DATA SCIENCE 1

Bevestigende analyse

Wouter Deketelaere

Karel de Grote Hogeschool

WAAROM

Waarom doen we dit?

Wat willen we bereiken

Wat willen we als resultaat

ANALYSE VAN GEGEVENS (DOELSTELLINGEN)	ANALYSE VAN GEGEVENS (TECHNIEKEN)
BESCHRIJVENDE ANALYSE	Wat gebeurt er?
VERKENNENDE ANALYSE	Waarom gebeurt dat?
BEVESTIGENDE ANALYSE	**Weten we zeker dat dit gebeurt**
PREDICATIEVE ANALYSE	Wat zal er gebeuren?
NORMATIEVE ANALYSE	Hoe kunnen we het waarmaken
OPERATIONEEL ONDERZOEK	Heuristische optimalisatie

Hoe doen we dit?

Welke stappen nemen we

Welke technieken gebruiken we

HOE

LEERDOELEN

Na deze les …

- weet je wat een **toevalsexperiment** is
- begrijp je wat de uitkomsten van een toevalsexperiment zijn
- weet je wat bedoeld wordt met een **gebeurtenis en een elementaire gebeurtenis**
- ken je de eigenschappen van een **kans**
- kan je een kans berekenen met de **regel van Laplace**
- kan je een kans berekenen op basis van een **kruistabel**
- weet je wat een **tegengestelde gebeurtenis** is
- weet je wat **afhankelijke en onafhankelijke** gebeurtenissen zijn
- begrijp je wat een **toevalsvariabele** is en waarom we het gebruiken
- begrijp je wat een **voorwaardelijke kans** is
- ken je de **som- en productregel** voor het berekenen van kansen
- kan je voorwaardelijke kansen berekenen
- kan je de **som- en productregel** toepassen wanneer dat nodig is

LEERDOELEN

Na deze les …

- weet je wat een **kansverdeling** is
- kan je een kanstabel of kansboom opstellen
- ken je het verschil tussen een **discrete en continue** kansverdeling
- ken je de eigenschappen van **kansverdelingen**
- begrijp je het verschil tussen **kans en kansdichtheid**
- weet je wat bedoeld wordt met een **verwachtingswaarde**
- kan je een verwachtingswaarde berekenen of laten berekenen met Orange
- weet je wat bedoeld wordt **variantie of standaardafwijking**
- kan je de variantie of standaardafwijking berekenen of laten berekenen met Orange
- weet je wat de **normale verdeling** is en waarvoor ze gebruikt wordt
- weet je wat een **z-score** is
- kan je een **z-score berekenen**
- weet je wat de **standaardnormale verdeling** is en waarvoor ze gebruikt wordt

LEERDOELEN

Na deze les …

- weet je wat **studentverdeling** is en waarvoor ze gebruikt wordt.
- begrijp je de **overeenkomst** tussen een standaardnormale en een studentverdeling
- weet je wat **chi-kwadraatverdeling** is en waarvoor ze gebruikt wordt
- weet je wat een **Fisher-verdeling** is en waarvoor ze gebruikt wordt
- weet je welke **parameters** de verschillende verdelingen hebben
- begrijp je de **invloed** van de parameters op deze verdelingen
- kan je kansen berekenen voor bovenstaande kansverdeling m.b.v. **GeoGebra**

BEVESTIGENDE ANALYSE

TOEVAL OF NIET?

- Is dit patroon toeval?
- Wat is de kans dat we dit waarnemen?
- Welke echte conclusies kunnen we hieruit trekken?

KANSREKENEN

- Grondbeginselen van waarschijnlijkheidstheorie
- Enkele wetten van waarschijnlijkheid

KANSVERDELINGEN

- Normale verdeling
- Standaard normale verdeling
- Studentverdeling
- Chi-kwadraat verdeling

STEEKPROEVEN

- Populatie vs. steekproef
- Eigenschappen van steekproeven

BETROUWBAARHEIDSINTERVALLEN

- Binnen welke grenzen vallen nieuwe metingen?

HYPOTHESETESTS

- Kunnen we beweringen verifiëren?
- Wanneer kunnen we beweringen weerleggen?

TECHNIEKEN

- Statistieken

WETEN WE ZEKER DAT HET GEBEURT?

BEVESTIGENDE ANALYSE

I

WELK PROBLEEM WILLEN WE OPLOSSEN

Kunnen we garanties geven?

II

KANSREKENEN

Wat is de kans?

III

KANSVERDELINGEN

Het grotere plaatje

IV

STEEKPROEVEN

Informatie verzamelen

V

BETROUWBAARHEID

Grenzen stellen

VI

HYPOTHESES TOETSEN

Beweringen controleren

I

WELK PROBLEEM

willen we oplossen?

UPTIME HIGH AVAILIBITY SERVER

UPTIME VAN JE SERVER

Wat is de **beschikbaarheid** van een server?

REALIABILTY: de **kans** dat het element niet faalt

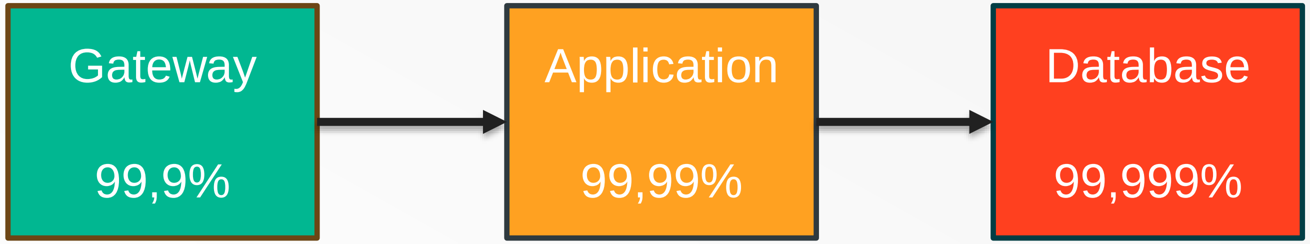
MAINTAIBILITY: de **kans** dat de server na falen succesvol hersteld werd

AVAILIBILITY: de **kans** dat een server op het gegeven moment niet faalt en niet wordt hersteld na een faling.

IT SERVICE AVAILIBILTY

De availability van een IT service hangt af van de availability van elke van de componenten.

Voorbeeld – Kubernetes Cluster



Availability Cluster = 0.999 x 0.9999 x 0.99999 = 0.9989
99.89 % = $\pm$10 uur downtime / jaar

Hoe berekenen we deze kansen?
Kunnen we dit garanderen/bewijzen?

WELK PROBLEEM WILLEN WE OPLOSSEN

Kunnen we garanties geven?

I

KANSREKENEN

Wat is de kans?

II

KANSVERDELINGEN

Het grotere plaatje

III

BEVESTIGENDE ANALYSE

IV

STEEKPROEVEN

Informatie verzamelen

V

BETROUWBAARHEID

Grenzen stellen

VI

HYPOTHESES TOETSEN

Beweringen controleren

II

KANSREKENEN

BEVESTIGENDE ANALYSE

TOEVAL OF NIET?

- Is dit patroon toeval?
- Wat is de kans dat we dit waarnemen?
- Welke echte conclusies kunnen we hieruit trekken?

KANSREKENEN

- Grondbeginselen van waarschijnlijkheidstheorie
- Enkele wetten van waarschijnlijkheid

KANSVERDELINGEN

- Normale verdeling
- Standaard normale verdeling
- Studentverdeling
- Chi-kwadraat verdeling

STEEKPROEVEN

- Populatie vs. steekproef
- Eigenschappen van steekproeven

BETROUWBAARHEIDSINTERVALLEN

- Binnen welke grenzen vallen nieuwe metingen?

HYPOTHESETESTS

- Kunnen we beweringen verifiëren?
- Wanneer kunnen we beweringen weerleggen?

TECHNIEKEN

- Statistieken

WETEN WE ZEKER DAT HET GEBEURT?

1

Wat is een kans?

WELKE KANS?

Een toevalsexperiment kan verschillende uitkomsten geven ondanks dezelfde beginsituatie.

Voorbeelden

- Gooien van een dobbelsteen
- Het IQ van een willekeurige INF1 student
- Test hoe lang de voeding van een server werkt, ...

Toch zijn er **regelmatigheden** in de resultaten als je veel metingen doet:

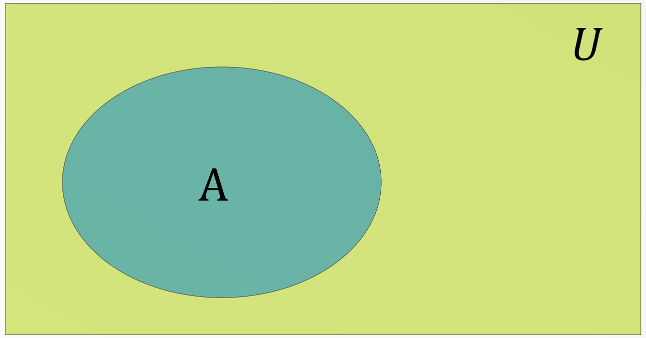
- Hoe vaak komen waarden voor? (frequenties)
- Het gemiddelde van de waarden situeert zich rond een bepaalde waarde
- De spreiding van de waarden
- ...

KANSEN MET VERZAMELINGENLEER

Wet van Laplace

- **Verzameling** U met alle mogelijke uitkomsten
- **Elementaire gebeurtenis** A_i is één uitkomst van U
- **Gebeurtenis** A is verzameling van gewenste elementaire uitkomsten

→ **Kans** $P(A)$ dat de gebeurtenis optreedt is $P(A) = \frac{\#A}{\#U}$



KANSEN MET VERZAMELINGENLEER

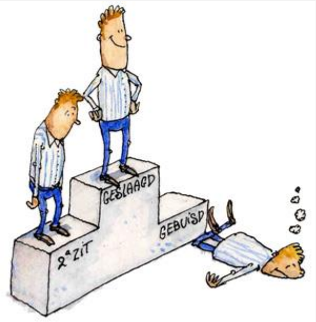
Voorbeeld

$U \ = \{ \, alle \ studenten \ INF1 \, \}$,

$\#U \ = \ 310$

$A \ = \{ \, geslaagde \ studenten \ INF1 \}$,

$\#A \ = 127$



Wat is de kans dat een willekeurig gekozen student uit U geslaagd is?

$$P(A) = \frac{\#A}{\#U} = \frac{127}{310} = 0{,}41$$

KANSEN MET VERZAMELINGENLEER

Voorbeeld

Kans om 7 ogen te gooien met 2 dobbelstenen?

$U \; = \{\, alle\ uitkomsten\ met\ 2\ dobbelstenen \,\},$

$\#U \; = 36$

$A \; = \{\, uitkomsten\ met\ 7\ ogen\},$

$\#A \; = 6$

	Gebeurtenis A	#A
2	(1,1)	1
3	(1,2); (2,1)	2
4	(1,3); (2,2); (3,1);	3
5	(1,4); (2,3), (3,2); (4,1)	4
6	(1,5); (2,4); (3,3); (4,2); (5,1)	5
7	(1,6); (2,5); (3,4); (4,3); (5,2); (6,1)	6
8	(2,6); (3,5); (4,4); (5,3); (6,2)	5
9	(3,6); (4,5); (5,4); (6,3)	4
10	(4,6); (5,5); (6,4)	3
11	(5,6); (6,5)	2
12	(6,6)	1
	#U (= TOTAAL)	**36**

- Wat is de kans om 3 ogen te gooien? $P(A) = \frac{\#A}{\#U} = \frac{2}{36} = \frac{1}{18}$
- Wat is de kans om 10 ogen te gooien? $P(A) = \frac{\#A}{\#U} = \frac{3}{36} = \frac{1}{12}$

KANSEN MET VERZAMELINGENLEER

Voorbeeld

Je kan kansen soms ook gemakkelijk aflezen uit kruistabellen.

	Wit merk (White label)	Geen wit merk (Private label)	
Slechte koeling	1498	1513	**3011**
Goede koeling	504	6485	**6989**
	2002	**7998**	**10000**

Wat is de kans dat een PC van een wit merk een slechte koeling heeft?

$U = \{ PC's\ van\ wit\ merk \},\ A = \{PC's\ van\ wit\ merk\ met\ slechte\ koeling \},$

$\#U = 2002, \#A = 1498$

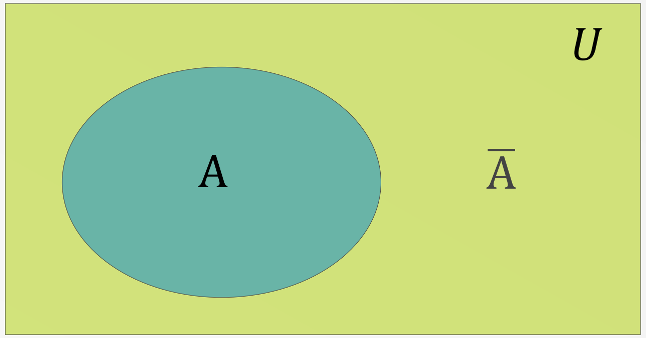
De kans is dus: $\mathrm{P(A)} = \frac{1498}{2002} = 0{,}75$

KANSEN MET VERZAMELINGENLEER

Voorbeeld

TEGENGESTELDE GEBEURTENIS

Kans op de tegengestelde gebeurtenis $\bar{A}$



De kans dat A of $\bar{A}$ optreedt is gelijk aan 1

$$P(A) + P(\bar{A}) = 1$$

De kans dat $\bar{A}$ optreedt is dus

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

Is soms eenvoudiger te bepalen/berekenen

Opmerkingen:

- wordt ook wel de complementaire gebeurtenis genoemd
- A en $\bar{A}$ zijn uitsluitende gebeurtenissen

EIGENSCHAPPEN VAN KANSEN

Een kans is een getal tussen 0 en 1

Een kans van 0 betekent dat de gebeurtenis niet zal plaatsvinden, en een kans van 1 betekent dat de gebeurtenis zeker zal plaatsvinden.

$$0 \leq P(A) \leq 1$$

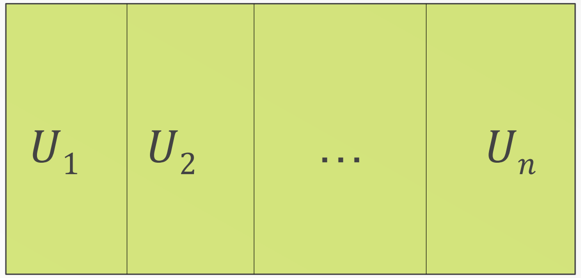
De som van de kansen van alle mogelijke uitkomsten is 1

Als $U_1, U_2, \ldots, U_n$ alle mogelijke uitkomsten zijn van een experiment, dan is

$$U_1 + U_2 + \cdots + U_n = U$$

en

$$P(U_1) + P(U_2) + \cdots + P(U_n) = 1.$$



TOEVALSVARIABELEN

Onhandige notatie

P(7 ogen te gooien met 2 dobbelstenen) → $\boxed{P(X = 7)}$

P(4 ogen of minder te gooien met 2 dobbelstenen) → $\boxed{P(X \leq 4)}$

P(meer dan 10 ogen gooien met 2 dobbelstenen) → $\boxed{P(X > 10)}$

Toevalsvariabele

Definieer een variabele X (of Y, Z) als

X = aantal ogen bij het gooien met 2 dobbelstenen

Y = product van het aantal ogen bij het gooien van 2 dobbelstenen

Z = het verschil van het aantal ogen bij het gooien van 2 dobbelstenen

TOEVALSVARIABELEN

Toevalsvariabelen hebben waarden

P(student is een meisje) of P(student is een jongen)

- Toevalsvariabele S = “biologisch geslacht van student is” met waarden $S = \{meisje, jongen\}$
- $P(S = meisje)$

P(draagt een bril)

- Toevalsvariabele B = “persoon draagt” met waarden $B = \{bril, geen\ bril\}$
- $P(B = bril)$

P(even aantal ogen bij het gooien van een dobbelsteen)

- Toevalsvariabele D = “aantal ogen” met waarden $D = \{1,2,3,4,5,6\}$
- $P(D = \{2,4,6\})$

2

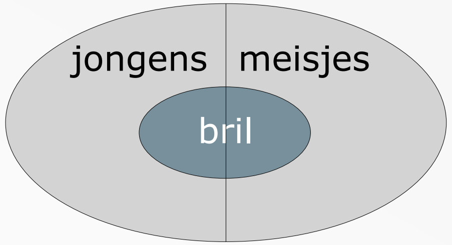
REKENEN MET KANSEN

Productregel, afhankelijke gebeurtenissen, voorwaardelijke kansen, somregel

VOORWAARDELIJKE KANS

Het optreden van gebeurtenis A kan afhankelijk zijn van het optreden van gebeurtenis B.

