b en b/2 zijn alleszins fout). In deze nieuwe notatie is het verwachte bedrag voor beide spelers gelijk aan =

$$\frac{1}{2} \text{ x B} + \frac{1}{2} \text{ x 2B} = \frac{3}{2} \text{ x B}$$

Indien X het bedrag in zijn eerste omslag is.
Wanneer hij een omslag kiest is $E(X) = 1,5B$
Indien hij wisselt in Y (het bedrag in de 'tweede' omslag) dan is $E(Y) = E(X) = 1,5B$
→ wisselen heeft geen effect op de verwachte waarde.

Merk op dat indien Eddy's redenering juist was, zich het volgende 'fantastische' scenario voordoet. Eddy ruilt en verhoogt zijn verwachte winst (vermenigvuldigen met 1.25).
De tegenspeler wil dan ook opnieuw ruilen om een hogere winst te hebben.
Dan opnieuw, wil Eddy ruilen, ...
Na n ruiloperaties neemt het bedrag de waarde $1,25^n$ x b aan. In de limiet is dit oneindig.

8. De dichtheid p van een continue stochast X is als volgt:

$$p : \mathbb{R} \to \mathbb{R} : x \to p(x) = \begin{cases} ax^2 & \text{voor } 0 \le x \le 1 \\ 0 & \text{elders} \end{cases}$$

- Toon aan dat a = 3,
- bereken de verwachte waarde van X,
- bereken de standaardafwijking van X,
- bereken de kans dat X een waarde aanneemt tussen 0 en 1/2,
- bereken de mediaanwaarde.

■ We maken gebruik van de twee voorwaarden voor een dichtheid. Opdat p een dichtheidsfunctie zou zijn, moet $p \ge 0$ en $\int p = 1$. De eerste voorwaarde is voldaan zodra $a \ge 0$. De tweede voorwaarde luid:

$$\int p = \int_0^1 ax^2 \text{ dx} = \frac{a}{3}x^3\Big|_0^1 = \frac{a}{3}1^3 - \frac{a}{3}0^3 = \frac{a}{3} = 1$$

De laatste gelijkheid impliceert dat a = 3.

■ We gebruiken de formule bovenaan op pagina 3 van het formularium:

$E[g(X)] = \int_{-\infty}^{+\infty} g(x)p(x) \text{ dx}$ omdat we weten dat X continu is.

In dit geval weten we dat $g(x) = \text{X}$ en $0 \leq x \leq 1$:

$$E[X] = \int_0^1 xp(x) \text{ dx} = \int_0^1 (x)(3x^2) \text{ dx} = \int_0^1 3x^3 \text{ dx} = \frac{3}{4}x^4\Big|_0^1 = \frac{3}{4}$$

■ We gebruiken de formule van pagina 3 van het formularium:

Variantie: $\sigma^2 = \mu_2' - (\mu_1')^2$

dit is gelijk aan: $\sigma^2 = E(X^2) - \big(E(X)\big)^2$

$$E(X^2) = \int_0^1 3x^4 \text{ dx} = \frac{3}{5}x^5\Big|_0^1 = \frac{3}{5}$$

Bijgevolg: $\sigma^2 = E(X^2) - \big(E(X)\big)^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80}$

De standaardafwijking is dus gelijk aan: $\sigma = \sqrt{\frac{3}{80}}$

■ Omdat X continu is, is de kans de oppervlakte onder de curve tussen 0 en 0,5:

$$P(0 \leq X \leq 0{,}5) = \int_0^{0{,}5} p = \int_0^{0{,}5} 3x^2 \text{ dx} = x^3\Big|_0^{0{,}5} = \frac{1}{8}$$

■ De mediaan het punt dat 50% van de data links laat. Dus de oppervlakte onder de curve links (en rechts) van de mediaan is gelijk aan 0,5.

$$\int_0^m p = 0{,}5 \Rightarrow x^3\Big|_0^m = m^3 = 0{,}5 \Rightarrow m = 0{,}5^{1/3}$$

9. De dichtheid p van een continue stochast X is als volgt:

$$p : \mathbb{R} \to \mathbb{R} : x \to p(x) = \begin{cases} x/a & \text{voor } 0 \leq x \leq 1 \\ 0 & \text{elders} \end{cases}$$

- Toon aan dat a = 2,
- bereken de verwachte waarde van X,
- bereken de standaardafwijking van X,
- bepaal de momentgenererende functie MX.

■ We maken opnieuw gebruik van de twee voorwaarden voor een dichtheid. Opdat p een dichtheidsfunctie zou zijn, moet $p \geq 0$ en $\int p = 1$. De eerste voorwaarde is voldaan zodra $a \geq 0$. De tweede voorwaarde luid:

$$\int p = \int_0^a \frac{x}{a}\,\mathrm{dx} = \frac{1}{2a}x^2\Big|_0^a = \frac{a^2}{2a} = 1$$

De laatste gelijkheid impliceert dat a = 2.

■ We gebruiken opnieuw de formule bovenaan op pagina 3 van het formularium:

$E[g(X)] = \int_{-\infty}^{+\infty} g(x)p(x)\,\mathrm{dx}$ om dat we weten dat X continu is.

In dit geval weten we dat $g(x)$=X en $0 \le x \le 1$:

$$E[X] = \int xp(x)\,\mathrm{dx} = \int_0^2 x^2/2\,\mathrm{dx} = \frac{1}{6}x^3\Big|_0^2 = \frac{4}{3}$$

■ We gebruiken opnieuw de formule van pagina 3 van het formularium:

Variantie: $\sigma^2 = \mu_2' - (\mu_1')^2$

dit is gelijk aan: $\sigma^2 = E(X^2) - \big(E(X)\big)^2$

$$E(X^2) = \int_0^2 x^3/2\,\mathrm{dx} = \frac{1}{8}x^4\Big|_0^2 = 2$$

Bijgevolg: $\sigma^2 = E(X^2) - \big(E(X)\big)^2 = 2 - \frac{16}{9} = \frac{2}{9}$

De standaardafwijking is dus gelijk aan: $\sigma = \sqrt{\frac{2}{9}}$

■ Momentgenererende functie:

$M_X(t) = E(e^{tX})$: in het formularium p3

en $E(e^{tX}) = \int_{-\infty}^{+\infty} g(x)p(x)\,\mathrm{dx}$: ook in het formularium p3

Dus: $M_X(t) = \int e^{xt}p = \frac{1}{2}\int_0^2 e^{xt}x\,\mathrm{dx}$

We passen partiele integratie toe:

$$\int e^{xt}x\,\mathrm{dx} = \frac{1}{t}\int x\,d(e^{tx}) = \frac{1}{t}\left[xe^{tx} - \int e^{tx}\,\mathrm{dx}\right] = \frac{1}{t}\left[xe^{tx} - \frac{1}{t}e^{tx}\right]$$

Aldus,

$$M_X(t) = \frac{1}{2t}\left[xe^{tx} - \frac{1}{t}e^{tx}\right]_{x=0}^{2} = \frac{1}{2t}\left[2e^{2t} - \frac{1}{t}(e^{2t} - 1)\right]$$

10. Van de stochast X is geweten dat $E\left[(X-1)^2\right] = 10$, en dat $E\left[(X-2)^2\right] = 6$. Bereken de variantie van X.

We kunnen beide gegeven vergelijkingen herschrijven:

$E\left[(X-1)^2\right] = 10$ → $E\left[X^2 - 2X + 1\right] = 10$ → $E(X^2) - 2E(X) + 1 = 10$

$E\left[(X-2)^2\right] = 6$ → $E\left[X^2 - 4X + 4\right] = 6$ → $E(X^2) - 4E(X) + 4 = 6$

Nu hebben we een stelsel van twee vergelijkingen in twee onbekende ($E(X^2)$ en $E(X)$) en lossen we op.

We vinden: $E(X)=\frac{7}{2}$ en $E(X^2)=16$

Bijgevolg: $\text{var}(X)=E(X^2)-(E(X))^2=16-\frac{49}{4}=\frac{15}{4}$

11. Het gemiddelde resultaat μ op een examen is 65 punten met een standaardafwijking gelijk aan 10. Een examenkopij wordt lukraak uit de stapel examens getrokken. Geef een benedengrens voor de kans dat deze examenkopij een score heeft tussen 45 en 85.

We maken gebruik van de Stelling van Chebyshev. Zij X de stochast die het resultaat op het examen weergeeft, dan;

$$P\left(|X-\mu|\le k\sigma\right)\ge 1-k^{-2}$$

Stel: $\mu=65$, $\sigma=10$: (dit is gegeven)

en k =2 omdat we tussen 45 en 85 zoeken dus twee maal σ in iedere richting van μ .

Dan: $P\left(|X-65|\le 20\right)\ge 1-\frac{1}{4}$

Met andere woorden: $P(45\le X\le 85)\ge 3/4=75\%$

12. Bespreek de verdeling van de volgende stochasten.
• Van alle piloten zijn er 40% ouder dan 40 jaar. Zij X het aantal piloten ouder dan 40 in een lukraak getrokken steekproef van 15 piloten.
• Een verkoper verkoopt aan 20% van de bezoekers aan zijn winkel. Op een zekere ochtend telt hij het aantal bezoekers dat aan de eerste koper voorafgaat. Zij Y dit aantal.
• In een kamer staan 6 vrouwen en 4 mannen. Drie mensen worden geselecteerd om een comité te vormen. Zij Z het aantal vrouwen in dit comité.

■ $X\sim b(n=15,p=0,4)$
Er zijn twee mogelijke uitkomsten. De steekproef is van lengte 15 met een 40% kans op succes.

■ $Y\sim Geo(p=0,2)$

■ $Z\sim Hyp(n=3,a=6,b=4)$
we trekken n=3 knikkers uit een set van a=6 witte en b=4 zwarte.

13. Een Bernoulli experiment b(1, p) wordt n keer herhaald. Het verwacht aantal successen is gelijk aan 3 met een variantie gelijk aan 2.1. Bepaal de waarde van p.

We hebben een stelsel van twee vergelijkingen in n en p . We lossen dit stelsel op:

$$X\sim b(n,p)$$
$$E(X)=3=np$$
$$\text{var}(X)=2,1=E(X^2)=npq=np(1-p)=np-npp$$

Oplossing: substitueer vergelijking 1 in vergelijking 2
3(1-p)=2,1 of 1-p=0,7 of nog p=0,3
Gebruik dit resultaat in de eerste vergelijking: 0,3n=3 of n=10

14. De vraag naar een bepaald tijdschrift aan een dagbladstand wordt gegeven door volgende dichtheid:

aantal gevraagde exemplaren	0	1	2	3	4	5	6
waarschijnlijkheid	0,1	0,2	0,3	0,1	0,1	0,1	0,1

De verkoopprijs bedraagt 2 USD per tijdschrift, de aankooprijs is 1 USD, en voor elk niet verkocht exemplaar wordt 0.10 USD terugbetaald.
• Bereken de verwachte waarde en de variantie van het gevraagd aantal tijdschriften.
• Stel dat de dagbladverkoper 6 exemplaren inkoopt. Bepaal de verwachte waarde en variantie van de nettowinst.
• Stel dat de dagbladverkoper 3 exemplaren inkoopt. Bepaal de verwachte waarde en variantie van de nettowinst.