Voorbeeld
Stel dat jongens veel vaker een bril dragen dan meisjes.



De kans op het voorkomen van een biologisch geslacht van (A) wordt mee bepaald door de het al dan niet waarnemen van een bril B.

Het al dan niet waarnemen van een bril (B) verandert de kans op het geslacht (A).

VOORWAARDELIJKE KANS

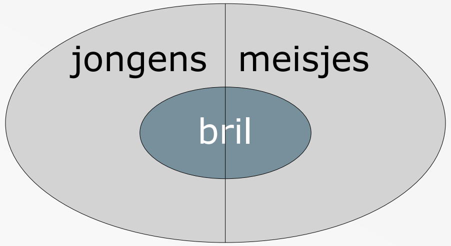
Notatie

$$P(A \mid B)$$

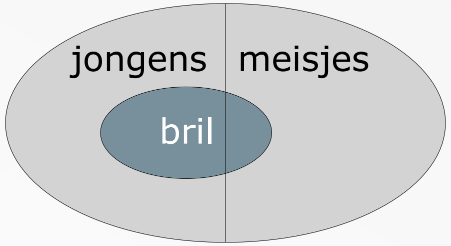
De verticale streep | lees je als gegeven, dus kans op optreden A gegeven B opgetreden is.

Voorbeelden

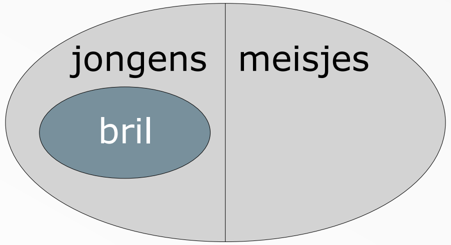
$$P(\,A = jongen \mid B \;= bril\,) = 0.5$$
$$P(\,A = meisje \mid B \;= bril\,) = 0.5$$



$$P(\,A = jongen \mid B \;= bril\,) = 0.75$$
$$P(\,A = meisje \mid B \;= bril\,) = 0.25$$



$$P(\,A = jongen \mid B \;= bril\,) = 1$$
$$P(\,A = meisje \mid B \;= bril\,) = 0$$



KANSEN MET VERZAMELINGENLEER

Voorbeeld

Je kan voorwaardelijke kansen ook gemakkelijk aflezen uit kruistabellen.

	Wit merk (White label)	Geen wit merk (Private label)	
Slechte koeling	1498	1513	**3011**
Goede koeling	504	6485	**6989**
	2002	**7998**	**10000**

Wat is de kans dat een PC van een wit merk (M=wit) een slechte koeling (K=slecht) heeft?

$$P(\,K = slecht \mid M = wit\,)$$

$$= \frac{1498}{2002}$$

2. Hoeveel binnen die kolom hebben een slechte koeling

1. Je weet al dat het wit merk is Dus enkel die kolom in rekening brengen

KANSEN MET VERZAMELINGENLEER

Voorbeeld

Je kan voorwaardelijke kansen ook gemakkelijk aflezen uit kruistabellen.

	Wit merk (White label)	Geen wit merk (Private label)	
Slechte koeling	1498	1513	**3011**
Goede koeling	504	6485	**6989**
	2002	**7998**	**10000**

Wat is de kans dat een PC met een slechte koeling (K=slecht) van een wit merk (M=wit) is?

$$P(\,M = wit \mid K = slecht\,)$$

$$= \frac{1498}{3011}$$

2. Hoeveel binnen die rij zijn van een wit merk

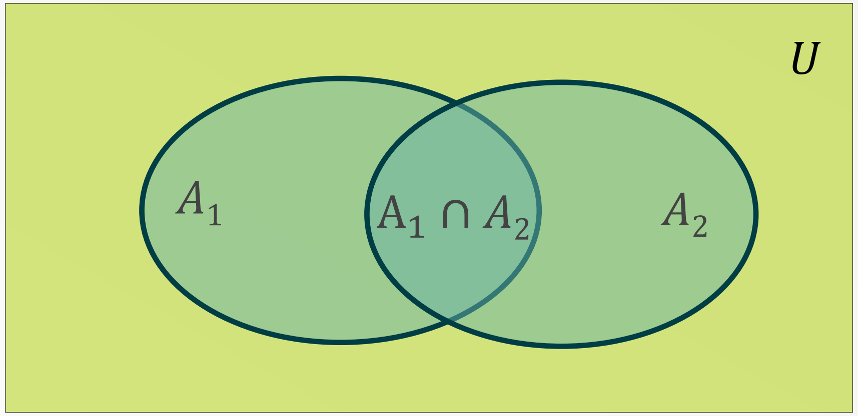
1. Je weet al dat er slechte koeling is. Dus enkel die rij in rekening brengen

PRODUCTREGEL

Stel dat we twee gebeurtenissen A_1 en A_2 hebben.
Mogelijk zijn deze gebeurtenissen afhankelijk van elkaar.

Wat is de kans dat deze gebeurtenissen tegelijk of in sequentie optreden?

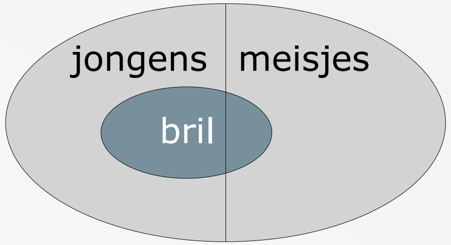
$$
\begin{aligned}
P(A) &= P(A_1 \cap A_2) \\
&= P(A_1\ EN\ A_2) \\
&= P(A_1) \cdot P(A_2 \mid A_1)
\end{aligned}
$$



PRODUCTREGEL

Voorbeeld

- kans dat iemand een bril draagt
  - $P(A_1 = bril\ ) = 0.2$
- kans dat iemand een meisje is
  - $P(A_2 = meisje\ ) = 0.5$
- kans op een meisje, gegeven dat die persoon een bril draagt
  - $P(A_2 = meisje \mid A_1 = bril) \ = \ 0.25$



- kans dat iemand een meisje met een bril is
  - $P(A_1 = bril \cap A_2 = meisje) \neq P(A_1 = bril) \cdot P(A_2 = meisje)$
    $\neq 0.2 \cdot 0.5$
  - $P(A_1 = bril \cap A_2 = meisje) = P(A_1 = bril) \cdot P(A_2 = meisje \mid A_1 = bril)$
    $= 0.2 \cdot 0.25$
    $= 0.05$

PRODUCTREGEL

Speciaal geval

Als de gebeurtenissen onafhankelijk zijn, dan geldt

$$P(A_2) = P(A_2 \mid A_1)$$

$$\downarrow$$

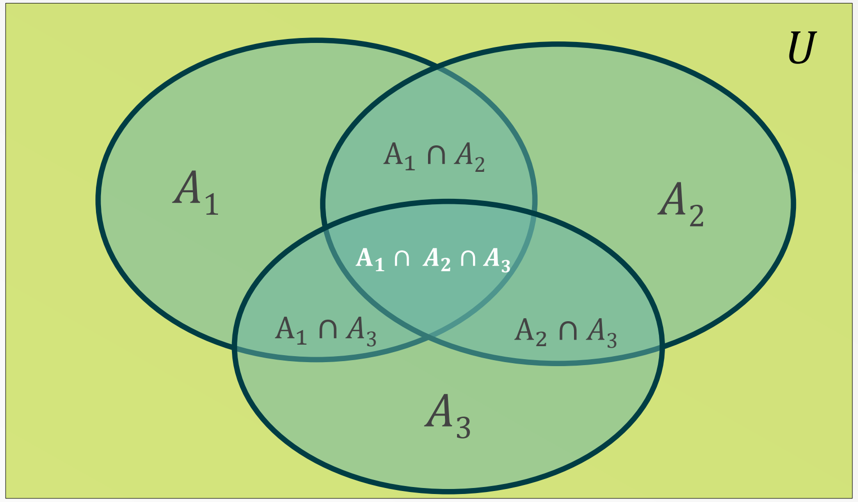
$$\begin{aligned} P(A) &= P(A_1 \cap A_2) \\ &= P(A_1) \cdot P(A_2 \mid A_1) \\ &= P(A_1) \cdot P(A_2) \end{aligned}$$

KETTINGREGEL

Stel dat we meerdere gebeurtenissen $A_1, A_2, \ldots, A_n$ hebben.
Mogelijk zijn deze gebeurtenissen afhankelijk van elkaar.

Wat is de kans dat alle gebeurtenissen tegelijk (of in sequentie) optreden?

$$
\begin{aligned}
P(A) &= P(A_1 \cap A_2 \cap \ldots \cap An) \\
&= P(A_1 \; EN \; A_2 \; EN \ldots EN \; A_n) \\
&= P(A_1) \cdot P(A_2 \mid A_1) \cdot \ldots \cdot P(A_n \mid A_{n-1}, \ldots, A_2, A_1)
\end{aligned}
$$



Venn-diagram voor 3 gebeurtenissen

Dit wordt dus snel ingewikkeld omwille van de afhankelijkheden.
We moeten al de afhankelijke kansen kennen
→ Vaak niet mogelijk om ze allemaal te kennen
→ Daarom vereenvoudigen met onafhankelijke kansen: $P(A) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)$

KETTINGREGEL

Speciaal geval

Als de gebeurtenissen onafhankelijk zijn, dan geldt:

$$P(A_2) = P(A_2 \mid A_1)$$
$$P(A_3) = P(A_3 \mid A_1, A_2)$$
$$\ldots$$
$$P(A_n) = P(A_n \mid A_{n-1}, \ldots, A_2, A_1)$$

$$\downarrow$$

$$P(A) = P(A_1 \cap A_2 \cap \ldots \cap An)$$
$$= P(A_1) \cdot P(A_2 \mid A_1) \cdot \ldots \cdot P(A_n \mid A_{n-1}, \ldots, A_2, A_1)$$
$$= P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n)$$

PRODUCTREGEL

Voorbeeld

Counter Strike

- 3 spelers
- speler 1 schiet **1 keer op 5 raak**
- speler 2 schiet **1 keer op 4 raak**
- speler 3 schiet **1 keer op 3 raak**
- één terrorist probeert door tunnel te geraken
- spelers kunnen elk maar 1 keer schieten
- **spelers beïnvloeden elkaar niet**

Wat is de kans dat de terrorist levend door de tunnel raakt?

PRODUCTREGEL

Voorbeeld

Dit is de kans dat speler 1 mist EN speler 2 mist EN speler 3 mist (A_1, A_2, A_3)

$$P(A) = P(A_1 \cap A_2 \cap A_3)$$

$$= P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_2, A_1)$$

$$= P(A_1) \cdot P(A_2) \cdot P(A_3)$$

$$= \frac{4}{5} \cdot \frac{3}{4} \cdot \frac{2}{3}$$

$$= \frac{24}{60} = 0.4$$

→ Terrorist heeft 40% kans dat hij door de tunnel raakt

SOMREGEL

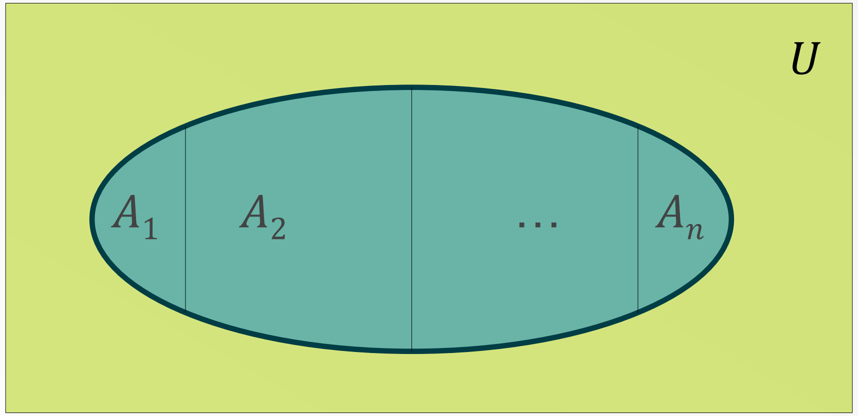
Een gebeurtenis A kan bestaan uit deelgebeurtenissen $A_1, A_2, \dots, A_n$

Stel dat de deelgebeurtenissen A_i niet overlappen, ze kunnen niet samen voorkomen, dus:

$$A_i \cap A_j = \emptyset$$
$$\forall \mathrm{i} = 1 \dots \mathrm{n}, \mathrm{j} = 1 \dots \mathrm{n} \text{ waarbij } i \neq j$$

De kans dat A, $P(A)$ optreedt is gelijk aan **som van de kans op de deelgebeurtenissen**.

$$P(A) = P(A_1 \; OF \; A_2 \; OF \dots OF \; A_n)$$
$$= P(A_1 \cup A_2 \cup \cdots \cup A_n)$$
$$= P(A_1) + P(A_2) + \dots + P(A_n)$$



SOMREGEL

Voorbeeld

Een boek kaarten, kies een kaart K.