■ X = het aantal verkochte dagbladen (discrete stochast)

X	0	1	2	3	4	5	6
P	0,1	0,2	0,3	0,1	0,1	0,1	0,1

1ste moment: $E(X)=0+0,2+0,6+0,3+0,4+0,5+0,6=2,6$

2de moment: $E(X^2)=0+0,2+1,2+0,9+1,6+2,5+3,6=10$

$\text{var}\, X=E(X^2)-\big(E(X)\big)^2=10-2,6^2=3,24$

De verwachte waarde is dus 2,6 met variantie 3,24

■ $E(\text{winst})=\sum \text{winst}.p(x)$

π_6 = winst bij 6 tijdschriften in voorraad
= verkoopprijs + niet verkochte exemplaren + kost van de aankoop
$=2X+0,1(6-X)-6=1,9X-5,4$

$E(\pi_6)=1,9E(X)-5,4=-0,46<0$: verlies

$\text{var}\,\pi_6=1,9^2\ \text{var}\, X=11,7$: formule $\text{var}\,\pi=a^2\ \text{var}\, x$

■ π_3 = winst bij 3 tijdschriften in voorraad
$=2X+0,1(3-X)-3=1,9X-2,7$ **als** $X<3$
=3 **als** $X\geq 3$

$E(\pi_3)=p_0\pi_3\big|_{x=0}+p_1\pi_3\big|_{x=1}+p_2\pi_3\big|_{x=2}+p_{\geq 3}^3$
= (0,1 x 2,7)-(0,2 x 0,8)+(0,3 x 1,1)+(0,4 x 3)
=1,1>0 : winst

$\text{var}\,\pi_3=4,82-1,21=3,61$

$E(\pi_3^2)=0,1(2,7)^2+0,2(0,8)^2+0,3(1,1)^2+0,4(3)^2$
$=4,82$

15. De kans op falen van een motor van een (tweemotorig of viermotorig) vliegtuig is gelijk aan p. Het falen van een motor is onafhankelijk van het al dan niet falen van de andere motoren. Een vlucht is succesvol indien minstens de helft der motoren niet faalt. Is het viermotorig vliegtuig steeds veiliger dan het tweemotorig?

X = het aantal motoren dat faalt

<u>2 motoren:</u>

$X \sim b(2, p)$: twee keer herhaald want er zijn 2 motoren

$P_2 = P(\text{vliegtuig faalt}) = P(X=2) = \binom{2}{2} p^2 = p^2$

<u>4 motoren:</u>

$X \sim b(4, p)$: vier keer herhaald want er zijn 4 motoren

$P_4 = P(\text{vliegtuig faalt}) = P(X \geq 3)$

$= P(X=3) + P(X=4)$

$= \binom{4}{3} p^3(1-p) + \binom{4}{4} p^4$

<u>Vraag:</u> Is de kans dat vliegtuig 2 faalt groter dan de kans dat vliegtuig 4 faalt? $p_2 > p_4$?

$p^2 > 4p^3(1-p) + p^4$

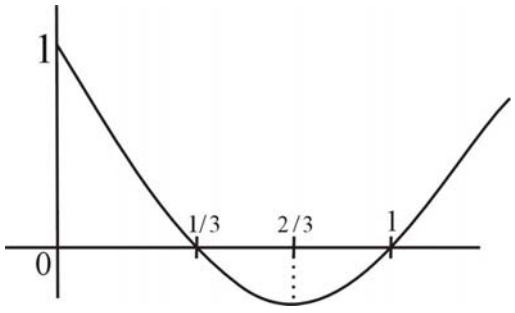
$1 > 4p(1-p) + p^2$

$1 > 4p - 4p^2 + p^2$

$f(p) = 3p^2 - 4p + 1 > 0$

$f'(p) = 6p - 4 = 0$

dus $p = 4/6 = 2/3$



→ de kans dat een 2-motorig vliegtuig faalt is kleiner dan de kans dat een 4-motorig vliegtuig faalt als en slechts als $1/3 < p < 1$.

16. De cumulatieve verdeling F van een continue stochast X is als volgt:

$$F : \mathbb{R} \to \mathbb{R} : x \to F(x) = \begin{cases} 1 - e^{-2x} & \text{voor } 0 \leq x, \\ 0 & \text{elders} \end{cases}$$

• Toon aan dat F inderdaad een cumulatieve verdeling is; m.a.w. verifieer of F voldoet aan alle voorwaarden opgelegd aan een cumulatieve verdeling (zie form p2,onderaan),
• bereken de kans dat X een waarde aanneemt strikt groter dan 2,
• bereken de kans dat X een waarde aanneemt strikt groter dan -3 en kleiner dan 4,
• bereken de mediaanwaarde,
• geef tenslotte het functievoorschrift van de dichtheidsfunctie.

■ De functie F voldoet aan alle voorwaarden. Zij is overal continu (dus ook rechtscontinu), meer nog zij is overal behalve in 0 afleidbaar. Zij is nergens dalend (de afgeleide is overal groter of gelijk aan nul).
De functie heeft het juiste asymptotische gedrag: F = 0 in $-\infty$ en F = 1 in $+\infty$.

■ $P(X>2)=1-F(2)=e^{-4}$

■ $P(-3<X\leq 4)=F(4)-F(-3)=1-e^{-8}$

■ m voldoet aan $F(m)=1/2$, waaruit $m=(\ln 2)/2$

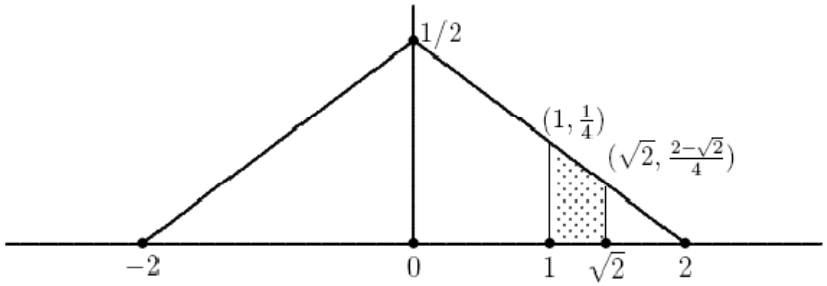
■ $p=F'$, bijgevolg $p(x)=2e^{-2x}$ voor $x\geq 0$ en $p(x)=0$ voor $x<0$

17. De dichtheid p van een continue stochast X is als volgt:

$$p:\mathbb{R}\to\mathbb{R}:x\to p(x)=\begin{cases}(2+x)/4 \text{ voor } -2\leq x\leq 0\\ (2-x)/4 \text{ voor } 0\leq x\leq 2\\ 0 \text{ elders}\end{cases}$$

• Maak een figuur van deze dichtheidsfunctie,
• bereken de verwachte waarde van X,
• bereken de kans dat X^2 een waarde aanneemt tussen 1 en 2.

■ dichtheid enkel op het interval $[-2,2]$ daarbuiten is het toch 0 :



■ De verwachte waarde is gelijk aan 0, vanwege symmetrie rond 0 (vgl met standaard normale verdeling: $E(X)=0$

■ Hier moet je de vraagstelling eerst in termen van X schrijven.
$P(1\leq X^2\leq 2)=P(-\sqrt{2}\leq X\leq -1)+P(1\leq X\leq\sqrt{2})=2P(1\leq X\leq\sqrt{2})$
Gebruik de optelregel en de symmetrie.
De kans $\alpha=P(1\leq X\leq\sqrt{2})$ is gelijk aan de gearceerde oppervlakte in de figuur en kan gemakkelijk berekend worden (je kan hier integratie omzeilen).

$$\alpha=\frac{1}{2}\text{ x }(\sqrt{2}-1)\text{ x }(\frac{1}{4}+\frac{2-\sqrt{2}}{4})=\frac{4\sqrt{2}-5}{8}$$

De gezochte kans is dan gelijk aan $2\alpha=\sqrt{2}-1,25$

18. Een militieplichtige wenst nadere informatie omtrent zijn vermoedelijke oproepingsdatum en neemt daarvoor telefonisch contact op met de bevoegde dienst. De telefoonlijn is evenwel in 95% van de gevallen bezet.
• Indien hij zes keer telefoneert, hoe groot is de kans dat hij minstens een keer iemand aan de lijn krijgt?
• Indien hij zes keer telefoneert, hoe groot is de kans dat hij precies een keer iemand aan de lijn krijgt?
• Wat is de kans dat hij pas bij zijn zesde poging iemand aan de lijn krijgt?

In de plaats van de kans dat hij minstens een keer iemand aan de lijn krijgt te berekenen zoeken we de kans op 6 mislukkingen te hebben:
De kans hier van is gelijk aan $0,95^6$. Het compliment hiervan geeft de gevraagde kans:
$1-0,95^6=0,265$

■ De kans om precies één keer antwoord te krijgen bij zes pogingen is gelijk aan:
$(6)(0,95^5)(0,05)=0,232$

■ De kans op vijf achtereenvolgende mislukkingen en vervolgens succes is gelijk aan:
$(0,95^5)(0,05)=0,039$

19. De exponentiele verdeling wordt vaak gebruikt om wachttijden te modelleren.
• Toon aan: zij X exponentieel verdeeld, zij t en s in $\mathbb{R}^+$, dan geldt:

$$P(X \geq s+t|X \geq s)=P(X \geq t)$$

We zeggen dat de exponentiele verdeling aan geheugenverlies lijdt.
• Is de exponentiele verdeling een goed model om 'de werkloosheidsduur van een werkzoekende' te beschrijven?

We bekijken nu een stochast Y die voldoet aan $Y^\beta \sim Exp(\lambda)$.
• Voor welke waarden van β geldt:

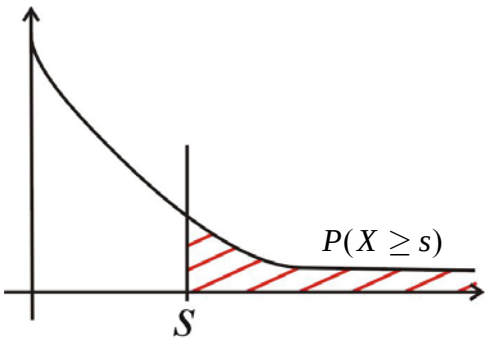
$$P(Y \geq s+t|Y \geq s) > P(Y \geq t)\,?$$

De verdeling van de stochast Y waarvoor de β de macht exponentieel verdeeld is, met $\beta>0$ noemt men de Weibull-verdeling. Merk op, indien β = 1, dan is Y exponentieel verdeeld.

■ <u>Bewijs</u> dat het model geen geheugen heeft.
$P(X \geq s+t|x \geq s)$
<u>vb.</u> wat is de kans dat we 10 dagen moeten wachten als we al 7 hebben gewacht

$$=\frac{\int_{s+t}^{\infty} \lambda e^{-\lambda x}\ \mathrm{dx}}{\int_{s}^{\infty} \lambda e^{-\lambda x}\ \mathrm{dx}}=\frac{\left[-e^{-\lambda x}\right]_{s+t}^{\infty}}{\left[-e^{-\lambda x}\right]_{s}^{\infty}}$$

$$=\frac{e^{-\lambda(s+t)}}{e^{-\lambda s}}=e^{-\lambda t}=P(X \geq t)$$



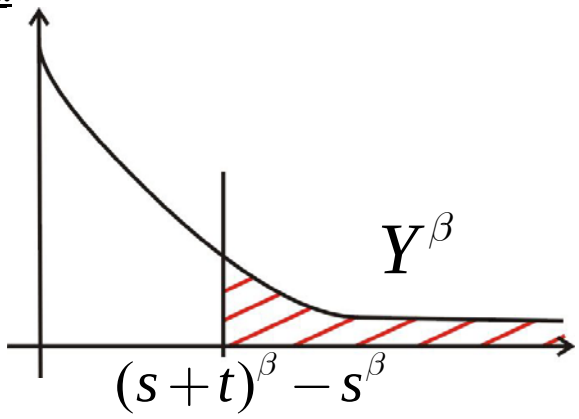
→ dus $P(X \geq s+t|x \geq s)=P(X \geq t)$
→ de exponentiele verdeling heeft geen geheugen

■ Neen, iemand die reeds lange tijd werkloos is heeft meer kans om werkloos te blijven.

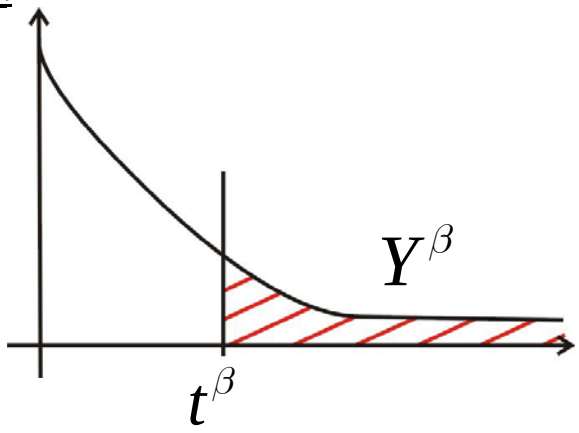
■ $P(Y \geq s+t|Y \geq s)$
$= P(Y^\beta \geq (s+t)^\beta|Y^\beta \geq s^\beta)$
$= P(Y^\beta \geq (s+t)^\beta - s^\beta)$

Je weet dat $P(Y^\beta \geq (s+t)^\beta - s^\beta)$ groter is dan $P(Y^\beta \geq t^\beta)=P(Y \geq t)$

1.



2.



Vraag: is 1>2?
als en slechts als:

$(s+t)^\beta - s^\beta \leq t^\beta$

$(s+t)^\beta \leq s^\beta + t^\beta$

$\boxed{0 < \beta < 1}$

→ nu kan je werkloosheid modelleren

20. Op een nacht is een taxi betrokken bij een verkeersongeval. De taxi-chauffeur pleegt vluchtmisdrijf. Een getuige echter verklaard dat de taxi groen was. Dit is belangrijke informatie: in deze stad zijn er precies drie taxibedrijven met elk een specifieke kleur van wagens (met name, groen, rood, en blauw). De manager van Taxi-Groen betwist de verklaring van de getuige en vraagt een onderzoek naar het gezichtsvermogen van de getuige. Uit de test die het gerechtshof laat uitvoeren, blijkt dat de getuige elke kleur in 80% van de gevallen juist identificeert. Verder, werd rood in 5% van de gevallen als blauw herkend (en bijgevolg, in 15% van de gevallen als groen). Blauw werd in 8% van de gevallen als rood herkend (dus, in 12% van de gevallen als groen). Groen, tenslotte, werd in 10% van de gevallen als blauw herkend. Bepaal de kans dat de taxi, gezien door de getuige, inderdaad groen was. Maak hierbij gebruik van het feit dat het marktaandeel van Taxi-Rood 60% bedraagt, van Taxi-Groen 30%, en van Taxi-Blauw 10%.

We zoeken dus de kans dat de taxi effectief groen was gegeven dat de getuigen verklaart dat de taxi groen was: (een kansboom tekenen kan ze zaak vergemakkelijken)

$$P(\text{groen} \mid \text{getuigen verklaart groen}) = \frac{P(\text{getuigen verklaart groen} \mid \text{groen}).P(\text{groen})}{P(\text{getuigen verklaart groen})}$$
$$= \frac{(0,30 \text{ x } 0,80)}{(0,60 \text{ x } 0,15) + (0,30 \text{ x } 0,80) + (0,1 \text{ x } 0,12)}$$
$$= 0,70$$

Dus met 70% kans heeft hij groen gezien.

Oefenzitting 4 – Dichtheden, kansmodellen, vectoren van lengte 2,….

1. De telefoonoproepen die in een telefooncentrale aankomen gedurende één minuut volgen een Poisson-verdeling met parameter $\lambda = 8$. Wat is de kans dat in één minuut:
- geen oproepen aankomen,
- juist 8 oproepen aankomen,
- minstens 8 oproepen aankomen.

Zij X de stochast die het aantal telefoonoproepen per minuut weergeeft, dan geldt:
$X \sim Pois(8)$

We vinden in het formularium de dichtheid: $p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$

■ Voor de kans dat er geen oproepen aankomen is k = 0:

$$p_0 = P(X = 0) = \frac{8^0}{0!} e^{-8} = 0,000335$$

■ Voor de kans dat er juist 8 oproepen aankomen is k = 8:

$$p_8 = P(X = 8) = \frac{8^8}{8!} e^{-8} = 0,139587$$

■ De kans dat er minstens 8 oproepen aankomen:

$$P(X \geq 8) = p_8 + p_9 + ... = 1 - P(X \leq 7) = 1 - (p_0 + p_1 + ... + p_7) = 0,547036$$

Alternatief: $X \approx N(8,8)$, waaruit $\frac{X-8}{\sqrt{8}} \approx N(0,1)$

$$P(X \geq 7,5) = P\left(\frac{X-8}{\sqrt{8}} \geq \frac{7,5-8}{\sqrt{8}} = -0,176\right) = 0,5673$$

Merk op dat de 8 vervangen werd door 7,5 (continuiteitscorrectie)

De benadering van de Poisson-verdeling door de normale verdeling (zie cursustekst pagina 189, eigenschap 10.4.1 (2), zie ook formularium pagina 4, lijn -5) geeft hier een duidelijke onderschatting (van 2%). De waarde $\lambda = 8$ is nog te klein om tot op twee decimalen nauwkeurig te werken.

2. De verkoop van uurwerken bij een bepaalde juwelier verloopt volgens een Poissonproces met een gemiddelde van vijf uurwerken per werkdag (veronderstel vijf werkdagen per week). Elke maandag wordt de voorraad van de juwelier aangevuld tot 25 uurwerken.
- Hoe groot is het risico dat de juwelier uit voorraad loopt.
- Hoe groot moet de voorraad zijn indien hij dit risico wil beperken tot 10%.

Zij X_i de stochast die het aantal verkochte uurwerken op dag i weergeeft, met $i = 1, 2, 3, 4, 5$ (1=maandag…). Dan geldt dat X_i Poissonverdeeld is met $\lambda = 5$.

De verkoop per week wordt voorgesteld door de stochast $Y = X_1 + X_2 + ... + X_5$.

In de veronderstelling dat X_i onderling onafhankelijk zijn, volgt mbv de momentgenererende functies dat Y Poissonverdeeld is met $\lambda = 25$.
Bijgevolg geldt $Y \approx N(25,25)$

■ We berekenen $P(Y > 25)$ via deze normale benadering;
Maar eerst moeten we een continuiteitscorrectie doen omdat we een discrete stochast (de Poisson) benaderen aan de hand van een continue stochast (de normale).
We maken een correctie van 0,5 en gebruiken dus 25,5 ipv 25.

$$P(Y \geq 25{,}5) = P\left(\frac{Y-\mu}{\sqrt{\sigma^2}} \geq \frac{25{,}5-\mu}{\sqrt{\sigma^2}}\right) = P\left(\frac{Y-25}{5} \geq \frac{0{,}5}{5} = 0{,}1\right) = \underline{0{,}4602}$$

■ Noteer met v de nodige voorraad. Los v op uit de vergelijking $P(Y \geq v + 0{,}5) = 0{,}1$
Herschaal naar de standaardnormale verdeling:

$$P\left(\frac{Y-25}{5} \geq \frac{v+0{,}5-25}{5}\right) = 0{,}1$$

Waaruit $\frac{v-24{,}5}{5} = 1{,}28$ of $v = 31$ (30,9)

Je kan deze oefening ook oplossen <u>zonder</u> gebruik te maken van de normale verdeling. Je moet dan wel een hele trein van Poisson-kansen $p_k = e^{-25}25^k / k!$ berekenen (en optellen); voor de eerste deelvraag heb je $p_0, p_1, ..., p_{24}$ nodig; voor de tweede deelvraag $p_0, ..., p_{31}$ (en dan vaststellen dat 31 als voorraad voldoende is). Dit doe je best via Excel waar het volstaat om de 'formule' voor p_0 in te typen en te copieren naar grotere waarden voor k.

3. Zij $X_1 \sim b(n, p)$ en $X_2 \sim b(m, p)$. Zij X_1 en X_2 onafhankelijk.
Bepaal $P(X_1 = x | X_1 + X_2 = z)$. Welk type verdeling is dit?

<u>vb.</u> Het tossen van één munt door twee personen. De stochastische variabele X telt het aantal successen bij persoon 1, en de stochastische variabele Y telt het aantal successen bij persoon 2. De stochastische variabelen X en Y zijn beide binominaal verdeeld. De vraag luidt: gegeven het totaal van z successen, bepaal de kans dar er x successen geteld werden bij persoon 1.

Indien de twee binomiaal experimenten tesamen z successen opleveren, dan is de kans dat x successen (van de z) werden voortgebracht door het eerste binomiaal experiment gegeven door:

$$P(X_1 = x | X_1 + X_2 = z) = \binom{n}{x}\binom{m}{zx} \Big/ \binom{n+m}{z}$$

→ dit is precies de hypergeometrische verdeling (met (n,a,b)=(z,n,m))

4. Het gemiddeld aantal noodoproepen tussen 18u en 22u 's avonds in een ziekenhuis bedraagt 5,2.
- Wat is de kans dat de eerste noodoproep ontvangen wordt tussen 18u15 en 18u45?
- Bereken ook de mediaantijd voor de eerste oproep.

■ We gebruiken de Poisson verdeling voor het aantal oproepen:
X = het aantal oproepen per 4 uur (22-18)
$X \sim Pois(5,2)$

Y = wachttijd tot het eerste oproep
$Y \sim Exp(5,2)$ indien 4 uren

$Y \sim Exp\left(\frac{5,2}{4} = 1,3\right)$ indien 1 uur

Kans dat het eerste tussen 18u15 en 18u45 ontvangen wordt:

$$P = P(0,25 \leq Y \leq 0,75)$$
$$= \int_{0,25}^{0,75} \lambda e^{-\lambda x} \text{ met } \lambda = 1,3$$
$$= \left[e^{-\lambda x}\right]_{0,25}^{0,75} = -e^{-3\lambda/4} + e^{-\lambda/4} = 0,345$$

→ de kans dat er een noodoproep ontvangen wordt tussen 18u15 en 18u45 bedraagt 35%

■ De mediaan van $Y \sim Pois(\lambda = 1,3)$
De kans dat de stochast Y waarden aanneemt kleiner dan de mediaan is 50%:

$$P(Y \leq m) = \frac{1}{2}$$
$$\int_0^m \lambda e^{-\lambda x} = \frac{1}{2}$$
$$= \left[-e^{-\lambda x}\right]_0^m$$
$$= 1 - e^{-\lambda m}$$
$$= e^{-\lambda m} = \frac{1}{2}$$
$$= -\lambda m = \ln\left(\frac{1}{2}\right)$$
$$= \lambda m = \ln 2$$
$$m = \frac{\ln 2}{\lambda}$$
$$\underline{m = 0,533} \text{ : ongeveer 32 minuten}$$

5. Een eerlijk muntstuk wordt meerdere malen opgeworpen. Zij X het aantal worpen tot de eerste keer kruis. Dus X neemt de waarde 11 aan, indien je eerst 10 keer munt gooit en de 11de keer kruis.
Iemand biedt je tegen betaling van een vergoeding het volgende spel aan: indien X = k ontvang je 2^k euro.
- Wat is je verwachte opbrengst ?
- Hoe laag moet de vergoeding zijn opdat je bereid zou zijn mee te spelen ?

(Sint-Petersburg paradox)

■ De maximale vergoeding die een deelnemer wil betalen = verwachte opbrengst
Y = opbrengst
Wat is $E(Y)$?

$P(Y=2^k)=p_k=2^{-k}$ (eerlijke munt) : discrete stochast
(als je 2 keer smijt heb je 0,5 kans $1/2^k$ enz)

$$E(Y)=p_1 2^1+p_2 2^2+...+p_k 2^k$$
$$=1+1+1+...+1$$
$$=\infty$$

→ "oneindig" grote inzet?
Hoe omzeil je dit probleem?

Methode 1:
De maximale vergoeding = verwachte nut
$U=\mathbb{R}^+\to\mathbb{R}:x\to\ln x$

$$E\big(U(Y)\big)=p_1\ln 2+p_2\ln 4+...+p_k\ln 2^k+...$$
$$=\sum_1^\infty\frac{\ln(2^i)}{2^i}$$
$$=\sum_1^\infty\frac{i}{2^i}\ln(2)$$
$$=2\ln 2 \text{ (eindig nut)}$$

Methode 2:
De maximale winst naar boven begrenzen
vb. winst $\le 2^{60}$

$$E(Y)=p_1 2^1+...+p_{59}2^{59}+(p_{60}+p_{61}+...)2^{60}$$
$$=59+\left(1+\frac{1}{2}+\frac{1}{4}+...\right)$$
$$=59+2=61 \text{ (het bedrag is nu ook eindig)}$$

6. De gemiddelde grootte van de miliciens in 1993 bleek 1m75 te zijn met een standaarddeviatie van 10cm.
- Wat is de kans dat een milicien groter is dan 1m85?
- Wat is de kans dat een milicien kleiner is dan 1m70?
- Wat is de kans dat de gestalte van een milicien ligt tussen 1m60 en 1m80?