Wat is de kans dat de kaart een K=**Aas** OF een K=**4** is?

- $P(K = Aas) = \frac{4}{52}$
- $P(K = 4) = \frac{4}{52}$

- $$P(K = Aas \cup K = 4) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = 0.154$$

ALGEMENE SOMREGEL

Stel dat twee deelgebeurtenissen overlappen dan wordt de regel:

$$P(A) = P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

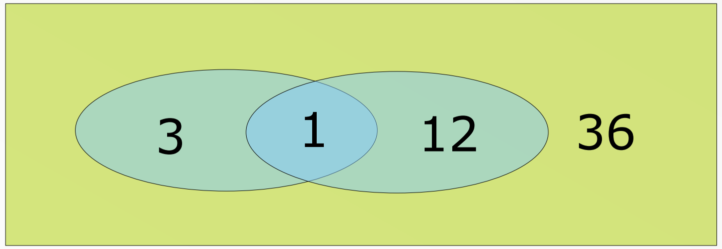
Voorbeeld

Kies één kaart K. Wat is de kans dat K=**Aas** (K_A) is of K=**Hartenkaart** (K_H)?

$K_{AH} = (\mathrm{K_A} \cup \mathrm{K_H})$ **maar** de doorsnede is niet leeg (nl. Hartenaas)

$\#K_A = 4, \;\; \#K_H = 13, \;\; \#(K_A\ en\ K_H) = 1$

$$P(K_{AH}) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = 30{,}8\%$$



ALGEMENE SOMREGEL

Stel dat drie(*) deelgebeurtenissen overlappen dan wordt de regel:

$$\begin{aligned} P(A) = P(A_1 \cup A_2 \cup A_3) = &+ P(A_1) + P(A_2) + P(A_3) \\ &- P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ &+ P(A_1 \cap A_2 \cap A_3) \end{aligned}$$

Voorbeeld

We gooien met 2 dobbelstenen. We winnen het spel als één van de dobbelstenen een 2, 3 of 4 is. Wat is de kans op deze gebeurtenis?

$P(D_1 = 2 \cup D_2 = 2) = P(D_1 = 3 \cup D_2 = 3) = P(D_1 = 4 \cup D_2 = 4) = \frac{11}{36}$ (reken dit na)

$P(D_1 = 2 \cap D_2 = 3) + P(D_1 = 3 \cap D_2 = 2) = \frac{2}{36}$ (hetzelfde voor de andere gevallen, 2 – 4 en 3 – 4)

$P(D_1 = 2 \cap D_2 = 3 \cap D_3 = 4) = 0$ (we gooien maar 2 dobbelstenen, dus kans is 0)

$$P(Win) = \frac{11}{36} + \frac{11}{36} + \frac{11}{36} - \frac{2}{36} - \frac{2}{36} - \frac{2}{36} + 0 = \frac{27}{36}$$

(*) formules voor meer dan twee overlappende deelgebeurtenissen bestaan ook.

ALGEMENE SOMREGEL

Voorbeeld

(2,3)
(3,2)
(2,4)
(4,2)
(3,4)
(4,3)
dubbel

(D1,D2)	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

(2,2)
dubbel

(3,3)
dubbel

(4,4)
dubbel

WELK PROBLEEM WILLEN WE OPLOSSEN

Kunnen we garanties geven?

I

KANSREKENEN

Wat is de kans?

II

KANSVERDELINGEN

Het grotere plaatje

III

BEVESTIGENDE ANALYSE

IV

STEEKPROEVEN

Informatie verzamelen

V

BETROUWBAARHEID

Grenzen stellen

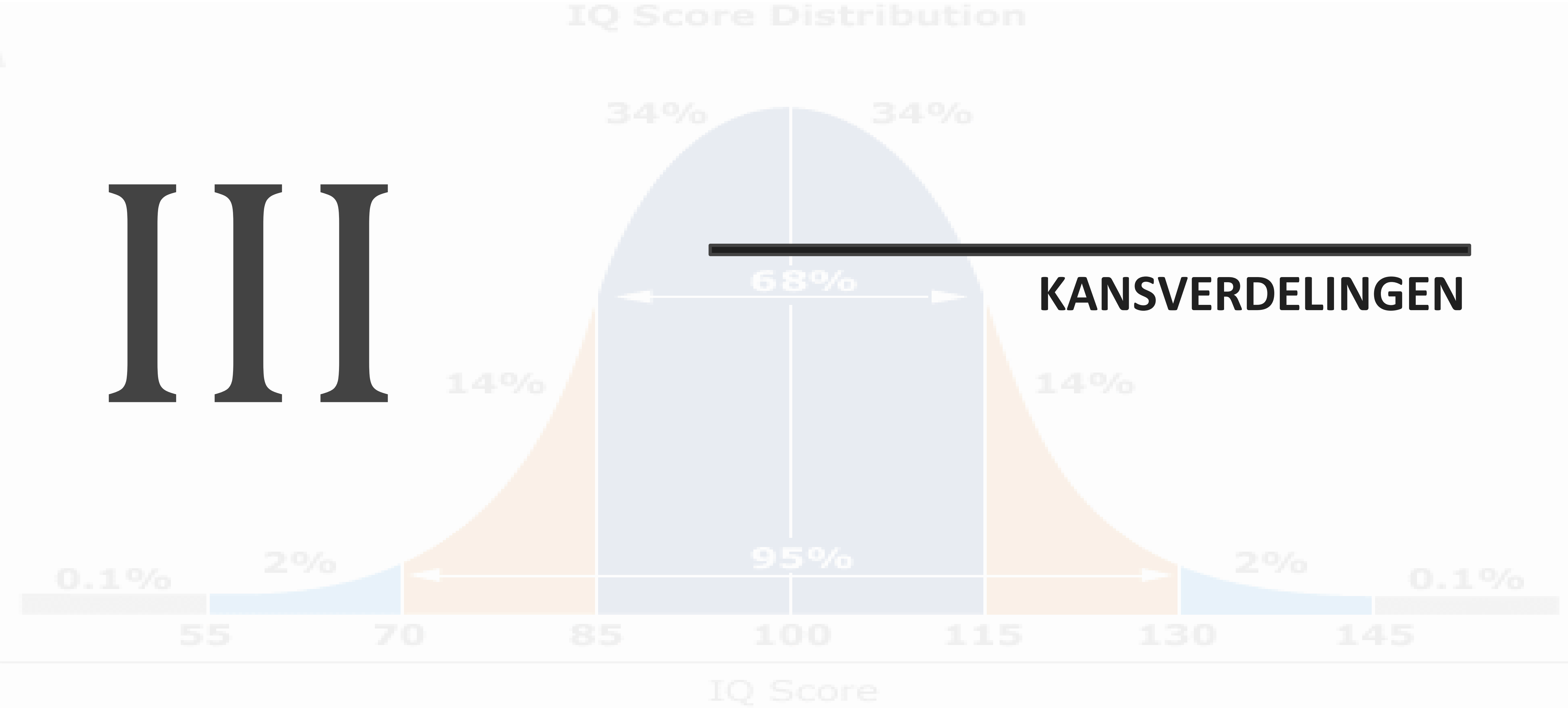
VI

HYPOTHESES TOETSEN

Beweringen controleren

III

KANSVERDELINGEN



BEVESTIGENDE ANALYSE

TOEVAL OF NIET?

- Is dit patroon toeval?
- Wat is de kans dat we dit waarnemen?
- Welke echte conclusies kunnen we hieruit trekken?

KANSREKENEN

- Grondbeginselen van waarschijnlijkheidstheorie
- Enkele wetten van waarschijnlijkheid

KANSVERDELINGEN

- Normale verdeling
- Standaard normale verdeling
- Studentverdeling
- Chi-kwadraat verdeling

STEEKPROEVEN

- Populatie vs. steekproef
- Eigenschappen van steekproeven

BETROUWBAARHEIDSINTERVALLEN

- Binnen welke grenzen vallen nieuwe metingen?

HYPOTHESETESTS

- Kunnen we beweringen verifiëren?
- Wanneer kunnen we beweringen weerleggen?

TECHNIEKEN

- Statistieken

WETEN WE ZEKER DAT HET GEBEURT?

1

Wat is een kansverdeling?

WAT IS EEN KANSVERDELING?

Kunnen we patronen vinden die schuilgaan achter toevalsexperimenten?

<table>
<tr><th>X_i</th><th>#</th><th>F_i</th><th colspan="2">f_i</th></tr>
<tr><td>2</td><td>|||</td><td>3</td><td>3/50</td><td>0,06</td></tr>
<tr><td>3</td><td>||</td><td>2</td><td>2/50</td><td>0,04</td></tr>
<tr><td>4</td><td>||||| ||</td><td>7</td><td>7/50</td><td>0,14</td></tr>
<tr><td>5</td><td>||||| |||||</td><td>10</td><td>10/50</td><td>0,20</td></tr>
<tr><td>6</td><td>||||</td><td>4</td><td>4/50</td><td>0,08</td></tr>
<tr><td>7</td><td>|||||</td><td>5</td><td>5/50</td><td>0,10</td></tr>
<tr><td>8</td><td>|||||</td><td>5</td><td>5/50</td><td>0,10</td></tr>
<tr><td>9</td><td>||||| ||</td><td>7</td><td>7/50</td><td>0,14</td></tr>
<tr><td>10</td><td>|||</td><td>3</td><td>3/50</td><td>0,06</td></tr>
<tr><td>11</td><td>||||</td><td>4</td><td>4/50</td><td>0,08</td></tr>
<tr><td>12</td><td></td><td>0</td><td>0/50</td><td>0,00</td></tr>
<tr><td></td><td></td><td>50</td><td>50/50</td><td>1,0000</td></tr>
</table>

- Wat als we een experiment (heel) veel keer zouden kunnen herhalen?
  - Tabel opstellen met de **relatieve frequenties per uitkomst.**
- Kunnen we door te **redeneren** een patroon vinden achter het experiment?
  - Tabel opstellen met alle mogelijke uitkomsten en wat de kans is op een uitkomst

→ Tabel met kansen = beperkt aantal waarden

→ Discrete kansverdeling

WAT IS EEN KANSVERDELING?

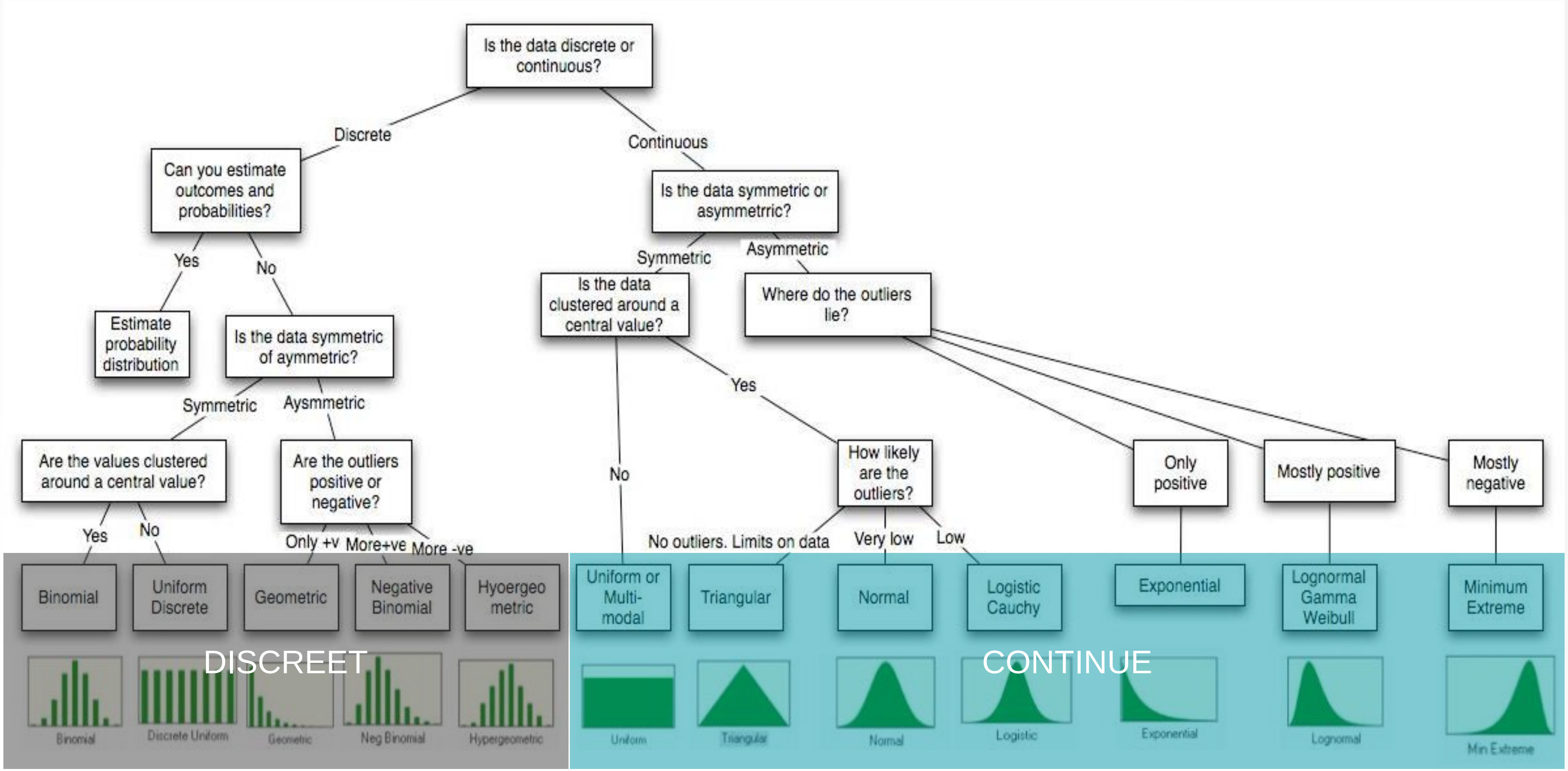
Kunnen we patronen vinden die schuilgaan achter toevalsexperimenten?

- Wat als er **teveel/oneindig** veel mogelijke uitkomsten zijn?
  - Tabel opstellen is altijd maar benadering.
  - Wiskundig model gebruiken

→ **Wiskundig model = oneindig veel waarden**

→ **Continue kansverdeling**

WAT IS EEN KANSVERDELING?