Gemiddelde = $\mu=175$
Standaard deviatie = $\sigma=10$
X = de grootte van de miliciens in cm in 1993
$X\sim N(\mu,\sigma^2)$ dus $X\sim N(175,100)$

■ $P(X>185)$?
$$P\left(\frac{X-\mu}{\sigma}>\frac{185-\mu}{\sigma}\right)$$

$$P\left(\frac{X-175}{10}>\frac{185-175}{10}\right)=P\left(\frac{X-175}{10}>1\right)$$
$$=P(Z\geq 1)\ :\text{tabel}$$
$$=0{,}1587=15\%$$

■ $P(X<170)$?

$$P\left(\frac{X-\mu}{\sigma}<\frac{170-\mu}{\sigma}\right)$$
$$P\left(\frac{X-175}{10}<\frac{170-175}{10}\right)=P\left(Z<0{,}5\right)$$

Maar de tabel genereert enkel bovenstaartkansen, dus:

$$P\left(Z>0{,}5\right)=0{,}3085=30\%$$

■ $P(160<X<180)$?

$$P\left(\frac{160-\mu}{\sigma}<\frac{X-\mu}{\sigma}<\frac{180-\mu}{\sigma}\right)$$
$$=P\left(\frac{160-175}{10}<Z<\frac{180-175}{10}\right)$$
$$=P(-1{,}5<Z<0{,}5)$$
$$=1-0{,}3035-0{,}0668$$
$$=0{,}6247=62\%$$

7. De gemiddelde leeftijd van een CRT-buis voor TV's bedraagt 1200 uren met een standaarddeviatie van 200 uren. Hoeveel bedraagt het 0.8 kwantiel m.a.w. bepaal de leeftijd die 80% van de buizen zou moeten bereiken?

X = de levensduur van een CRT-buis (in uren)
Levensduur is normaal verdeeld:

$$X\sim N(1200,\sigma^2=200^2)$$

We zoeken de waarde van x die voldoet aan:

$$P(X\geq x)=P\left(\frac{X-\mu}{\sigma}\geq\frac{x-\mu}{\sigma}\right)=P\left(\frac{X-1200}{200}\geq\frac{x-1200}{200}\right)=0{,}84$$
$$P\left(Z<\frac{x-1200}{200}\right)=-0{,}84$$
$$\frac{x-1200}{200}=-0{,}84$$
$$x=1032\,\text{uuren}$$

→80% van de buizen bereikt een leefttijd van 1032 uuren

8. In het Groothertogdom Luxemburg is het maandinkomen normaal verdeeld met gemiddelde 100.000 BEF en een standaarddeviatie van 10.000 BEF. Het totaal aantal inkomenstrekkers bedraagt 200.000.
• Bereken het aantal inkomenstrekkers dat een inkomen ontvangt kleiner dan 85.000 BEF of groter dan 115.000 BEF.
• Binnen welke inkomensgrenzen (symmetrisch rond het gemiddelde) ligt 50 percent van de inkomenstrekkers?

$X \sim N(100000, \sigma = 10000)$

■ $P(X < 85000) = P(X > 115000)$: het is symmetrisch

$$P\left(\frac{X-\mu}{\sigma} > \frac{115000-100000}{10000}\right) = P\left(Z > \frac{15000}{10000}\right) = 0,0668$$

→6,68% meer dan 115000 en 6,68% minder dan 85000
= 26720 personen.

■ Symmetrisch rond het gemiddelde dus we kunnen aan één kant werken:
$P(X > x) = 0,25$

$$P\left(\frac{X-\mu}{\sigma} > \frac{x-100000}{10000}\right) = P\left(Z > \frac{x-100000}{10000}\right) = 0,67$$

Inkomensgrenzen : 100000 ± 6740 BEF

9. Het gewicht van personen die een bepaalde lift gebruiken is normaal verdeeld met een gemiddelde van 70kg en een standaarddeviatie van 10kg. De lift kan maximaal 4 personen vervoeren, die samen niet meer dan 320kg mogen wegen. Het aantal personen door de lift per rit vervoerd is uniform verdeeld met een minimum van 1 en een maximum van 4.
Bereken het verwachte aantal overbelaste ritten per 100 ritten van de lift.

X_i = gewicht van individu i

$X_i \sim N(70, \sigma^2 = 100)$

$\frac{1}{4} = p_1 = p_2 = p_3 = p_4$ = kans dat i personen de lift nemen (uniform verdeeld)

P(lift is overbelast)=

$= p_1.P(X_1 \geq 320) + p_2.P(X_1 + X_2 \geq 320) + p_3.P(X_1 + X_2 + X_3 \geq 320) + p_4.P(X_1 + X_2 + X_3 + X_4 \geq 320)$

: we veronderstellen dat het gewicht van de personen onafhankelijk van elkaar is

$Y = X_1 + X_2 + X_3 + X_4 \sim N(280, \sigma^2 = 400)$

$$P(Y \geq 320) = P\left(\frac{Y-280}{20} \geq \frac{320-280}{20}\right) = P(Z \geq 2) \quad (\text{met } Z \sim N(0,1))$$

$P(Z \geq 2) = 0,023$: tabel

$P(\text{lift is overbelast}) = \frac{1}{4}(0,023) = \frac{1}{174}$

: het is alleen mogelijk om overbelast de zijn met 4 personen, de kans om overbelast te zijn met 3 personen is zo klein dat wij er geen aandacht aan besteden

10. Romeo en Julia spreken 's ochtends af om in de Alma te gaan eten ergens tussen 12u en 13u. Ze vergeten echter een exact tijdstip te zetten (en ze bezitten geen van beiden een GSM). Onderstel dat ze lukraak arriveren tussen twaalf en één en dat ze maximaal tien minuten wachten bij de ingang van de Alma. Bepaal de kans dat ze elkaar ontmoeten aan de ingang van de Alma.

$X_R =$ het tijdstip Romeo komt toe bij Alma

$X_J =$ het tijdstip Julia komt toe bij Alma

X_R en X_J zijn onafhankelijk

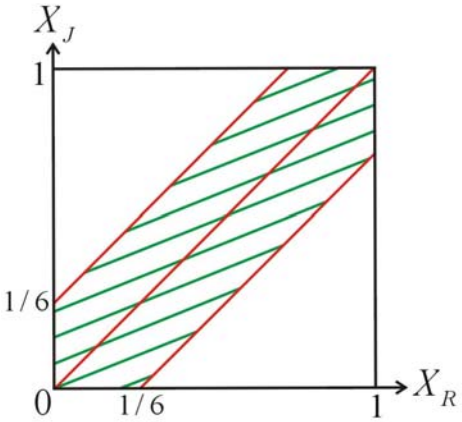
$X_R, X_J \sim U[0,1]$

0 = 12:00 uur

1 = 13:00 uur

$P(\text{Romeo + Julia ontmoeten elkaar}) = P\left(\left|X_R - X_J\right| < \frac{1}{6}\right)$

= de kans dat het verschil van aankomst Romeo en Julia minder dan 10 minuten (1/6) is.



= groene oppervlakte

$= 1 - \left(\frac{5}{6}\right).\left(\frac{5}{6}\right)$

$= \frac{11}{36}\ (\approx 30,6)$

11. Een bedrijf produceert elektrische toestellen voor thuisgebruik. De levensduur van deze toestellen is exponentieel verdeeld met een gemiddelde van 3 jaar. Het bedrijf geeft één jaar garantie op de toestellen.
Wat is het verwacht percentage toestellen dat vervangen moet worden op kosten van de onderneming?

$X : \Omega \to \mathbb{R}$

$\omega \to X(\omega)$: levensduur van toestel ω in jaren

$X \sim Exp(\lambda)$

Bij de exponentielde verdeling is : $\mu = 1/\lambda$ dus $\lambda = 1/\mu$

$X \sim Exp\left(\frac{1}{3}\right)$

Men geeft een garantie als het binnen het jaar kapot is dus als X =1

$$P(X<1) = \int_0^1 \lambda e^{-\lambda x} = \int_0^1 \frac{1}{3} e^{-x/3}$$
$$= \frac{1}{3}\int_0^1 e^{-x/3} = \frac{1}{3}\left[\frac{e^{-x/3}}{-1/3}\right]_0^1$$
$$= -1\left(e^{-\frac{1}{3}} - e^0\right)$$
$$= -e^{-\frac{1}{3}} + 1$$
$$= 0{,}3834 = \underline{38\%}$$

12. Het bedrag dat uitgegeven wordt voor een middagmaal is normaal verdeeld met gemiddelde 220 BEF en standaardafwijking 28 BEF. Tijdens een bepaalde dag zijn er 450 personen die 240 BEF of meer uitgegeven hebben.
Hoeveel personen hebben tijdens die dag een middagmaal genomen?

X = bedrag dat wordt uitgegeven voor een middagmaal

$X \sim N(220, \sigma^2 = 28^2)$

450 personen die 240BEF of meer uitgegeven hebben:

$$P(X \geq 240) = P\left(\frac{X-\mu}{\sigma} \geq \frac{240-220}{28}\right) = P\left(Z \geq \frac{5}{7}\right) = 0{,}239$$

Deze 23,9% correspondeert met een groep van 450 personen.
De 100% correspondeert dan met een groep van 1883 personen.

13. In een bedrijf worden fessen machinaal gevuld. De inhoud van een fles zou 500 cl moeten bedragen. De feitelijke inhouden zijn normaal verdeeld met een standaardafwijking van 5 cl bij een vaste instelling van het vulapparaat. De firma stelt het op prijs dat met een kans van 99% een gevulde fles ten minste 500 cl bevat.
Op welke inhoud moet het vulapparaat worden ingesteld?

X = de inhoud van de flessen in cl

$X \sim N(\mu, 5^2)$

→we zoeken μ.

De fles moet met een kans van 99% ten minste 500cl bevatten:

$$P(X \geq 500) = 0,99$$
$$P\left(\frac{X-\mu}{\sigma} \geq \frac{500-\mu}{5}\right) = 0,99$$
$$P\left(Z \geq \frac{500-\mu}{5}\right) = 0,99$$
$$\frac{500-\mu}{5} = -2,326$$
$$\underline{\mu = 511,63}$$

→de vulapparaat moet worden ingesteld op 511,63 cl

16. De stochasten X en Y zijn onafhankelijk en zijn beiden uniform verdeeld over de verzameling $\{1,2,3,4,5\}$.

- Geef de gezamenlijke dichtheid $p_{X,Y}$
- Geef de marginale dichtheden p_X en p_Y
- Geef de dichtheid van de stochast X + Y .

X en $Y \sim U\{1,2,3,4,5\}$
X en Y zijn onafhankelijk

■ gezamenlijke dichtheid:
$p_{X,Y} = p_X \cdot p_Y$
Met $p_X(i) = \frac{1}{5}$ voor i = 1,2,…,5
$p_Y(i) = \frac{1}{5}$ voor i = 1,2,…,5
$p_{X,Y}(i,j) = p_X(i).p_Y(i) = \frac{1}{25}$

■ marginale dichtheden p_X en p_Y
$p_X(i) = \frac{1}{5}$ voor i = 1,2,…,5
$p_Y(i) = \frac{1}{5}$ voor i = 1,2,…,5

■ dichtheid van de stochast X + Y:

X \ Y	**1**	**2**	**3**	**4**	**5**
1	2	3	4	5	6
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10

X+Y	p_{X+Y}
2	1/25
3	2/25
4	3/25
5	4/25
6	5/25
7	4/25
8	3/25
9	2/25
10	1/25

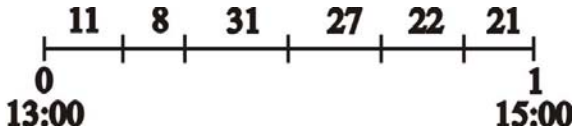
vb. 2 trekken met een kans 1/25
$P(X = 1 \text{ en } Y = 1)$
→ discrete stochast
→ gewichten tellen op tot 1

17. Je landt in de JFK-luchthaven in New York. Vanuit de luchthaven kun je de metro naar Manhatten nemen. Je informeert bij het loket naar de wachttijd tot de eerstvolgende metro.
De dame aan het loket is niet de eerste de beste. Haar antwoord luidt: het aantal metro's dat hier toekomt, volgt een Poisson-proces met een verwachte waarde van drie per uur.
• Op basis van deze informatie, hoe lang ga je moeten wachten tot de metro komt ?
Kun je argumenteren dat de verwachte wachttijd tien minuten bedraagt ?
• Welke stochast beschrijft de tijdsduur tussen de aankomsten van twee opeenvolgende metro's?

Veronderstel dat het tijdstip van uw aankomst op het metroperron lukraak getrokken wordt tussen 13.00 en 15.00, en dat de aankomsttijden van de metro als volgt zijn:
12:44, 13:11, 13:19, 13:50, 14:17, 14:39, 15:04
Deze getallen werden gesimuleerd via EXCEL.
• Bereken de verwachte wachttijd op basis van deze gegevens.
• Kun je nu intuitie aanbrengen waarom de verwachte wachttijd meer dan tien minuten bedraagt?

■ X = aantal treinen dat toekomt in het station per uur
$X \sim Pois(3)$
Verwachte waarde = 3 treinen per uur
= 1 per 20 minuten
Jij komt lukraak aan dus $\overline{X} = 10$ minuten

■ Y = tijdsduur tussen de aankomsten van twee opeenvolgende metro's
$Y \sim Exp(3)$
$E(Y) = \frac{1}{3}$

■ 

W = wachttijd gegeven uniform binnenkomen tussen 0 en 120 (13u00 en 15u00)

Verwachte wachttijd is de wachttijd maal de kans op die wachttijd:
$$E(W) = \left(\frac{11}{120}.\frac{11}{2}\right) + \left(\frac{8}{120}.4\right) + \left(\frac{31}{120}15,5\right) + \left(\frac{27}{120}.13,5\right) + \left(\frac{22}{120}.11\right) + \left(\frac{21}{120}.(10,5+4)\right)$$
De kans om in een slecht interval te komen is groter dan de kans om in een gunstig interval (minder dan 10 minuten wachttijd) te komen.
→dus; $E(W) > 10$

18. Moeilijk. Bij het simuleren van het vorige probleem, moet je een trekking doen uit een exponentieel verdeelde stochast. Jammer genoeg is de exponentiele verdeling niet opgenomen in de random-number-generation van EXCEL.
Deze oefening presenteert een methode om dergelijke problemen te omzeilen.

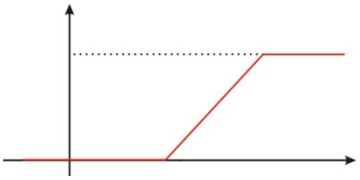
- Beschouw een continue stochast X met verdelingsfunctie F.
  Bepaal de verdeling (en/of dichtheid) van de stochast Y = F(X).
- Veronderstel nu dat de stochast U uniform verdeeld is over het interval [0, 1].
  Bepaal de verdeling (en/of dichtheid) van de stochast $F^{-1}(U)$
- Geef de transformatie die een rij van lukrake getallen uit [0, 1] (uniforme verdeling) omzet naar een rij van lukrake getallen getrokken uit een exponentiele verdeling met verwachtingswaarde 20.

X is een continue stochastische variabele
Dichtheid p_X, verdeling F_X

$Y = F_X(x)$ is een stochastische variabele $0 \le Y \le 1$

■ Vraag: bepaal de verdeling (en/of dichtheid) van Y:

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(F_X(x) \le y) \\ &= P(X \le F_X^{-1}(y)) \\ &= F_X\left(F_X^{-1}(y)\right) \\ &= y \end{aligned}$$



$Y \sim U[0,1]$: een uniforme verdeling

■ $U \sim U[0,1]$

$Z = F_X^{-1}(U)$ is een stochastische variabele $-\infty < F_X^{-1}(U) < +\infty$

$$\begin{aligned} F_Z(Z) &= P(Z \le z) \\ &= P(F_X^{-1}(U) \le z) \\ &= P(U \le F_X(z)) \\ &= F_X(z) \end{aligned}$$

! Z en X hebben dezelfde verdeling

Trek lukrake getallen tussen $[0,1]$

$y_1, y_2, ..., y_n$
Dan $F_X^{-1}(y_1), F_X^{-1}(y_2), ..., F_X^{-1}(y_n)$
Zijn lukrake getallen uit $X \sim Exp(\lambda)$

19. Uit een enquete (Nederland) blijkt dat economie studenten ongeveer 25 uur van de hun toebedeelde 40-urige werkweek besteden aan studie. Desalniettemin verwacht de student een redelijke kans op slagen. Je kan vergelijken met een werknemer die ongeveer halftijds werkt en toch de volle honderd procent van het loon wil opstrijken. Studenten beschouwen het systeem van herexamens als extra ruimte en stimulans voor de driedaagse werkweek. Om de studenten aan te zetten tot een regelmatige en tijdige studie 'leeft' het voorstel om de tweede zittijd af te schaffen. De tweede zittijd zou vervangen worden door een loterij waarbij een niet geslaagde student een kans p heeft op een herexamen. Op die manier wordt een incentief gegeven aan de student. Ook de docent doet profijt: minder (her-)examens betekent immers meer tijd voor het uitvinderschap. De volgende pay-off-matrix geeft de uitkomsten voor de twee groepen van spelers, docenten en studenten. Docenten kunnen de waarde van p bepalen. Studenten kunnen kiezen tussen voltijds studeren en een driedaagse werkweek.

		Student **Voltijds (q)**	**Deeltijds (1-q)**
Docent	Herexamen (p)	(2,2)	(-1,3)
	Geen Herexamen (1-p)	(1,1)	(0,-1)

De eerste (tweede) coordinaat in een koppel verwijst naar de pay-off voor de docent (student).

• Geef een interpretatie aan deze matrix. In de bovenstaande contekst, wat is de 'logica' achter deze cijfers ?

• Toon aan dat de matrix geen Nash-evenwicht in zuivere strategieen heeft.

• Indien de student een gemengde strategie aanhoudt met q de kans op een voltijdse werkweek, bepaal dan het Nash-evenwicht in gemengde strategieen. Interpreteer je oplossing.

(Uitkomst: p = 2=3 en q = 1=2.)

Bron: C.G. De Vries (december 2004) \Stimulans en kans", Erasmus Universiteit Amsterdam, http://www.few.eur.nl/few/people/cdevries/oratie_nalwebfeb.pdf

■ We beginnen bij (0,-1) : de student heeft baat bij voltijds studeren
→ we bewegen naar (1,1) : de docent heeft baat bij herexames
→ we bewegen naar (2,2) : de student heeft baat bij deeltijds studeren
→ we bewegen naar (-1,3) : de docent heeft baat bij een herexames
→ we zijn terug in (0,1)
Er is geen Nash evenwicht in zuivere strategieen. Er is altijd incentief om te bewegen.

■ Evenwicht in mixed strategies
* p versus 1-p = studenten indifferent maken

$$E\left(\text{payoff student} \mid \text{voltijds}\right) = 2p + 1(1-p)$$

$$E\left(\text{payoff student} \mid \text{deeltijds}\right) = 3p + (-1)(1-p)$$

indifferent maken → $2p + 1 - p = 3p - 1 + p \;\Rightarrow\; p^* = \frac{2}{3}$

dus de docenten moeten een p kiezen van $\frac{2}{3}$

* q versus 1-q = docent indifferent maken

$$E\left(\text{payoff docent} \mid \text{herexamens}\right) = 2q - 1(1-q)$$

$$E\left(\text{payoff docent} \mid \text{geen herexamens}\right) = 1q - 0(1-q)$$

indifferent maken → $2q - 1 + q = q \;\Rightarrow\; q^* = \frac{1}{2}$

Oefenzitting 5 – Statistische besluitvorming

1. Om de kwaliteit van de ijzererts te onderzoeken baseert de koper zich op de zuiverheidsgraden $Y_1, Y_2, ..., Y_{40}$ van 40 lukraak getrokken stalen ijzererts. De stochasten Y_i zijn onafhankelijk en hebben allen de volgende dichtheid:

$$p : \mathbb{R} \to \mathbb{R} : x \to p(x) = \begin{cases} 3x^2 & \text{indien } 0 \le x \le 1 \\ 0 & \text{elders} \end{cases}$$

De erts wordt aanvaard indien $\overline{Y} \ge 0,7$. Bepaal de kans $P(\overline{Y} \ge 0,7)$.

We gaan gebruik maken de Centrale Limiet Stelling (CLS) en de normale verdeling. Hiervoor moeten we eerst de gemiddelde en variantie vinden.

1^{ste} centrale moment : $E(Y_i) = \int_{-\infty}^{+\infty} x p_X = \int_0^1 x 3x^2 = \int_0^1 3x^3 = \left[\frac{3x^4}{4}\right]_0^1 = \frac{3}{4} = \mu$

2^{de} centrale moment : $E(Y_i^2) = \int_{-\infty}^{+\infty} x^2 p_X = \int_0^1 x^2 3x^2 = \int_0^1 3x^4 = \left[\frac{3x^5}{5}\right]_0^1 = \frac{3}{5}$

$$\text{var}(Y_i) = E(Y_i^2) - (E(Y_i))^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80} = \sigma^2$$

We gebruiken nu CLS:

$$\overline{Y} = \frac{Y_1 + Y_2 + ... + Y_n}{n} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}} \approx N(0,1)$$

We kunnen nu de kans zoeken:

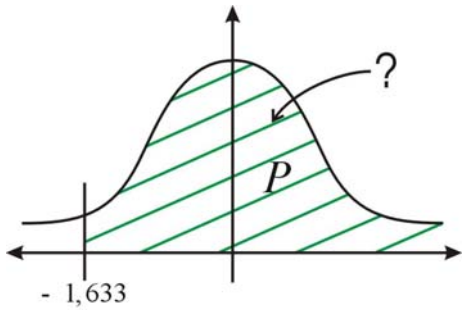
$$P = P(\overline{Y} \ge 0,7)$$

$$= P\left(Z = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}} \ge \frac{0,7 - \frac{3}{4}}{\sqrt{\frac{30}{80}} / \sqrt{40}} = -1,633\right)$$

$$= 1 - 0,0516$$

$$= 0,949$$

$$P(\overline{Y} \ge 0,7) = 94,9\%$$



3. Het bedrag dat klanten hebben uitstaan bij een welbepaalde bank bedraagt gemiddeld 200 000 BEF en heeft een standaardafwijking van 26 000 BEF. Een willekeurige steekproef van 36 klanten wordt getrokken.
Wat is de kans dat het gemiddelde bedrag dat deze 36 klanten hebben uitstaan gelegen is tussen 192 000 en 212 000 BEF?

Ω = verzameling van klanten bij een bepaalde bank

$X : \Omega \to \mathbb{R}$

$\omega \to X(\omega)$ = spaarvolume van klant ω bij deze bank in BEF

$\mu = 200000$

$\sigma = 26000$

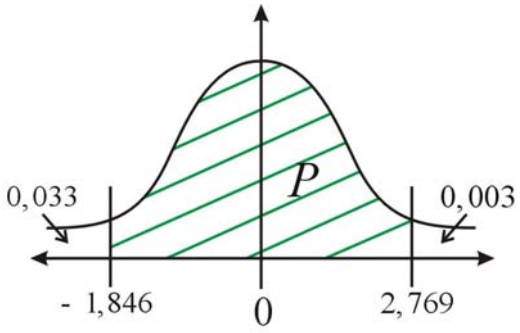
We gebruiken de Centrale Limiet Stelling:

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \approx N(0,1)$$

We transformeren naar een standaard normale verdeling en zoeken de kans:

$$\begin{aligned} P &= P(192000 \le \bar{X} \le 212000) \\ &= P\left(\frac{192000 - \mu}{\sigma / \sqrt{n}} \le Z \le \frac{212000 - \mu}{\sigma / \sqrt{n}}\right) \\ &= P\left(\frac{192000 - 200000}{26000 / 6} \le Z \le \frac{212000 - 2000000}{26000 / 6}\right) \\ &= P\left(-\frac{(8)(6)}{26} \le Z \le \frac{(12)(6)}{26}\right) \\ &= P\left(-1,846 \le Z \le 2,769\right) \\ &= 1 - 0,033 - 0,003 \\ &= 0,964 \end{aligned}$$



De kans dat het gemiddelde bedrag dat deze 36 klanten hebben uitstaan gelegen is tussen 192 000 en 212 000 BEF is 96,4%.