2

DISCRETE KANSVERDELINGEN

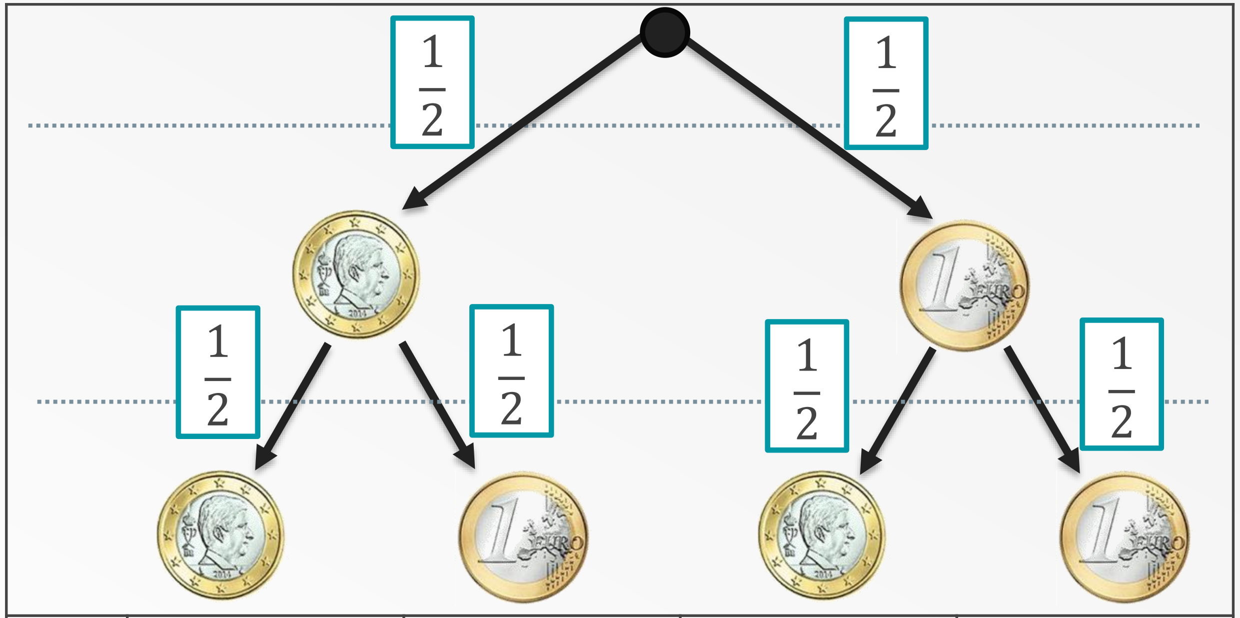
WAT IS EEN KANSVERDELING?

Toevalsexperiment

Tweemaal opgooien van een munt

uitkomst	1e munt	2e munt	#M	#K
u_1			0	2
u_2			1	1
u_3			1	1
u_4			2	0

Kansboom



#M	0	1	1	2
#K	2	1	1	0
kans	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

WAT IS EEN KANSVERDELING?

Discrete toevalsvariabele

X = Aantal keer kop wanneer we muntstuk tweemaal opwerpen

uitkomst	munt 1	munt 2	#M	#K	kans	$X(u_i) = x_i$
u_1			0	2	$\frac{1}{4}$	$X(u_1) = 2$
u_2			1	1	$\frac{1}{4}$	$X(u_2) = 1$
u_3			1	1	$\frac{1}{4}$	$X(u_3) = 1$
u_4			2	0	$\frac{1}{4}$	$X(u_4) = 0$

→

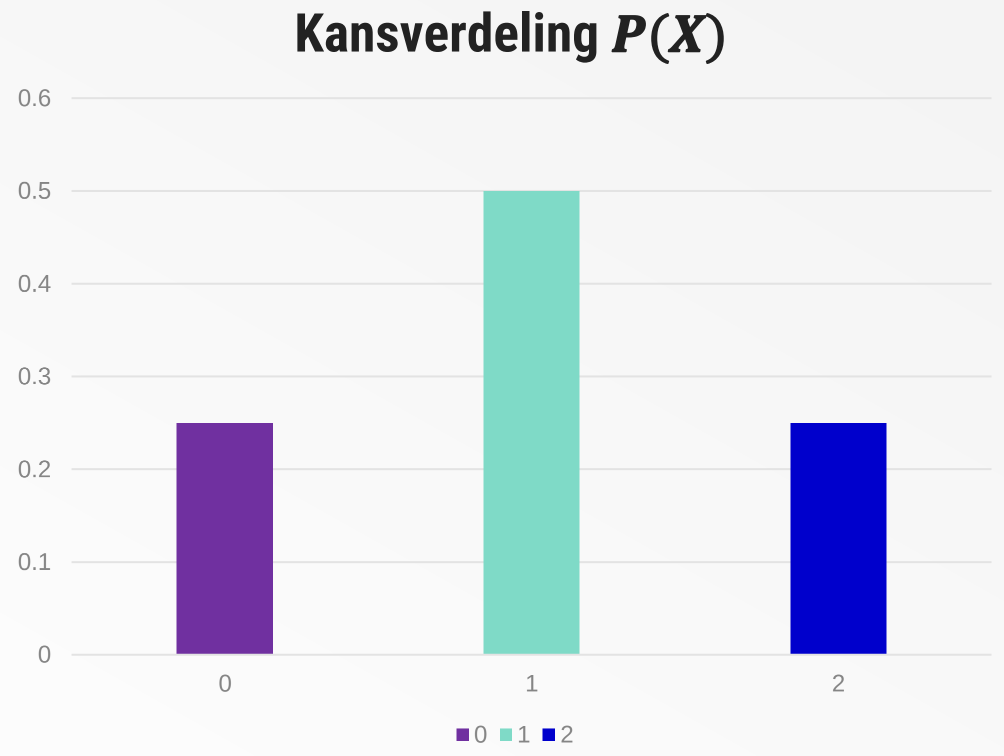
Kansverdeling $P(X)$ van X	
$X = x_i$	Kans $P(X = x_i)$
$X = 2$	$P(X = 2) = \frac{1}{4}$
$X = 1$	$P(X = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$
$X = 0$	$P(X = 0) = \frac{1}{4}$

WAT IS EEN KANSVERDELING?

Discrete toevalsveranderlijke

X = “Aantal keer kop wanneer we muntstuk tweemaal opwerpen”

Kansverdeling $P(X)$ van X	
$X = x$	Kans
$X = 2$	$P(X = 2) = \frac{1}{4}$
$X = 1$	$P(X = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$
$X = 0$	$P(X = 0) = \frac{1}{4}$

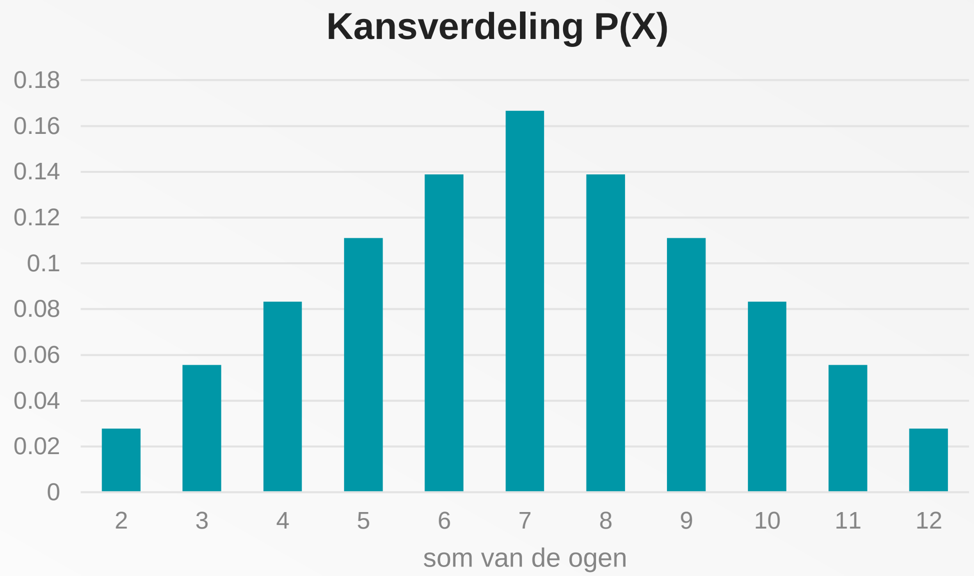


WAT IS EEN KANSVERDELING?

Discrete toevalsvariabele

X = "Som van het aantal ogen op twee dobbelstenen"

uitkomsten	#u	x_i	$P(X = xi)$	
(1,1)	1	2	1/36	0,0278
(1,2); (2,1)	2	3	2/36	0,0556
(1,3); (2,2); (3,1);	3	4	3/36	0,0833
(1,4); (2,3), (3,2); (4,1)	4	5	4/36	0,1111
(1,5); (2,4); (3,3); (4,2); (5,1)	5	6	5/36	0,1389
(1,6); (2,5); (3,4); (4,3); (5,2); (6,1)	6	7	6/36	0,1667
(2,6); (3,5); (4,4); (5,3); (6,2)	5	8	5/36	0,1389
(3,6); (4,5); (5,4); (6,3)	4	9	4/36	0,1111
(4,6); (5,5); (6,4)	3	10	3/36	0,0833
(5,6); (6,5)	2	11	2/36	0,0556
(6,6)	1	12	1/36	0,0278
#U (= TOTAAL)	**36**	**36**	**36/36**	**1,0000**



Eigenschappen van kansverdelingen

EIGENSCHAPPEN VAN KANSVERDELINGEN

Kansen en relatieve frequenties zijn gelijkaardig

- Zijn er dan een gemiddelde, mediaan, modus?

  → centrummaten van een kansverdeling

- Is er een IQR, variantie, standaardafwijking

  → spreidingsmaten van een kansverdeling

CENTRUMMATEN

Gemiddelde van kansverdeling X heet de verwachtingswaarde (expectation)

Notatie

$\mathbb{E}[X]$ of μ

Definitie

$$\mu = \mathbb{E}[X] = \sum_{x_i} x_i \cdot \mathrm{P}(X = x_i)$$

verwachtingswaarde = som van (waarde · kans op die waarde)

Voor het gooien met 1 dobbelsteen is dat dus:

$$\mu = \sum_{x_i} x_i \cdot P(X = x_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$$

gemiddeld gooi dus 3.5 met 1 dobbelsteen

Voor het gooien met 2 dobbelstenen is dat dus:

$$\mu = \sum_{x_i} x_i \cdot P(X = xi) = 2 \cdot \frac{1}{36} + \ldots + 7 \cdot \frac{6}{36} + \ldots + 12 \cdot \frac{1}{36} = 7$$

gemiddeld gooi dus 7 met 2 dobbelstenen

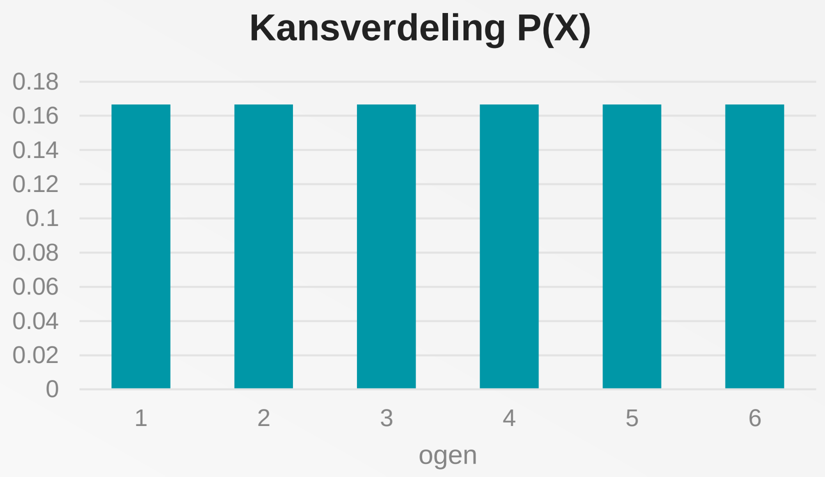
CENTRUMMATEN

Voorbeeld

Voor het gooien met 1 dobbelsteen is dat dus:

$$\mu = \sum_{x_i} x_i \cdot P(X = x_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$$

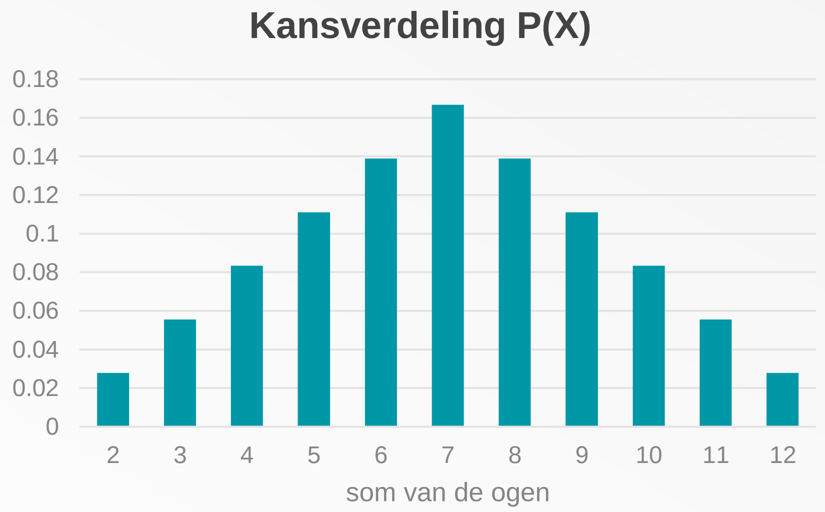
gemiddeld gooi je dus 3.5 met 1 dobbelsteen



Voor het gooien met 2 dobbelstenen is dat dus:

$$\mu = \sum_{x_i} x_i \cdot P(X = xi) = 2 \cdot \frac{1}{36} + \ldots + 7 \cdot \frac{6}{36} + \ldots + 12 \cdot \frac{1}{36} = 7$$

gemiddeld gooi je dus 7 met 2 dobbelstenen



CENTRUMMATEN

Voorbeeld

Multiplechoicevraag met giscorrectie

- keuze uit 4 antwoorden
  - juist antwoord = 1 punt
  - fout antwoord = $-\frac{1}{4}$ punt
- kans op een goed antwoord is p

score	kans
1	p
$-\frac{1}{4}$	$1-p$

Stel dat je er niks van kent.

Wat is p dan? $\boxed{p = \frac{1}{4}}$

Wat is dan de verwachte score? $\boxed{\mu = 1 \cdot \frac{1}{4} - \frac{1}{4} \cdot \frac{3}{4} = \frac{1}{16}}$

SPREIDINGSMATEN

Spreiding rond het gemiddelde van kansverdeling X heet de variantie.