4. Een student wil voor zijn eindwerk het gemiddelde van een normaal verdeelde stochast benaderen via een 95%-b.i. met behulp van een steekproef. Bepaal de (minimale) lengte van de steekproef, indien hij een intervalbreedte van 1 eenheid wil toestaan bij een standaardafwijking van 2 eenheden.

$X \sim N(\mu, \sigma^2)$

95% betrouwbaarheids interval : $\alpha = 0,05$

Zoek n zodat I = 1 bij $\sigma = 2$

We gebruiken de Centrale Limiet Stelling voor de steekproef:

$$\bar{X} = \frac{X_1 + X_2 + ... + X_n}{n} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \approx N(0,1)$$

We kennen de kans en zoeken dus de waarde van n:

$$P = P(-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq Z_{\alpha/2}) = 0,95$$

$$\mu = \bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Dit is hetzelfde als: $\mu = \bar{X} \pm I$

$$I = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

We willen de waarde van n vinden als I =1:

$$1 = 1,96 \frac{2}{\sqrt{n}}$$

$$n = 15,3664$$

→de minimale lengte van de steekproef is 15.

5. De scores op een welbepaald examen worden verondersteld normaal verdeeld te zijn met een gemiddelde van 70 (op 100) en een standaardafwijking van 18.
Een willekeurige steekproef van 50 studenten genereert een gemiddelde score van 67. Kan men op basis hiervan besluiten dat het werkelijke gemiddelde lager ligt dan 70?

X = de score op het examen
We veronderstellen dat X normaal verdeeld is:

$$X \sim N(\mu, \sigma^2 = 18^2)$$

$H_0 : \mu = 70$: dit willen we verwerpen, dat het gemiddelde gelijk is aan 70

$H_1 : \mu < 70$: dit willen we bewijzen, dat het gemiddelde lager ligt dan 70

Geobserveerde data: n = 50, $\bar{x} = 67$.

Toetsstochast:

$$\bar{X}|_{H_0} \sim N(70, \sigma^2 = 18^2 / n)$$

We transformeren dit naar de standaard normale verdeling:

$$Z = \frac{\bar{X}|_{H_0} - 70}{18 / \sqrt{50}} \sim N(0,1)$$

Geobserveerde waarde: $Z_{obs} = \frac{67 - 70}{18 / \sqrt{50}} = -1,1785$

De kans dat $Z \leq -1,1785$ wordt gegeven door $\alpha = 0,1193$ (van de tabel)

Besluit: we kunnen H_0 niet verwerpen ($1 - \alpha = 88,07\%$ wordt als te klein aanzien).
M.a.w. het verschil tussen de geobserveerde 67 en de verwachte 70 is niet significant of niet betekenisvol.

7. Een advertentie voor een dactylo-cursus vermeldt dat de cursisten leren typen aan een snelheid van 30 woorden/min. Een test bij 15 ex-cursisten levert een gemiddelde op van slechts 28,5 woorden/min. Met een standaardafwijking van 2,5 woorden/min.
Kan u de inhoud van de advertentie bijtreden bij een betrouwbaarheidsniveau van 95%?

Ω = verzameling van dactylo- cursisten
$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$: aantal woorden die cursist ω kan typen per minuut

Toets:
$H_0 : \mu = 30$
$H_1 : \mu < 30$: wat men wil "bewijzen" door H_0 te verwerpen

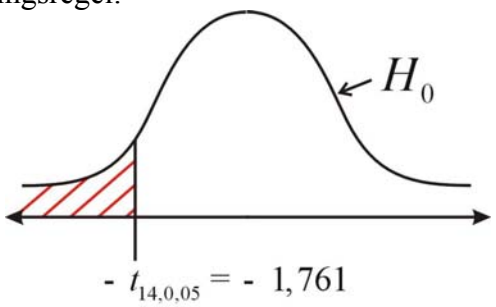
Toetsstochast : $\bar{X}|_{H_0} \sim N\left(30, \frac{\sigma^2}{n}\right)$

$$Z_{obs} = \frac{\bar{X} - 30}{\sigma / \sqrt{n}}|_{H_0} \sim N(0,1)$$

σ^2 kennen we niet dus moeten schatten met S^2
→ hier door hebben we dikkere staarten nodig dan de normale verdeling dus gebruiken we de T verdeling

$$T = \frac{\bar{X} - 30}{S / \sqrt{n}}|_{H_0} \sim t_{n-1} = t_{14}$$

Beslissingsregel:

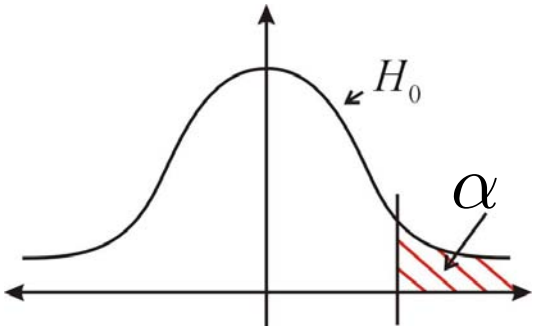


Uitvoeren van de toets:
$T|_{obs} = \frac{28,5 - 30}{2,5 / \sqrt{15}} = -2,324 < -t_{14,0,025} = -1,761$
→ $-2,324$ ligt ruim in het kritisch gebied (het rode gebied) dus kunnen we H_0 verwerpen.
→ we hebben "bewezen" dat H_0 verworpen mag worden op een niveau van $-t_{14,0,025}$ dus met 97,5% zekerheid

8. Bij de productie van houten stoelen mag de vochtigheidsgraad van het hout niet hoger zijn dan 10%. Bij een nieuwe lading hout wordt van een lukraak getrokken staal de vochtigheidsgraad gemeten.
Formuleer een nul- en alternatieve hypothese omtrent de bruikbaarheid van het hout. Interpreteer de type I en de type II fout. Welke van deze fouten ziet de leverancier (resp. de producent) liever niet gemaakt worden?

$X(\omega)$ = vochtigheidsgraad van hout staal ω

$H_0 : E(X) \leq 0{,}10$: bruikbaar, vochtigheidsgraad laag genoeg
$H_1 : E(X) > 0{,}10$: niet bruikbaar, vochtigheidsgraad te hoog



Type-1-fout (kans α)
Hout is droog, maar door staal trekking wordt het niet aanvaard
→ de leverancier verliest

Type-2-fout (kans β)
Hout is nat, maar door staal trekking wordt het toch aanvaard
→ de stoelmaker verliest

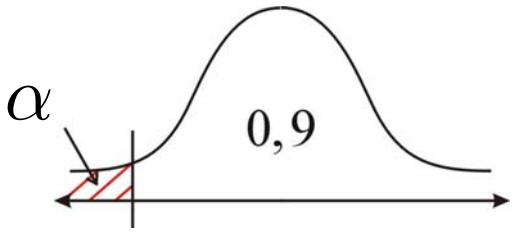
10. Een bedrijf is er in geslaagd om de levensduur van een welbepaald toestel te verhogen en wil dit in haar brochure onderstrepen. Een steekproef uit deze toestellen levert de volgende levensduren op:

100, 20, 90, 119, 105, 103, 93, 130, 112, 107 uren

Hoe kan het bedrijf de reclame-boodschap formuleren, rekening houdend met een 90% zekerheid dat er niet gelogen wordt.

Ω = verzameling van welbepaalde toestellen
$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$: levensduur van toestel ω uitgedrukt in uren

■ stel $X \sim N(\mu, \sigma^2)$
→ een éénzijdig betrouwbaarheids interval voor μ



■ $Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

We kennen σ niet dus we gebruiken S als een schatter en gebruiken dus de T verdeling:

$T = \dfrac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}$

$P\left(\dfrac{\bar{X} - \mu}{S / \sqrt{n}} \leq t_{9,0,1}\right) = 0,09$

: normaal gebruiken we $\geq$, maar omdat er $\bar{X} - \mu$ staat moeten we $\leq$ gebruiken

■ betrouwbaarheids interval: $\dfrac{\bar{X} - \mu}{S / \sqrt{n}} \leq t_{9,0,1}$

$\bar{X} - \mu \leq \dfrac{S}{\sqrt{n}} t_{9,0,1}$

$-\mu \leq \dfrac{S}{\sqrt{n}} t_{9,0,1} - \bar{X}$

$\mu \geq -\dfrac{S}{\sqrt{n}} t_{9,0,1} + \bar{X}\big|_{obs} = 84,9$

→ met 90% zekerheid is μ groter dan 84,9 = 90% van de toestellen haalt 84,9 uren

11. Een psycholoog wenst de gemiddelde reactietijd te berekenen. Hij gaat ervan uit dat de standaardafwijking van de reactietijd 0,05 seconden bedraagt. Hoe groot moet de steekproef zijn opdat de psycholoog met 95% zekerheid kan zeggen dat zijn geschat gemiddelde met minder dan 0,01 seconde afwijkt van het ware gemiddelde?

X = reactietijd

$\Omega : X \to \mathbb{R}$

$\omega \to X(\omega)$: reactietijd van persoon ω in seconden

We veronderstellen dat X normaal verdeeld is en gebruiken de CLS voor $\bar{X}$:

$X \sim N(\mu, \sigma^2)$

$\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$

We gaan dit transformeren naar de standaard normale verdeling:

$Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

$P\left(-z_{\alpha/2} \leq \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2}\right)$

→ $\mu = \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

$\mu - \bar{X} = \pm 1,96 \frac{0,05}{\sqrt{n}}$

$\left|\mu - \bar{X}\right| = 1,96 \frac{0,05}{\sqrt{n}}$

Sinds het geschat gemiddelde met minder dan 0,01 seconde afwijkt van het ware gemiddelde weten we dat : $\left|\mu - \bar{X}\right| < 0,01$

$(1,96)(0,05) < 0,01\sqrt{n}$

$\frac{(1,96)(0,05)}{0,01} < \sqrt{n}$

$n > 9,8^2$

$n > 96,04$

→ de steekproef met tenminste 97 mensen bevatten.

12. De lampen uit bedrijf A (B) hebben een gemiddelde levensduur van 1400 (1200) uren met standaardafwijking van 200 (100) uren. Neem uit elk bedrijf een steekproef van lengte 125. Wat is de waarschijnlijkheid dat de steekproef uit merk A een gemiddelde levensduur heeft van 160 uren meer dan die uit merk B?

X = levensduur van een lamp van merk A
$\mu_X = 1400$ en $\sigma_X^2 = 200^2$

Y = levensduur van een lamp van merk B
$\mu_Y = 1200$ en $\sigma_Y^2 = 100^2$

Er wordt gevraagd naar het verschil in gemiddelde levensduur van twee grote steekproeven ($n_A = n_B = n = 125$) dus kunnen we CLS gebruiken:

$\bar{X} \approx N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$

$\bar{Y} \approx N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$

Het verschil van 2 onafhankelijke normaal verdeelde stochasten blijft normaal:

$\bar{X} - \bar{Y} \approx N\left(\mu_X - \mu_Y, \frac{\sigma_X^2 + \sigma_Y^2}{n}\right)$

We transformeren dit naar de standaard normale verdeling:

$Z = \frac{\bar{X} - \bar{Y} - 200}{20} \approx N(0,1)$

We kunnen nu de gevraagde kans brekenen:

$$P = P(\bar{X} - \bar{Y} \geq 160) = P\left(Z \geq \frac{160-200}{20} = -2\right) = 0,9772$$

→ de kans dat de steekproef uit merk A een gemiddelde levensduur heeft van 160 uren meer dan die uit merk B is 97,72%

13. Het stemmenaandeel p van een politieke partij wordt onderzocht via een opiniepeiling. Toon aan dat, om p te kennen op $\pm 3\%$ nauwkeurigheid (met een betrouwbaarheid $1-\alpha = 95\%$) de lengte van de steekproef minstens 1000 moet zijn.