Notatie

$\mathbb{E}[(X-\mu)^2]$ of σ^2

Definitie

$$\sigma^2 = \mathbb{E}[(X-\mu)^2] = \sum_{x_i} (x_i - \mu)^2 \cdot \mathrm{P}(X = x_i)$$

$$\sigma = \sqrt{\sigma^2}$$

CENTRUMMATEN

Voorbeeld

Voor het gooien met 1 dobbelsteen is dat dus:

$$\sigma^2 = \sum_{x_i} (x_i - \mu)^2 \cdot P(X = x_i) = (1 - 3.5)^2 \cdot \frac{1}{6} + \cdots + (6 - 3.5)^2 \cdot \frac{1}{6} = \frac{35}{12}$$

De spreiding rond het gemiddelde (de variantie) is dus gelijk aan 2.912

Voor het gooien met 2 dobbelstenen is dat dus:

$$\sigma^2 = \sum_{x_i} (x_i - \mu)^2 \cdot P(X = x_i) = (2 - 7)^2 \cdot \frac{1}{36} + \ldots + (7 - 7)^2 \cdot \frac{6}{36} + \ldots + (12 - 7)^2 \cdot \frac{1}{36} = 5{,}8333$$

De spreiding rond het gemiddelde (de variantie) is dus gelijk aan 5.83

4

CONTINUE KANSVERDELINGEN

DE NORMALE VERDELING

Toevalsexperiment
We selecteren willekeurig een mannelijke student van 20 jaar op KdG.

X = De lengte van de student

X is een continue toevalsvariabele met $\mu = 180 \text{ cm}$ en $\sigma = 10 \text{ cm}$

- Wat is de kans dat hij kleiner is dan 140 cm?
$$P(X < 140)$$
- Wat is de kans hij minstens 180 cm groot is?
$$P(X \geq 180)$$
- Wat is de dat hij exact 160 cm groot is?
$$P(X = 160)$$

DE NORMALE VERDELING

X is normaal verdeeld met verwachtingswaarde $\mu = 180$ en standaardafwijking $\sigma = 10$

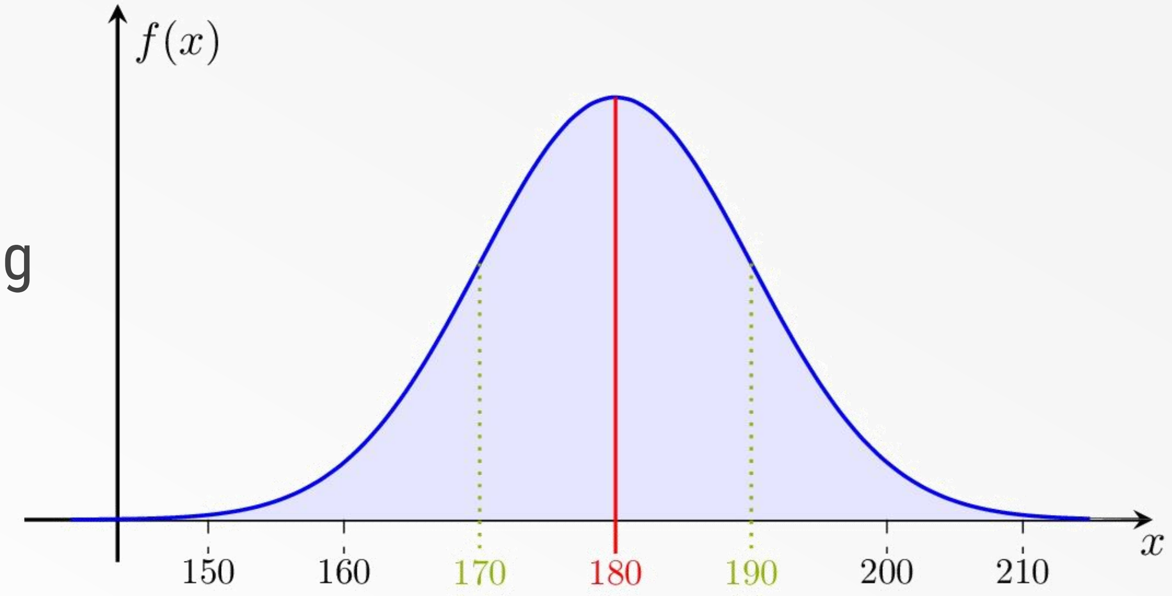
Notatie

$X \sim N(180, 10)$

Definitie van $N(180,10)$

180, 10 en $f(x)$ bepalen uitzicht van verdeling

$$f(x) = \frac{1}{10 \cdot \sqrt{2 \cdot \pi}} e^{\left(\frac{-(x-180)^2}{2 \cdot 10^2}\right)}$$



$f(x)$ is de **kansdichtheid**, niet de kans!

DE NORMALE VERDELING

Toevalsvariabele X is normaal verdeeld met verwachtingswaarde μ en standaardafwijking σ

Notatie

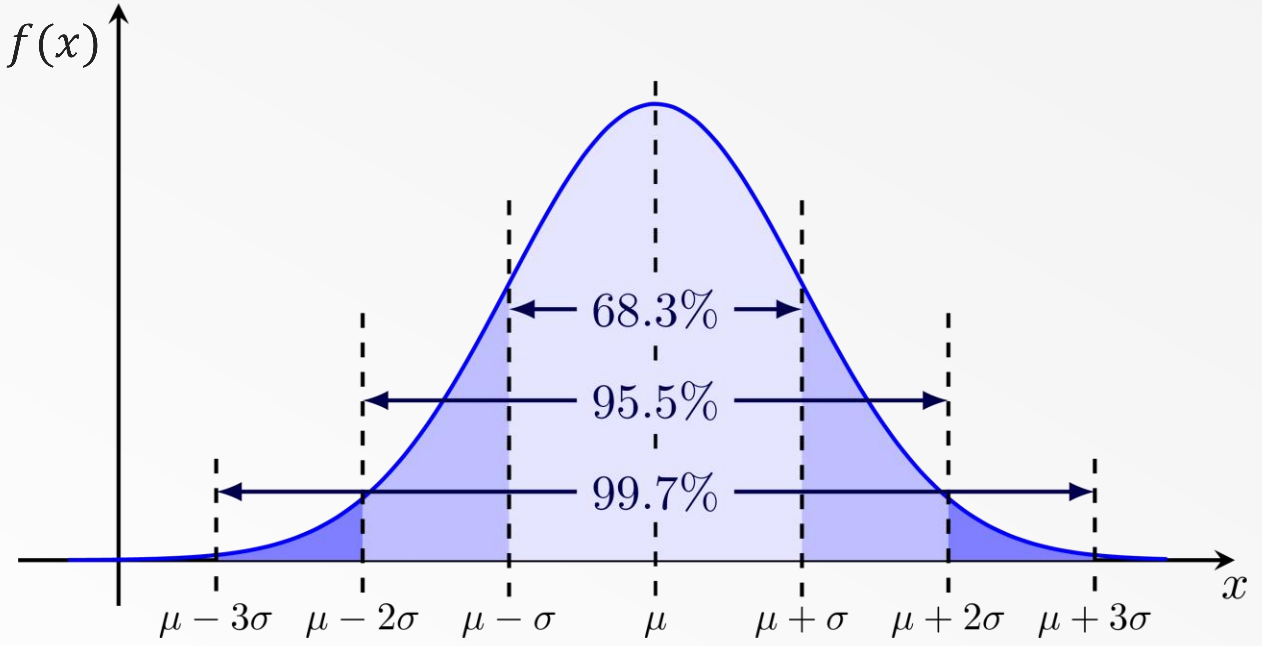
$X \sim N(\mu, \sigma)$

Definitie van $N(\mu, \sigma)$

μ, σ en $f(x)$ bepalen uitzicht van verdeling

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

$f(x)$ is de **kansdichtheid**, niet de kans!



DE NORMALE VERDELING

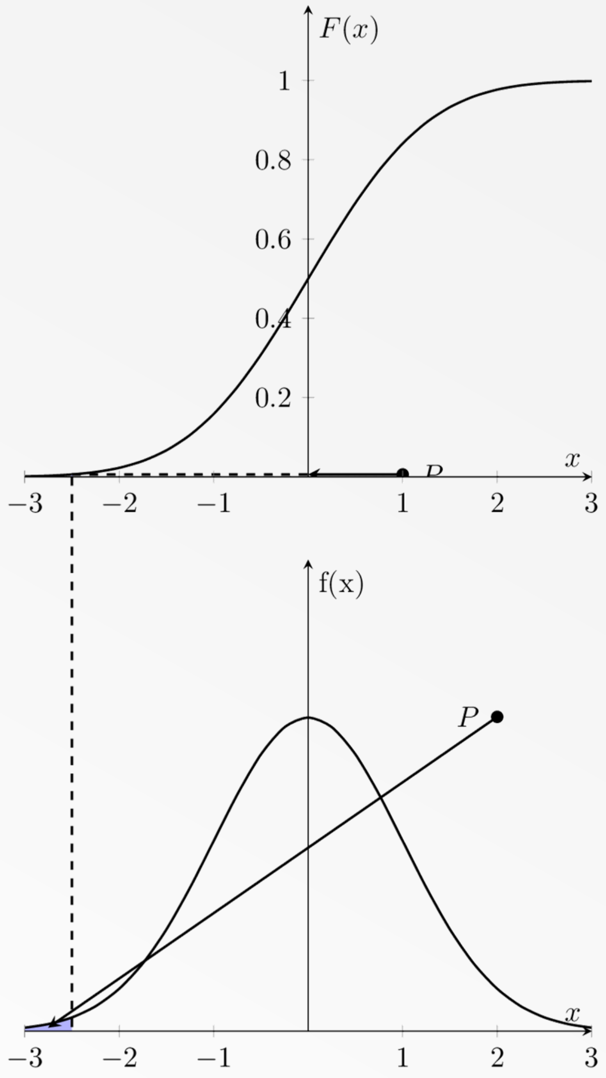
Kansdichtheid vs. kans

$f(x)$ is de **kansdichtheidsfunctie**

$F(x)$ is de cumulatieve **kansfunctie.**

De oppervlakte onder $f(x)$ is gelijk aan de

overeenkomstige waarde van $F(x)$.



DE NORMALE VERDELING

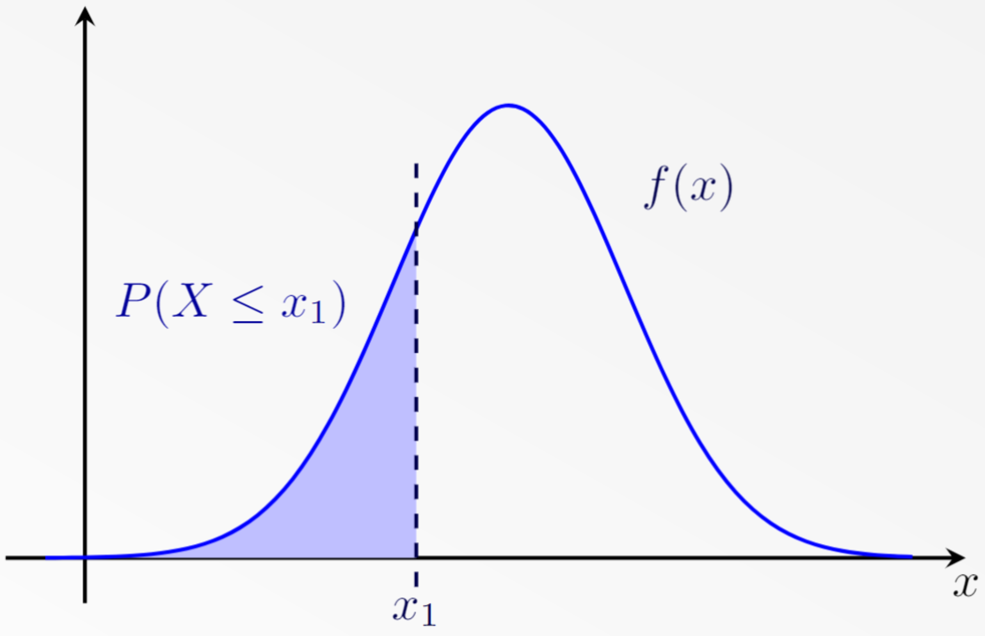
Kansen berekenen

De kans dat X kleiner is dan x_1 is

de oppervlakte onder de normaalcurve

tussen $-\infty$ en x_1.

$$\begin{aligned} P(X \leq x_1) &= P(X \leq x_1) - P(X \leq -\infty) \\ &= P(X \leq x_1) - 0 \end{aligned}$$



DE NORMALE VERDELING

Kansen berekenen

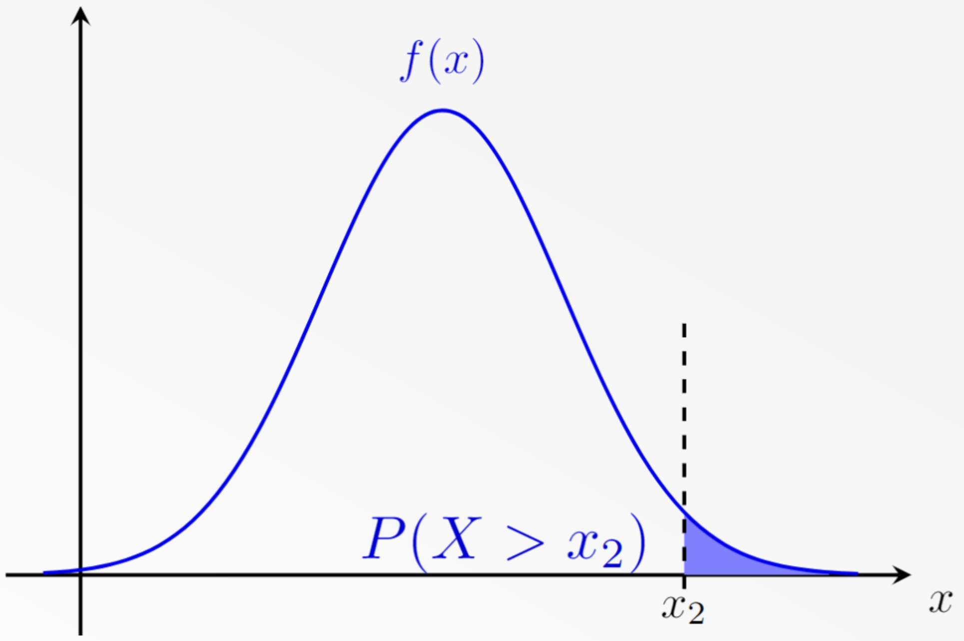
De kans dat X groter is dan x_2 is

de 1 - oppervlakte onder de normaalcurve

tussen $-\infty$ en x_2.

$$P(X > x_2) = 1 - P(X \leq x_2)$$

→ complementaire kans



DE NORMALE VERDELING

Kansen berekenen

De kans dat X tussen x_1 en x_2 ligt is de oppervlakte onder de normaalcurve tussen x_1 en x_2.

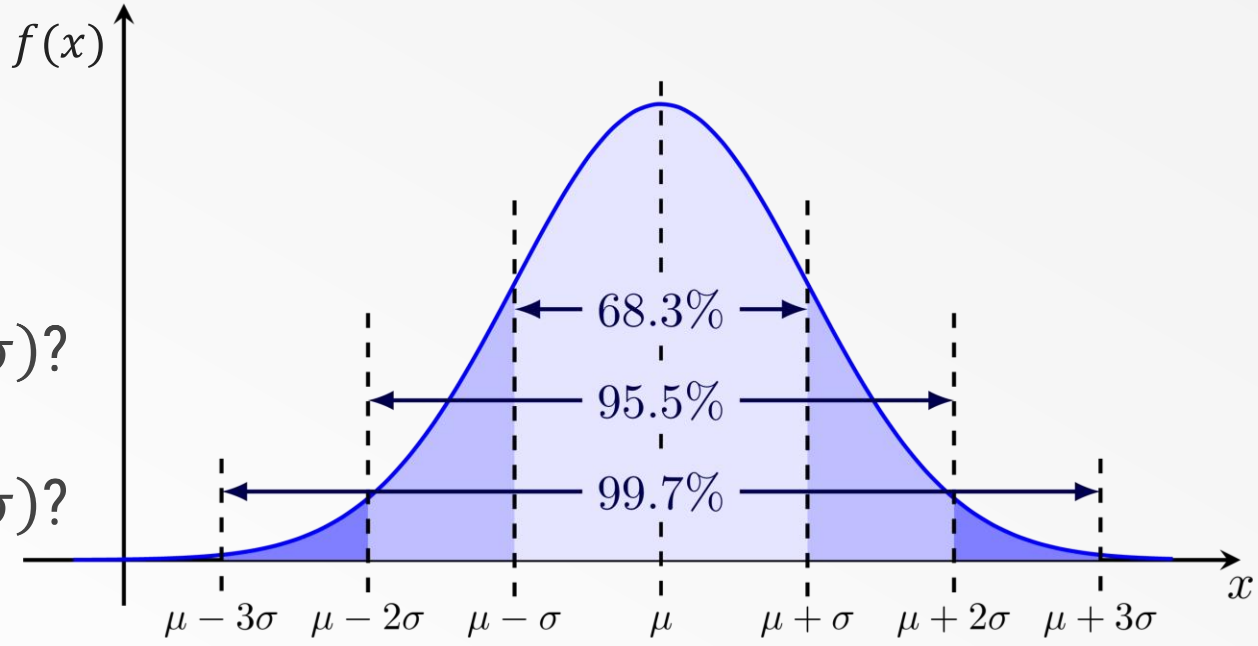
Wat is de kans dat $P(X \leq \mu)$?

Wat is de kans dat $P(\mu - \sigma < X \leq \mu + \sigma)$?

Wat is de kans dat $P(\mu - 2\sigma < X \leq \mu + 2\sigma)$?

Wat is de kans dat $P(\mu - 3\sigma < X \leq \mu + 3\sigma)$?

Wat is de kans dat $P(-\infty < X < +\infty)$?



DE NORMALE VERDELING

Voorbeeld

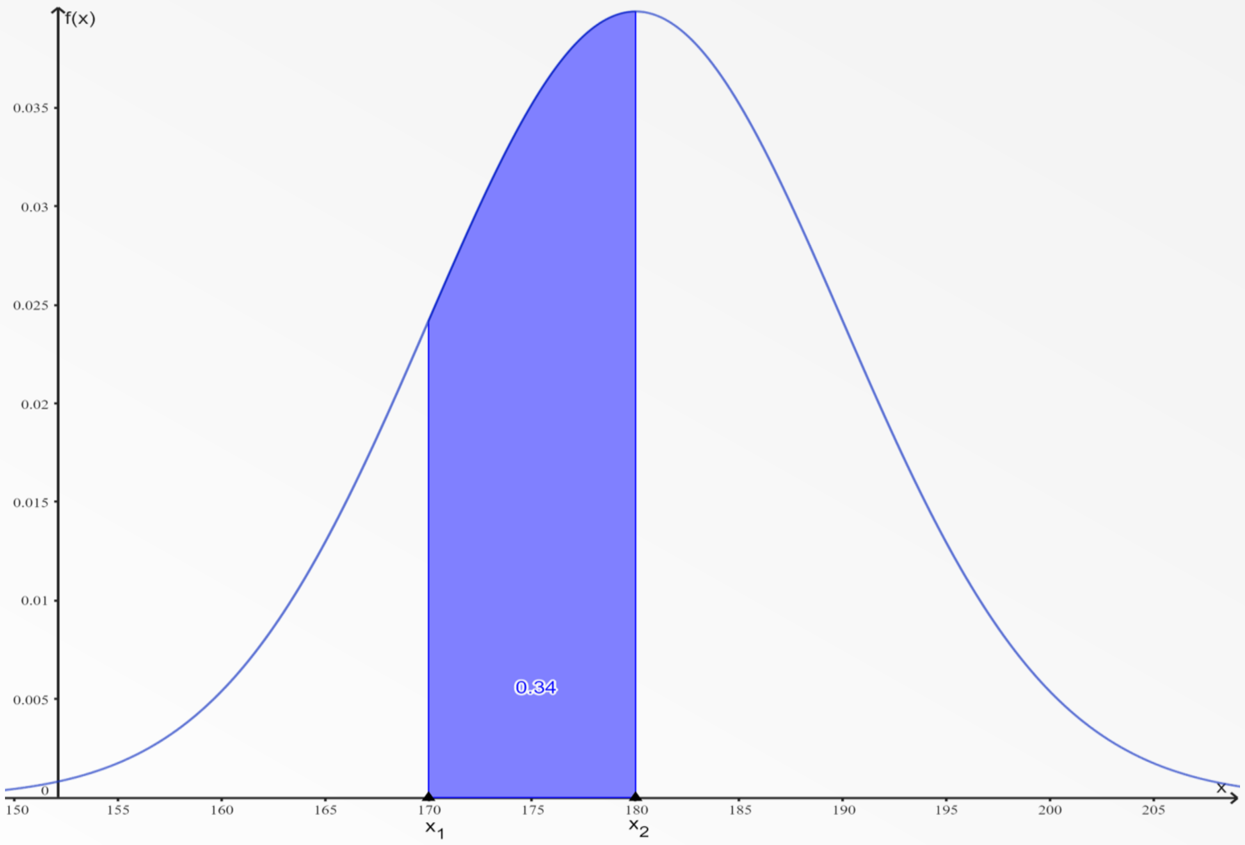
X = de lengte van een mannelijke KdG-student van 20 jaar

$X \sim N(180,10)$

Wat is de kans dat deze student tussen de 170 en 180cm groot is?

$\boxed{P(170 < X < 180) = 0.34}$

$\boxed{P(X < 180) - P(X < 170) = 0.34}$



DE STANDAARDNORMALE VERDELING

Toevalsexperiment
We selecteren opnieuw willekeurig een mannelijke student van 20 jaar op KdG.

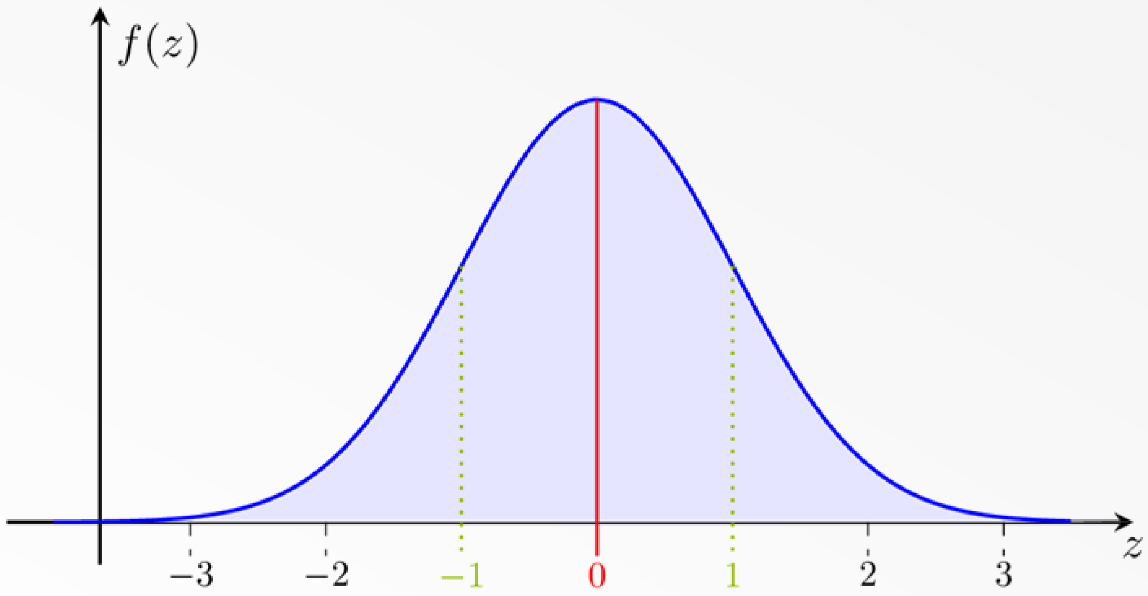
X = De lengte van de student

X is een continue toevalsvariabele met $\mu = 180$ cm en $\sigma = 10$ cm

Nieuwe toevalsvariabele

$$Z = \frac{X - \mu}{\sigma}$$

Z is een continue toevalsvariabele met
$\mu = 0$ cm en $\sigma = 1$ cm



DE STANDAARDNORMALE VERDELING

Voorbeeld

X = de lengte van een mannelijke KdG-student van 20 jaar

$X \sim N(180, 10)$

$Z \sim N(0, 1)$

Wat is de z-score van indien de student 170cm is?

$$z = \frac{x-\mu}{\sigma} = \frac{170-180}{10} = -1$$

Wat is de z-score van indien de student 180cm is?

$$z = \frac{x-\mu}{\sigma} = \frac{180-180}{10} = 0$$

Wat is de z-score van indien de student 210cm is?

$$z = \frac{x-\mu}{\sigma} = \frac{210-180}{10} = 3$$

Wat is jouw z-score?

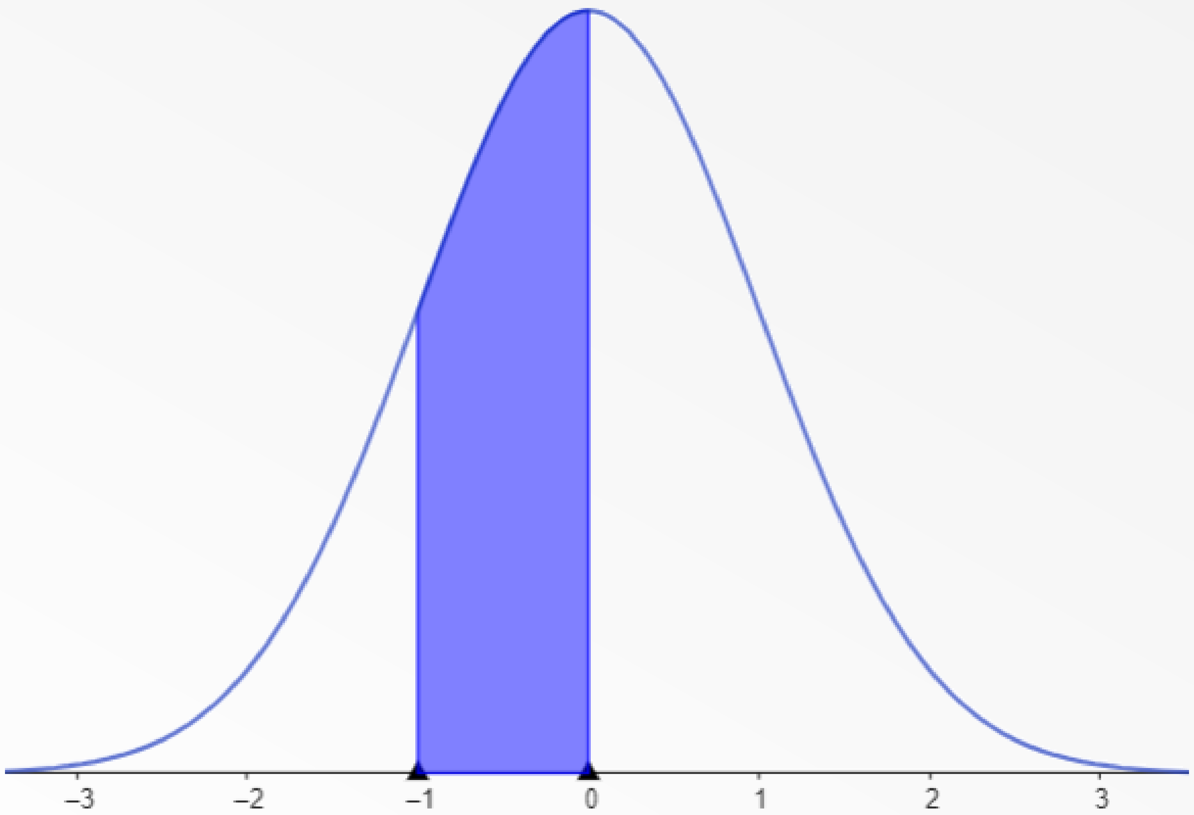
DE STANDAARDNORMALE VERDELING

Voorbeeld
X = de lengte van een mannelijke KdG-student van 20 jaar
$X \sim N(180, 10)$
$Z \sim N(0, 1)$

Wat is de kans dat deze student tussen een Z-score heeft tussen -1 en 0 groot is?

$\boxed{P(-1 < Z < 0) = 0.34}$

$\boxed{P(Z < 0) - P(Z < -1) = 0.34}$



DE χ^2- VERDELING

Voorbeeld
We doen een experiment waarbij we de eerlijkheid van een munt willen controleren.
We werpen het muntstuk 100 keer op.

categorie	# geobserveerd	# verwacht
	45	50
	55	50

Wat is de kans dat we 45 - 55 observeren als we 50-50 verwachten?