Ω = populatie van kiezers

$X : \Omega \to \mathbb{R}$

$\omega \to X(\omega)$ = 1 indien ω stemt voor een welbepaalde partij

= 0 indien ω niet stemt voor een welbepaalde partij

$X \sim b(1, p)$: er zijn twee mogelijke uitkomsten

We gebruiken CLS:

$$Z = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1)$$

We kennen p niet, dus omdat n groot is, kunnen we p schatten met $\hat{p}_{obs}$:

($\hat{p}_{obs}$ is consistent en daarom behouden we de standaard normale verdeling)

$$\frac{\bar{X} - p}{\sqrt{\frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}}} \approx N(0,1)$$: er blijft nog één 'p' in de vergelijking : één onbekende

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - p}{\sqrt{\frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}}} \leq z_{\alpha/2}\right) = 0,95$$

betrouwbaarheids interval voor p:

$$p = \bar{X} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}}$$

($\sqrt{\frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}}$ bereikt een maximum voor $\hat{p}_{obs} = \frac{1}{2}$)

Gegeven: $\alpha = 0,05$

$$I = z_{\alpha/2}\sqrt{\frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}} < \varepsilon \text{ : halve interval breedte}$$

$$\leq z_{\alpha/2}\frac{1}{2\sqrt{n}} < \varepsilon = 0,03$$

$$\frac{1,96}{(2)(0,03)} = \frac{z_{\alpha/2}}{2\varepsilon} \leq \sqrt{n}$$

Antwoord: n > 1068

14. Een vaas bevat 60 rode en 40 blauwe knikkers. Er worden twee steekproeven getrokken van lengte 30 (telkens met teruglegging). Bepaal de kans dat het aantal rode knikkers in deze steekproeven met minstens 8 eenheden verschilt.

Ω = vaas met knikkers

$X : \Omega \to \{0,1\}$

$\omega \to X(\omega)$ = 1 indien ω rood is

= 0 indien ω wit is

$X \sim b(1, p)$ met p=0,6

Twee steekproeven met teruglegging: n=30:

- Y_i = aantal rode knikkers in steekproef i

$Y_i \sim b(30, p)$

$\approx N\left(30p, \sigma^2 = 30p(1-p)\right)$: CLS

- $Y_1 - Y_2 \approx N\left(0, \sigma^2 = (2)(30)p(1-p)\right)$: bij onafhankelijkheid tellen de varianties op

$$Z = \frac{Y_1 - Y_2}{\sqrt{60p(1-p)}} \approx N(0,1)$$

- $P = P\left(|Y_1 - Y_2| \geq 7,5\right)$

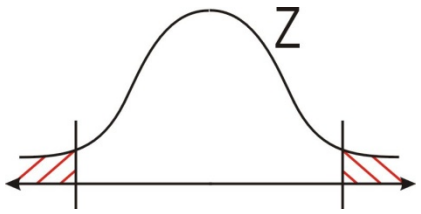
→ we gebruiken 7,5 ipv 8 : continuiteitscorrectie

→ omdat we overstappen van een discrete stochast naar een continue stochast

$$= P\left(|Z| \geq \frac{7,5}{\sqrt{(60)(0,6)(0,4)}} = 1,976\right)$$

$= 2P(Z \geq 1,976)$: **symmetrie**

$= 0,048$



15. Een kandidaat behaalde in een verkiezing 65% van de stemmen. Wat is de kans dat twee steekproeven van lengte 200 proporties voorspellen die met meer dan 10% verschillen?

$X \sim b(1, p) \qquad p = 0,65$

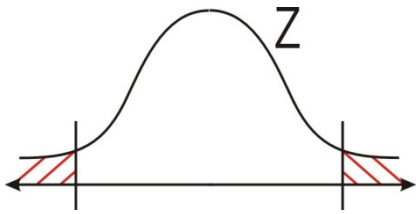
Twee onafhankelijke steekproeven n=200

Y_i = aantal stemmen voor die bepaalde kandidaat in steekproef i; i=1,2

- $\overline{Y}_i \approx N\left(p, \frac{p(1-p)}{n}\right)$: CLS

$$\overline{Y}_1 - \overline{Y}_2 \approx N\left(0, \frac{2p(1-p)}{n}\right)$$

- $Z = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{2p(1-p)}{n}}} \approx N(0,1)$



- $P = P\left(\left|\overline{Y}_1 - \overline{Y}_2\right| \geq 0,01\right)$

$$= P\left(\frac{\left|\overline{Y}_1 - \overline{Y}_2\right|}{\sqrt{\frac{2p(1-p)}{n}}} \geq \frac{0,01}{\sqrt{\frac{2p(1-p)}{n}}}\right)$$

$$= P\left(|Z| \geq \frac{0,01}{\sqrt{\frac{2p(1-p)}{n}}}\right)$$

$$= P\left(|Z| \geq \frac{0,01}{\sqrt{\frac{(0,65)(0,35)}{100}}}\right)$$

$$= P\left(|Z| \geq \frac{1}{\sqrt{(0,65)(0,35)}}\right)$$

$$= P\left(|Z| \geq 2,09\right) = 0,0183$$

$= 2P\left(|Z| \geq 2,09\right)$: symmetrie

De kans is dus gelijk aan 2 x 0,0183 = 0,0366

16. In deze oefening tonen we aan dat er geen onvertekende schatter bestaat voor de parameter λ in een exponentiele verdeling. Zij $X \sim Exp(\lambda)$.

• We weten dat $E(X) = 1/\lambda$. Waarom kunnen we niet besluiten dat 1/X een onvertekend schatter is van λ, i.e. E(1/X)= λ?

• Veronderstel dat de stochast $\hat{\lambda}(X)$ een onvertekend schatter is van λ, waarbij de functie $\hat{\lambda}: \mathbb{R}^+ \to \mathbb{R}^+$ continu is. Schrijf de gelijkheid neer waaraan $\hat{\lambda}$ moet voldoen.

• Werk verder met deze gelijkheid. Deel aan weerskanten door λ, en leid (partieel) af naar λ.

• Je bekomt: $\int_{x=0}^{\infty} \hat{\lambda}(x) x e^{-\lambda x} = 0$

• Besluit dat de functie $\hat{\lambda}$ samenvalt met de nulfunctie. Maak hierbij gebruik van de veronderstelling dat de functie $\hat{\lambda}: \mathbb{R}^+ \to \mathbb{R}^+$ continu is.

• Beeindig nu het bewijs(je).

■ We moeten aantonen dat er niet altijd een unbiased schatter bestaat.

Vraag: $E[X] = \frac{1}{\lambda}$ is dan $E\left[\frac{1}{X}\right] = \lambda$?

→**NEE** Waarom? De functie heeft een positieve kromming $f: X \to \frac{1}{X}$

Denk aan de ongelijkheid van Jensen:

$E(U(X))$

$U(E(X))$

Enkel bij nulkromming is de schatter niet vertekend

■ $E\left(\hat{\lambda}(x)\right) = \lambda$

$\int_0^{\infty} \hat{\lambda}(x) x e^{-\lambda x}\, dx = \lambda$

■ $-\int_0^{\infty} \hat{\lambda}(x) x e^{-\lambda x}\, dx = 0$

$I = \int_0^{\infty} \hat{\lambda}(x) x e^{-\lambda x}\, dx = 0$

\+ ++ 0 → kan niet! positieve functie = positieve oppervlakte

→ dus $\hat{\lambda}(x) = 0$ voor elke x

[Zo niet : $\hat{\lambda}(x_0) > 0$ met $x_0 > 0$

$\hat{\lambda}(x_0) > 0$ voor x in $[x_0 - \varepsilon, x_0 + \varepsilon]$

$I \geq \int_{x_0-\varepsilon}^{x_0+\varepsilon} \hat{\lambda}(x) x e^{-\lambda x} > 0$]

Oefenzitting 6 – Statistische besluitvorming, schatten van relaties

1. Een voedingsbedrijf wenst de hypothese te toetsen dat het gemiddelde gewicht van een varkentje, gevoed met een nieuw mengvoeder 160 pond bedraagt. De onderneming zal deze hypothese verwerpen indien het steekproefgemiddelde van 25 varkens minder dan 155 pond bedraagt. Op grond van vroegere testen weet men dat de standaardafwijking van de populatie 25 pond bedraagt.
• Bereken de waarschijnlijkheid van een type I fout.
• Indien het ('ware') populatiegemiddelde 156 pond bedraagt, bereken de kans van een fout van type II.

Ω = verzameling varkentjes gevoed met het nieuw mengvoeder
$X : \Omega \to \mathbb{R}$
$\omega \to X(\omega)$: het gewicht van varkentje ω in pond

$H_0 : \mu_X = 160$
$H_1 : \mu_X < 160$

Beslissingsregel: verwerp H_0 indien:
$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{25} < 155$ pond

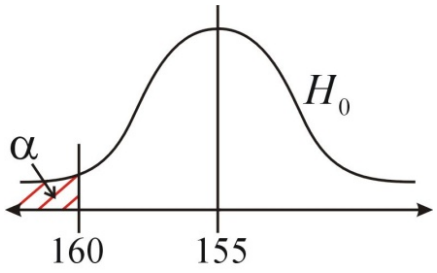
Bijkomend gegeven : $\sigma_X = 25$

■ Kans op Type-I fout:

Definitie van Type-I fout: $P(\text{type 1} - \text{fout}) = P(H_0 \text{ verwerpen} \mid H_0 \text{ waar})$

We veronderstellen dat $X \sim N$

$\overline{X}|_{H_0} \sim N\left(\mu_X, \sigma^2 = \frac{\sigma_X^2}{25}\right)$



$Z|_{H_0} = \frac{\overline{X} - \mu_X}{\sigma / \sqrt{n}}|_{H_0} = \frac{\overline{X} - 160}{5}|_{H_0} \sim N(0,1)$

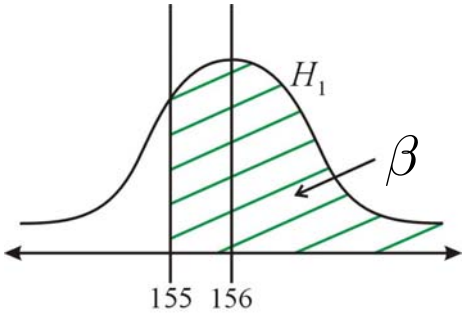
$$P(\overline{X} \le 155) = P\left(\frac{\overline{X} - \mu_X}{\sigma / \sqrt{n}} \le \frac{155 - 160}{5}\right) = P\left(Z \le \frac{155 - 160}{5} = -1|_{H_0}\right) = 0{,}1587$$

→ kans op Type-I fout = $\alpha = 15{,}8\%$

■ Kans op Type-II fout:

$P\left(\bar{X} > 155 \middle| \mu_X = 156\right)$

Stel : $\mu_X = 156$



Dan $\bar{X} \sim N\left(156, \sigma^2 = \frac{25^2}{25}\right)$

$Z = \frac{\bar{X} - 156}{5} \sim N(0,1)$

$P\left(Z > \frac{155-156}{5} = -0,2\right) = 0,579$

Kans op Type-II fout = $\beta = 57,9\%$

2. Om de hypothese te toetsen dat een muntstuk "eerlijk" is, wordt de volgende beslissingsregel gebruikt: aanvaard de hypothese indien het aantal keer kop in een enkele steekproef van 100 worpen gelegen is tussen 40 en 60, verwerp de hypothese indien het aantal keer kop niet in dit interval ligt.
Bepaal de kans dat de hypothese wordt verworpen hoewel ze eigenlijk correct is.