DE χ^2- VERDELING

Nieuwe toevalsvariabelen

O_i = Waargenomen frequentie van categorie i

E_i = Verwachte frequentie van categorie i

$Z_i = \frac{O_i - E_i}{\sqrt{E_i}}$ (genormaliseerde frequentie van categorie i)

$Z_i^2 = \frac{(O_i - E_i)^2}{E_i}$ (kwadrateren)

$Q = Z_1^2 + Z_2^2 + \cdots + Z_k^2$ (sommeren)

$Q = \sum_{i=1}^{k} Z_i^2$

$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ (maat voor de afwijking)

DE χ^2- VERDELING

Voorbeeld

We doen een experiment waarbij we de eerlijkheid van een munt willen controleren.
We werpen het muntstuk 100 keer op.

categorie	# geobserveerd	# verwacht
	45	50
	55	50

Wat is de kans dat we 45 - 55 observeren als we 50-50 verwachten?

DE χ^2- VERDELING

Hoe is Q verdeeld?

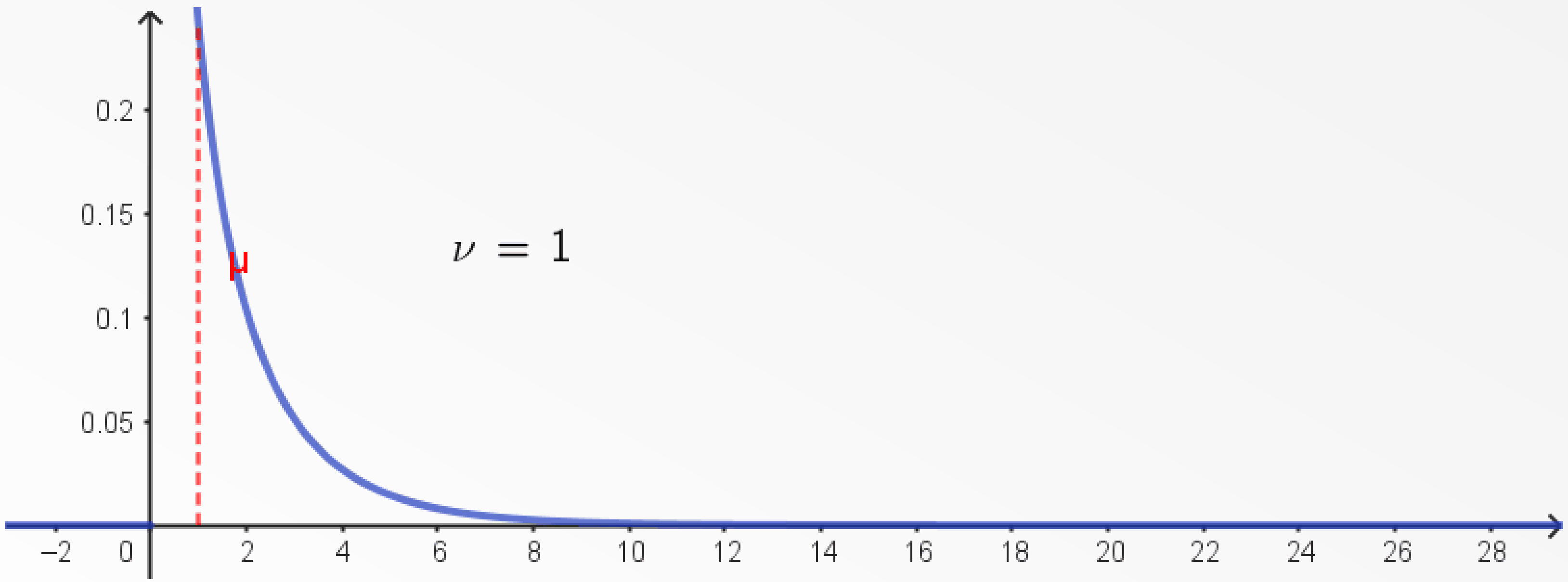
$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ is Chi-kwadraat verdeeld.

Notatie

$Q \sim \chi^2(\nu)$

Definitie van $\chi^2(\nu)$

ν en $f(x)$ bepalen uitzicht van verdeling



$\mathbb{E}[Q] = \mu = \nu$

$\mathbb{E}[(Q - E[Q])^2] = 2 \cdot k$

DE χ^2- VERDELING

Voorbeeld

Muntstuk dat we 100 keer opgooien.

$\rightarrow k = 2$

$\rightarrow Q = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = \frac{(45-50)^2}{50} + \frac{(55-50)^2}{50}$

$\rightarrow Q = \frac{25}{50} + \frac{25}{50} = 1$

$\rightarrow \nu = (\#rijen\ - 1) \cdot (\#kolommen\ - 1)$

$\rightarrow \nu = (2-1) \cdot (2-1) = 1$

Kans dat we dit resultaat uitkomen of nog uitzonderlijker

$\rightarrow P(\chi^2(\nu) > Q) = 1\ - P(\chi^2(\nu) \leq Q)$

$\rightarrow P(\chi^2(\nu) > Q) = 0.3172$

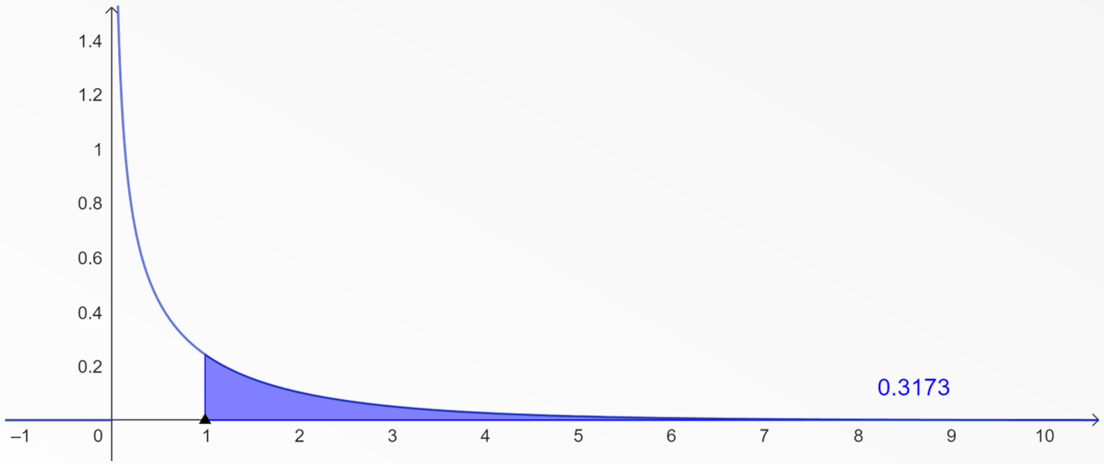
DE χ^2- VERDELING

Voorbeeld

We doen een experiment waarbij we de eerlijkheid van een munt willen controleren.
We werpen het muntstuk 100 keer op.

categorie	# geobserveerd	# verwacht
	45	50
	55	50

Wat is de kans dat we 45 - 55 observeren als we 50-50 verwachten?



DE F- VERDELING

Voorbeeld
We hebben de inhoud van 5 bierflesjes van machines A en 8 flesjes van machine B en vinden volgende standaardafwijkingen.

machine	n	s
A	$n_1 = 5$	$s_1 = 2.2$
B	$n_2 = 8$	$s_2 = 1.1$

Wat is de kans dat we dit verschil in standaardafwijking tegenkomen?

DE F- VERDELING

Nieuwe toevalsvariabelen

S_1^2 en S_2^2 : waargenomen varianties

We nemen nu de verhouding van deze

$$Y = \frac{S_1^2}{S_2^2}$$

Hoe is Y verdeeld?

Y is verdeeld volgens Fisher-verdeling

Notatie

$Y \sim F(\alpha, \beta)$

met α, β vrijheidsgraden

DE F- VERDELING

Voorbeeld

We hebben de inhoud van 5 bierflesjes van machines A en 8 flesjes van machine B en vinden volgende standaardafwijkingen.

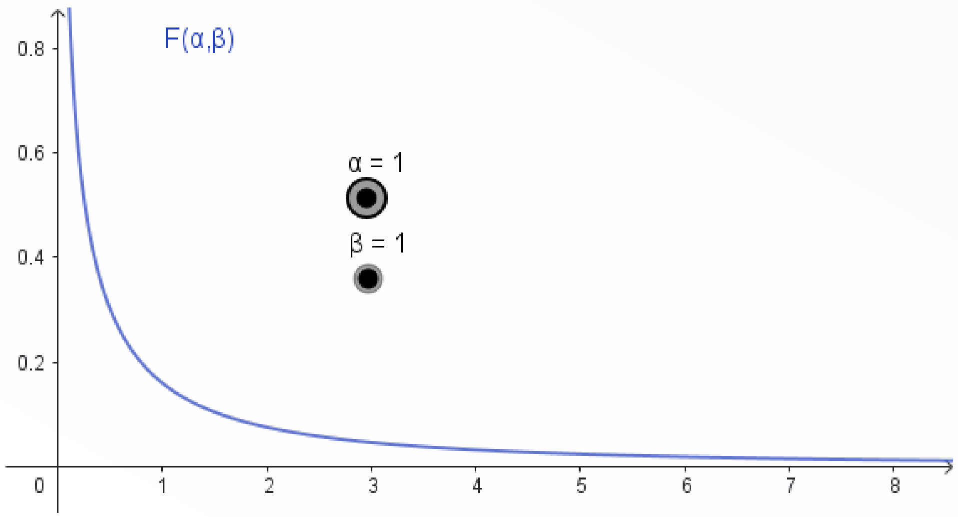
machine	n	s
A	$n_1 = 5$	$s_1 = 2.2$
B	$n_2 = 8$	$s_2 = 1.1$

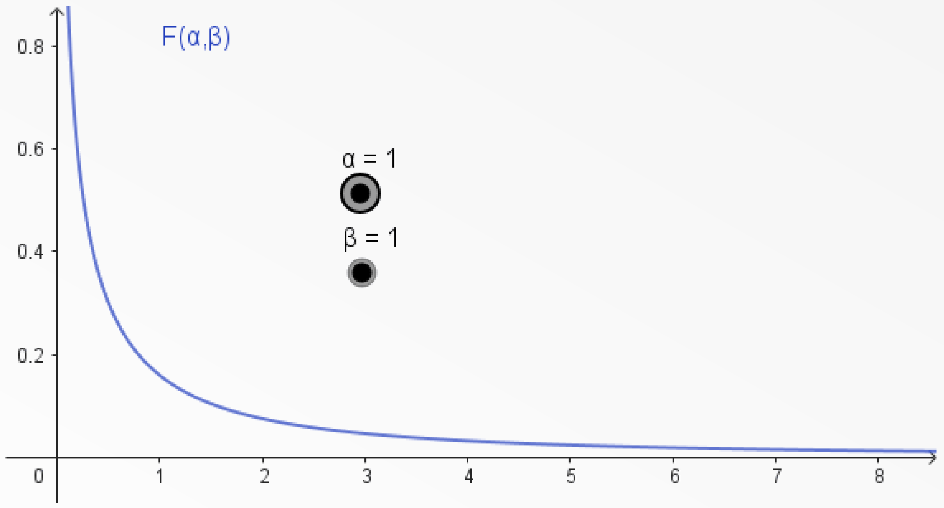
Wat is de kans dat we dit verschil in standaardafwijking tegenkomen?

DE F- VERDELING

α en β bepalen uitzicht van deze verdeling

$\alpha = n_1 - 1$ en $\beta = n_2 - 1$, bepaald door de steekproefgroottes.





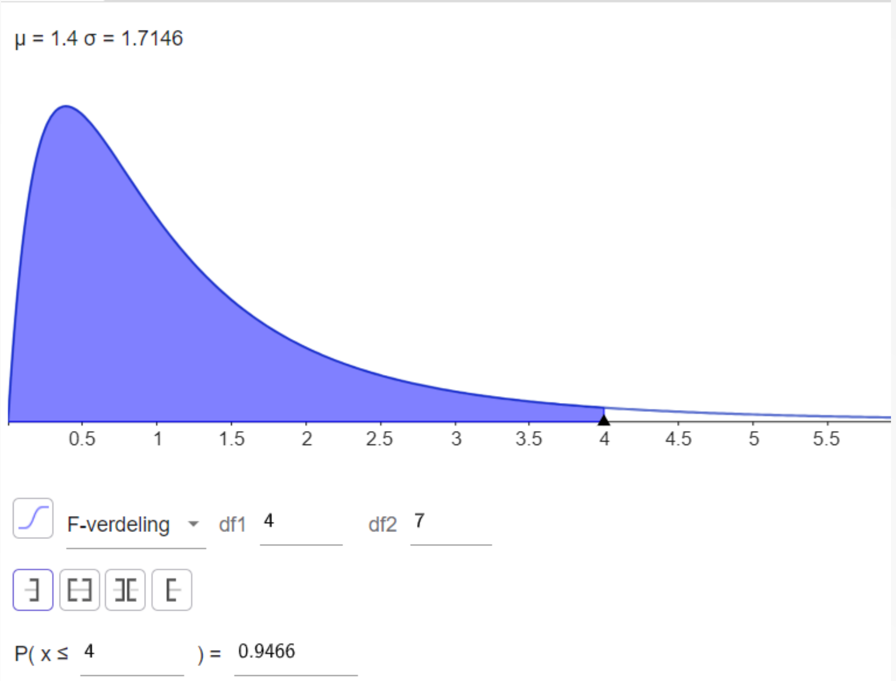
DE F- VERDELING

Voorbeeld
Wat is de kans dat dit verschil in standaardafwijking tegenkomen?

Oplossing

$$y = \frac{2.2^2}{1.1^2} = 4 \text{ (grootste waarde altijd als teller)}$$

$$\begin{aligned} P(F > y) &= 1 - P(F < y) \\ &= 1 - P(F < 4) \\ &= 1 - 0{,}9466 \\ &= 0{,}0534 \end{aligned}$$



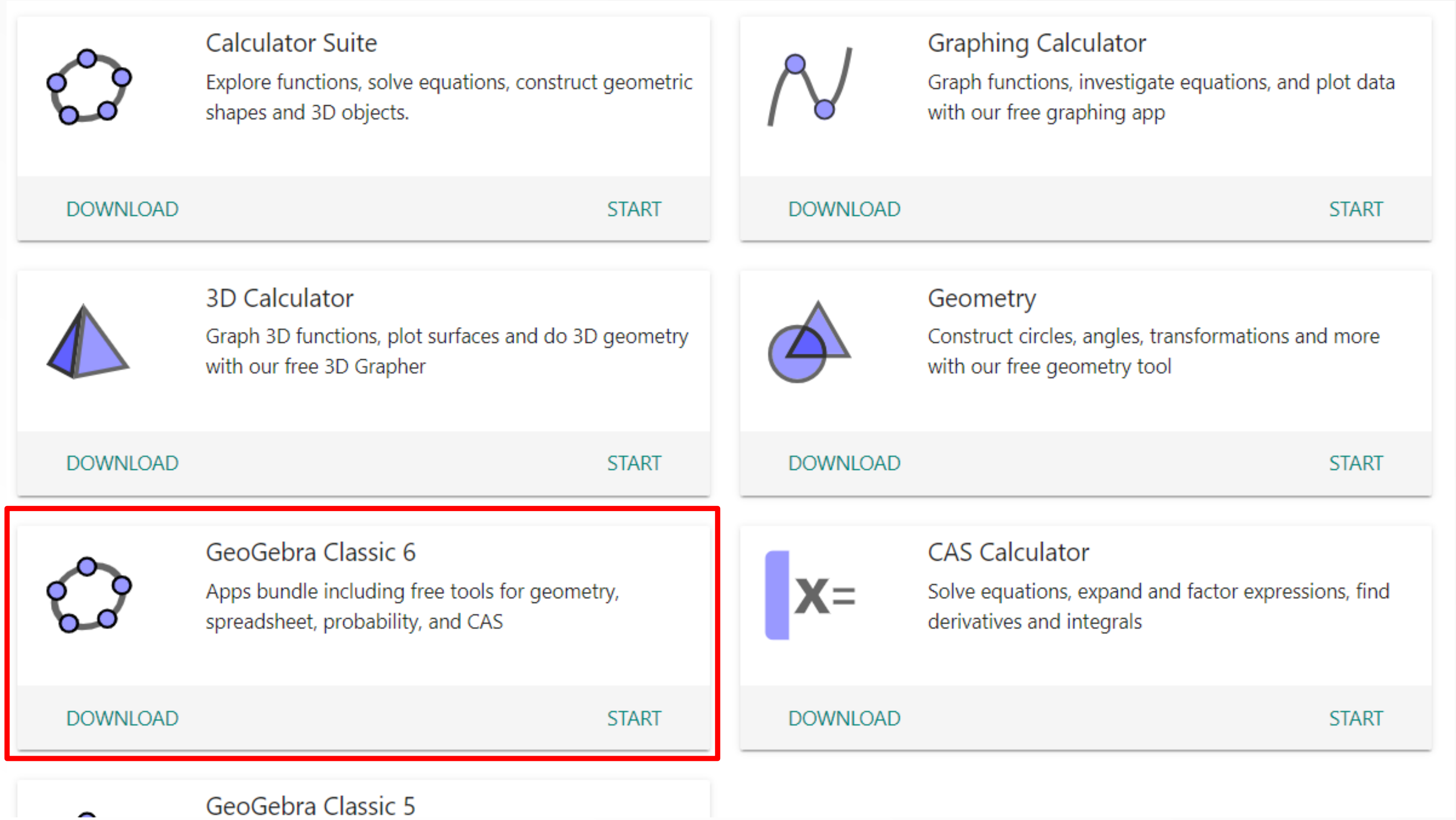
Praktijk

5

GEOGEBRA

Installeren

https://www.geogebra.org/download



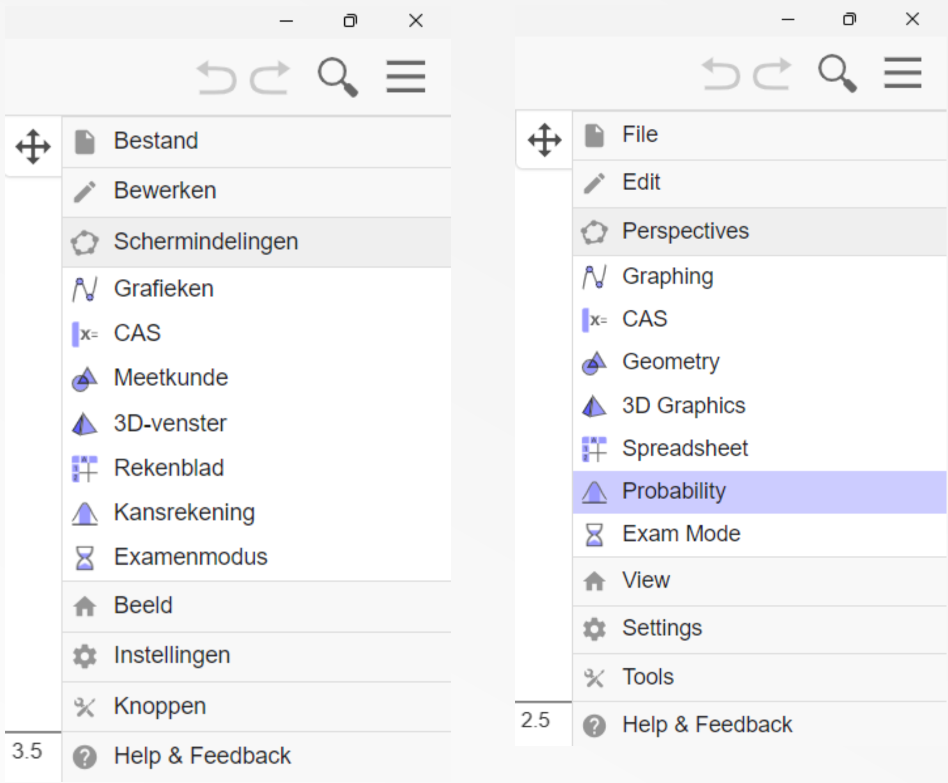
GEOGEBRA

Schermindelingen/Perspectives

Via

- Schermindelingen → Kansrekenen of
- Perspectives → Probability

overschakelen naar scherm voor **kansverdelingen**.



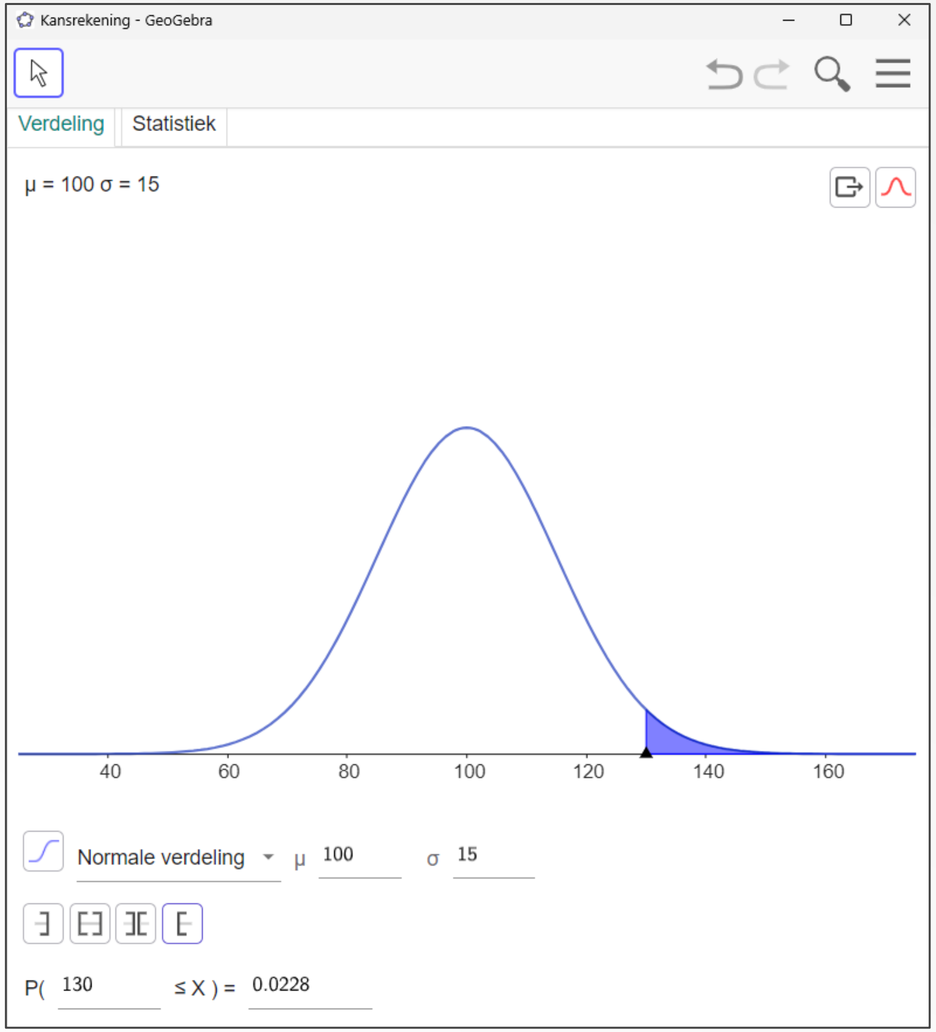
GEOGEBRA

Kans dat je hoogbegaafd bent?

- Kies Normale verdeling
- Stel $\mu = 100$ en $\sigma = 15$ (gemiddeld IQ en std. afw.)
- Kies voor groter dan interval [
- Vul 130 in

Kans dat iemand hoogbegaafd is 0.0228

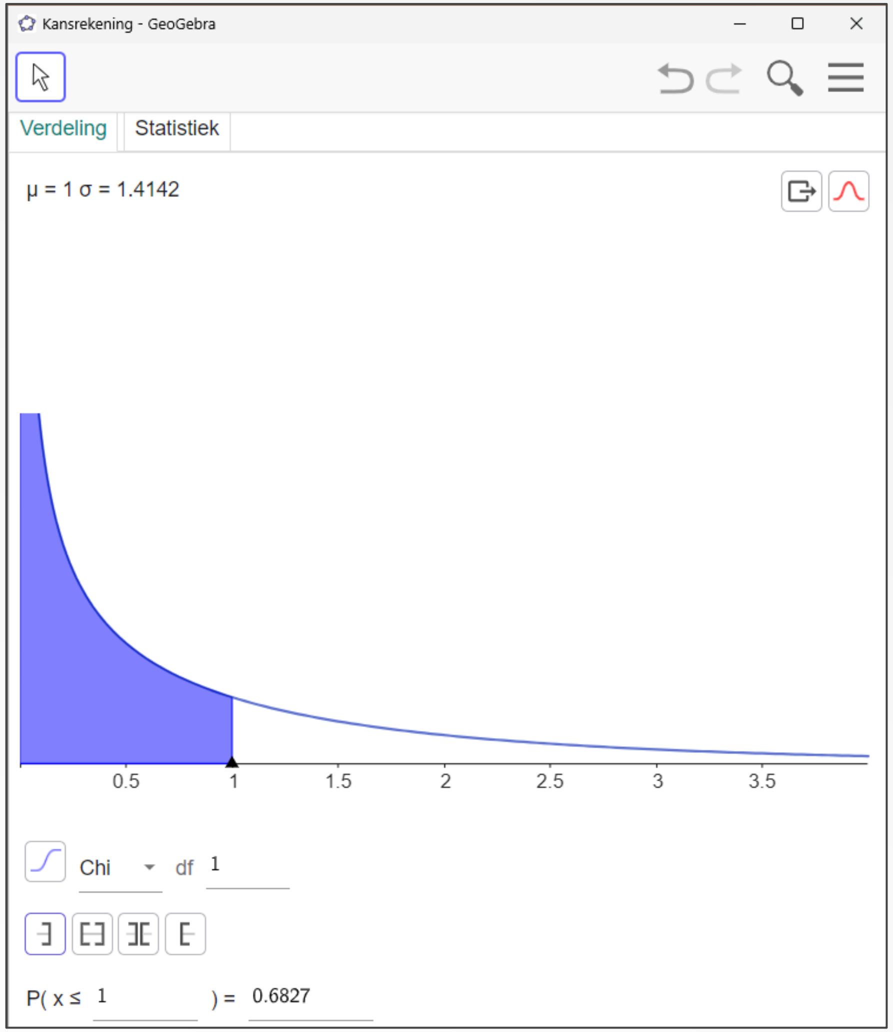
Bereken hoeveel hoogbegaafden in jouw klas zitten.



GEOGEBRA

Wat is de kans op deze uitkomst of minder extreem?

- Kies Chi verdeling
- Stel $df = 1$
- Kies voor groter dan interval]
- Vul 1 in



Kans op deze uitkomst of minder uitzonderlijk

- $P(\chi^2 \leq Q) = 0.6827$

Wat hebben we geleerd?

- **experiment**
  een activiteit die onder gecontroleerde omstandigheden wordt uitgevoerd
- **uitkomstenverzameling**
  alle mogelijke uitkomsten van een verzameling, genoteerd met U
- **gebeurtenis**
  een deelverzameling van de verzameling van alle uitkomsten van een experiment, genoteerd met $A, B, \dots$
- **onafhankelijke gebeurtenissen**
  als het optreden van één gebeurtenis geen effect heeft op de waarschijnlijkheid van het optreden van een andere gebeurtenis
- **afhankelijke gebeurtenissen**
  als twee gebeurtenissen NIET onafhankelijk zijn, dan zeggen we dat ze afhankelijk zijn
- **voorwaardelijke gebeurtenis**
  een gebeurtenis die zich voordoet als een andere gebeurtenis zich al heeft voorgedaan

Wat hebben we geleerd?

- **kans**
  de waarschijnlijkheid van een gebeurtenis, uitgedrukt als een getal tussen 0 en 1

- **toevalsvariabele**
  een variabele die een willekeurige uitkomst vertegenwoordigt in een kansexperiment, waarbij de waarde afhangt van toeval. Genoteerd met een hoofdletter, bv. X, Y, Z, Q, …

- **kansverdeling**
  een overzicht van alle mogelijke uitkomsten van een toevalsvariabele samen met de bijbehorende kansen of waarschijnlijkheden van die uitkomsten.

- **normaalverdeling**
  een symmetrische klokvormige kansverdeling die vaak wordt gebruikt om natuurlijke fenomenen te modelleren, waarbij de meeste waarnemingen zich dichtbij het gemiddelde bevinden, genoteerd als $N(\mu, \sigma)$

- **studentverdeling**
  een kansverdeling die wordt gebruikt in statistiek voor het schatten van de betrouwbaarheidsintervallen en het uitvoeren van hypothesetesten, vooral bij kleinere steekproeven waar de normaalverdeling niet altijd geschikt is, genoteerd als $T(\nu)$ (zie volgend onderdeel)

Wat hebben we geleerd?

- **Chi-kwadraatverdeling**
  Een verdeling die wordt gebruikt om de mate van overeenstemming van waargenomen verdelingen tot verwachte theoretische verdelingen te beoordelen.

- **Fisher-verdeling**
  Een verdeling die wordt gebruikt voor het berekenen van waarschijnlijkheden over verschillen in varianties.