X = aantal keer kop in 100 worpen
$X \sim b(100, p)$: twee mogelijke uitkomsten = Bernouilli

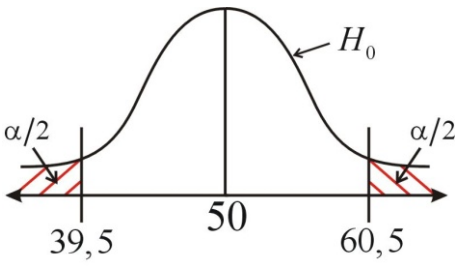
We maken gebruik van de central limiet stelling (CLS):
$X \approx N\left(100p, 100p(1-p)\right)$

$H_0 : p = 1/2$
$H_1 : p \neq 1/2$

Beslissingsregel: verwerp H_0 indein $X \notin [40,60]$

Kans dat de hypothese wordt verworpen hoewel ze eigenlijk correct is
= kans op Type-I fout

$X\big|_{H_0} \approx N\left(50, \sigma^2 = 25\right)$



We voeren een continuiteitscorrectie uit omdat we overgaan van een discrete stochast (Bernouilli) naar een continue stochast (Normale):
H_0 aanvaarden indien $X \in [39,5 : 60,5]$

$Z = \frac{X - 50}{5}\Big|_{H_0} \approx N(0,1)$

$$P\left(X \notin [39,5:60,5]\big|_{H_0}\right) = 2P\left(X \geq 60,5\big|_{H_0}\right)$$
$$= 2P\left(Z \geq \frac{10,5}{5} = 2,1\right) = 0,036$$

Kans op Type-I fout = $\alpha = 3,6\%$

3. Een steekproef van 250 eenheden wordt verdeeld over de mogelijke uitkomsten zoals weergegeven in de volgende tabel:

Mogelijke uitkomsten	0	1	2	3	4	5	6	7	8	9
Geobserveerde frequentie	17	31	29	18	14	20	35	30	20	36

De verwachte verdeling was een uniforme verdeling, of 25 observaties per mogelijke uitkomst.
Is de geobserveerde verdeling significant verschillend van de verwachte verdeling?

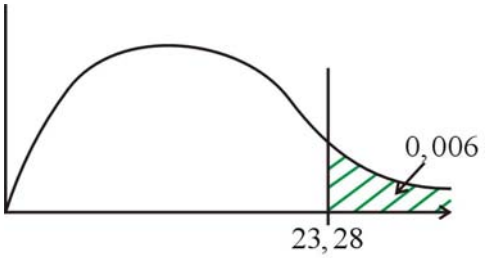
$H_0 : X \sim \mu$

O_k	17	31	29	18	14	20	35	30	20	36	=250
e_k	25	25	25	25	25	25	25	25	25	25	

De onderste rij is de expected value sinds we een uniforme verdeling verwachten.

Toetsstochast : $T = \sum_{k=1}^{10} \frac{(O_k - e_k)^2}{e_k}\Big|_{H_0} \sim \chi^2_{10-1}$: in het formularium p6

$$T_{obs} = \frac{8^2 + 6^2 + 4^2 + 7^2 + 11^2 + 5^2 + 10^2 + 5^2 + 5^2 + 11^2}{25}$$
$$= \frac{582}{25} = 23,28$$



$P(T \geq 23,28) = 0,006$
→ zeer klein

Verwerp H_0 met $\alpha = 0,7\%$

4. Een steekproef bij 320 families met 5 kinderen toont de volgende verdeling:

(aantal jongens , aantal meisjes)	(5,0)	(4,1)	(3,2)	(2,3)	(1,4)	(0,5)
aantal geobserveerde families	18	56	110	88	40	8

Is dit resultaat in overeenstemming met de hypothese dat de geboorten van jongens en van meisjes even waarschijnlijk zijn ($\alpha = 0.05$)?

$H_0 : X \sim b\left(5, \frac{1}{2}\right)$: experiment wordt 5 maal herhaald met 2 mogelijke uitkomsten

Kans 0 jongens: $p(k) = \binom{n}{k} p^k q^{n-k}$: dichtheid van Bernouilli : in het formularium p3

$$= \binom{5}{0} 0,5^0 0,5^5 = \frac{1}{32}$$

$e_k = p(k).x = \frac{1}{32}.320 = 10$

Kans 1 jongen: $P(X = 1) = \frac{5}{32}.320 = 50$

Kans 2 jongens: $P(X = 2) = \frac{10}{32}.320 = 100$

Kans 3 jongens: $P(X = 3) = 100$

Kans 4 jongens: $P(X = 4) = 50$

Kans 5 jongens: $P(X = 5) = 10$

} symmetrie van $X \sim b\left(5, \frac{1}{2}\right)$

Met deze informatie kunnen we nu een tabel maken:

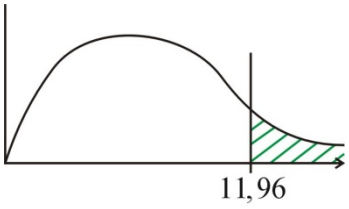
O_k	18	56	110	88	40	8
e_k	10	50	100	100	50	10

=320

Toetsstochast : $T = \sum \frac{(O_k - e_k)^2}{e_k}\Big|_{H_0} \sim \chi^2_{6-1}$: in het formularium p6

$T_{obs} = \frac{8^2}{10} + \frac{6^2}{50} + \frac{10^2}{100} + \frac{12^2}{100} + \frac{10^2}{50} + \frac{2^2}{10} = 11,96$

$P(T \geq 11,96) = 0,035$



→ verwerp H_0 zodra $\alpha > 0,035$

6. Beschouw het deterministische lineaire regressie probleem met één verklarende en één te verklaren variabele.
Toon aan dat de KK-schatters $\hat{a}$ en $\hat{b}$ voor het intercept a en de richtingscoefficient b, inderdaad de som der kwadraten van de residuele afwijkingen minimaliseert. Maw verifieer de voorwaarde van tweede orde.

$\underset{\hat{a},\hat{b}}{\text{minimaliseer}}\ f(\hat{a},\hat{b}) = \sum (y_i - \hat{a} - \hat{b}x_i)^2$

$$\frac{\partial f}{\partial \hat{a}} = -2\sum (y_i - \hat{a} - \hat{b}x_i) = 0$$

: voorwaarden van 1e orde

$$\frac{\partial f}{\partial \hat{b}} = -2\sum (y_i - \hat{a} - \hat{b}x_i) = 0$$

$\frac{\partial^2 f}{\partial \hat{a}^2} = 2n$: 2e afgeleide

$\frac{\partial^2 f}{\partial \hat{a}\partial \hat{b}} = 2\sum x_i$: gemengde afgeleide

$\frac{\partial^2 f}{\partial \hat{a}^2} = 2\sum x_i^2$

Hessiaan: $\begin{pmatrix} Z''_{XX} & Z''_{XY} \\ Z''_{YX} & Z''_{YY} \end{pmatrix} = \begin{pmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{pmatrix}$ positief definiet?

■ diagonaal elementen > 0 : OK

■ determinant (Hessiaan) $= n\sum x_i^2 - \left(\sum x_i\right)^2$

$= n\left(\sum x_i^2 - n\bar{x}^2\right)$

$= n\left(\sum (x_i - \bar{x})^2\right) > 0$: OK

Besluit: Hessiaan is idd positief definiet (we zoeken een minimum)

7. Een lineaire regressie $Y = \alpha + \beta x + U$ levert de volgende vergelijking op:

$$Y = -4 + 0,65x + U$$

Bovendien is er gegeven dat n = 30, $\bar{x} = 35.6$, $\Sigma U_i^2 = 1.43$, en dat $\Sigma(x_i - \bar{x})^2 = 2.9$.

- Bepaal een puntschatting voor Y wanneer x = 38,
- Bepaal de grenzen van 95%-b.i. rond Y wanneer x = 39,
- Bepaal de grenzen van 95%-b.i. rond E(Y) wanneer x = 39,
- Test de nulhypothese dat Y = 20 wanneer x = 39 (significantieniveau α = 0.02).

$Y = -4 + 0,65x + U$

n = 30, $\bar{x} = 35.6$, $\Sigma U_i^2 = 1.43$, $\Sigma(x_i - \bar{x})^2 = 2.9$

■ x = 38 : puntschatting voor Y (vertekend)

$Y = -4 + (0,65)(38) = 20,7$

■ x = 39 : puntschatting voor Y (vertekend)

$Y = 20,7 + 0,65 = 21,35$

■ <u>betrouwbaarheids interval voor E(Y):</u>

$$21,25 \pm \underbrace{\sqrt{\frac{1,43}{30-2}\left(\frac{1}{30}+\frac{(39-35,6)^2}{2,9}\right)}}_{0,928} \sim t_{28;0,025}$$

■ <u>voorspellings interval voor Y:</u>
betrouwbaarheids interval voor Y:

$$21,35 \pm \underbrace{\sqrt{\frac{1,43}{30-2}\left(1+\frac{1}{30}+\frac{(39-35,6)^2}{2,9}\right)}}_{1,037} \sim t_{28;0,025}$$

8. Beschouw de volgende data in (x, y).

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

A • Voer de regressie uit met x als verklarende variabele (i.e. $Y = a + bx$).
B • Voer de regressie uit met y als verklarende variabele (i.e. $X = a_0 + b_0 y$).
• Indien je de tweede vergelijking oplost naar y (i.e. $y = (X - a_0)/b_0$) wat we interpreteren als $Y = (x - a_0)/b_0$; bekom je dan hetzelfde resultaat als bij de eerste regressie? Kun je een verklaring geven?

$\sum x_i = 56 \quad \sum y_i = 40$
$\sum x_i^2 = 524 \quad \sum y_i^2 = 256 \quad \sum x_i y_i = 364$

<u>Deel A</u>

■ $Y = a + bx + U$

$\hat{b} = \dfrac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_j - \overline{x})^2}$: in het formularium p6

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

$$= \frac{(8)(364) - (56)(40)}{(8)(524) - 56^2}$$

$$= \frac{7}{11} = 0,636$$

■ $\hat{a} = \overline{Y} - \hat{b}\overline{x}$: in het formularium p6

$$= \frac{40 - (7/11)(56)}{8} = \frac{6}{11} = 0,545$$

[als je $\hat{b}$ kent, kan je $\hat{a}$ berekenen]

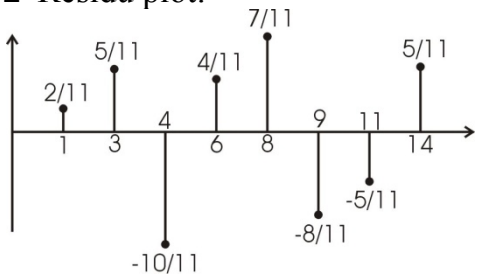
■ Kijken naar de plot van de residu's:

$$\overline{Y_i} = \hat{a} + \hat{b}x_i = \frac{6}{11} + \frac{7}{11}x_i$$

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9
$\hat{y}$	13/11	27/11	34/11	48/11	62/11	69/11	83/11	104/11

$$\sum y_i = \sum \hat{y}_i = 40$$

■ Residu plot:



→ dit is een goede residu plot

$$\sum \left(Y_i - \overline{Y}\right)^2 = \sum Y_i^2 - n\overline{Y}^2 = 256 - (8)(25) = 56$$

$$\sum \left(Y_i - \overline{Y}\right)^2 = 53,46$$

$$r^2 = \frac{53,46}{56} = 95,45\%$$

<u>Deel B:</u>

■ $X = a_0 + b_0 y$

$$\hat{b}_0 = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum y_j^2 - \left(\sum y_i{}^2\right)}$$

$$= \frac{(8)(364) - (56)(40)}{(8)(256) - 40^2}$$

$$= \frac{3}{2} = 1,50$$

■ $\hat{a}_0 = \overline{x} - \hat{b}_0\overline{y}$: in het formularium p6

$$= \frac{56 - (1,5)(40)}{8} = -0,50$$

[als we $\hat{b}$ kennen, kennen we $\hat{a}$]

Merk op:

* $\hat{Y}_i = \hat{a} + \hat{b}x_i = \frac{6}{11} + \frac{7}{11}x_i$

* $\hat{X}_i = \hat{a}_0 + \hat{b}_0 y_i = -\frac{1}{2} + \frac{3}{2}y_i$

→ dit zijn twee verschillende rechte
→ zowel intercept als helling zijn verschillend

y	1	2	4	4	5	7	8	9
x	1	3	4	6	8	9	11	14
$\hat{x}$	1	5/2	11/2	11/2	14/2	20/2	23/2	26/2

$$\sum X_i = \sum \hat{x}_i = 56$$

■ Residu plot: