

Data Science 2

Kansverdelingen

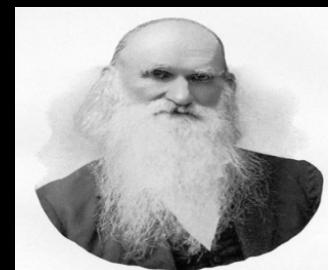
Wim De Keyser
Geert De Paepe
Jan Van Overveld



Quote van de week

"It is always probable that something improbable will happen."

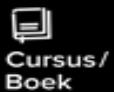
Logan E. Bleckley (1827-1907)



Agenda



1. Wat is een kansverdeling?
2. Verband met frequentietabel
3. Centrum- en spreidingsmaten van een kansverdeling
4. Veel voorkomende kansverdelingen
 - De binomiale verdeling
 - De normale verdeling
5. In de praktijk
6. In de media

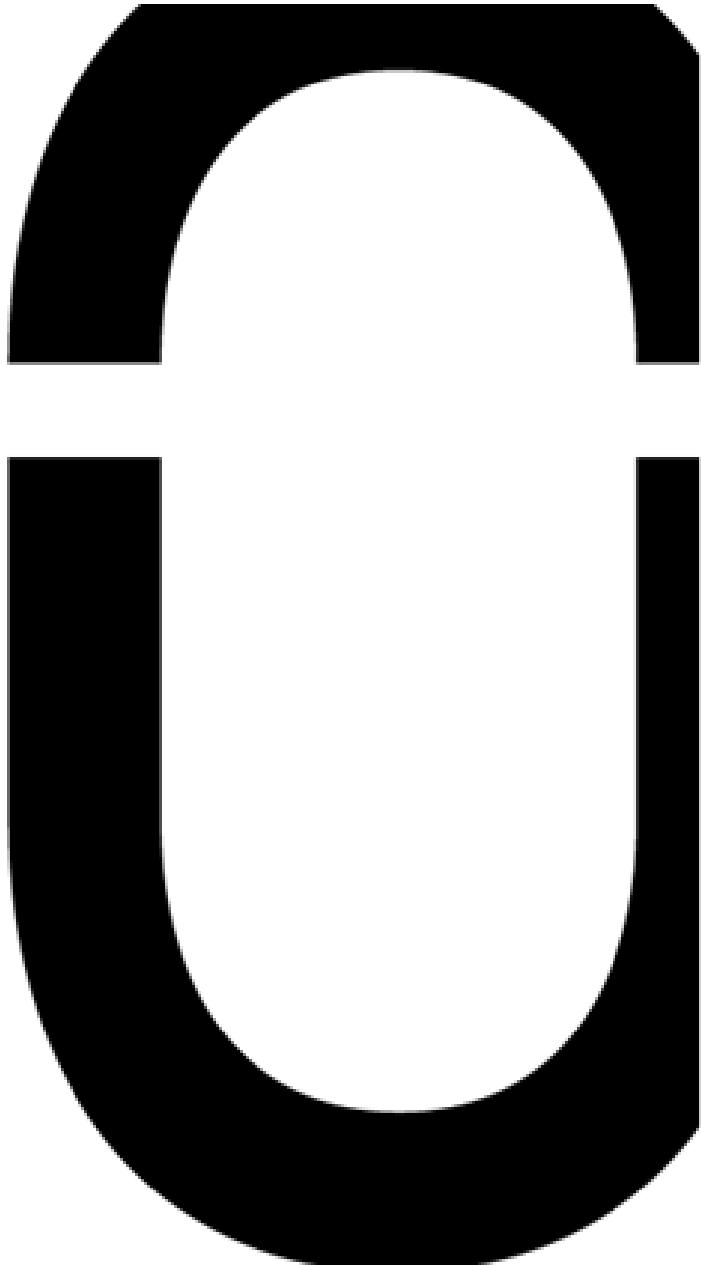


Tekst op Canvas: "Kansverdelingen.pdf"



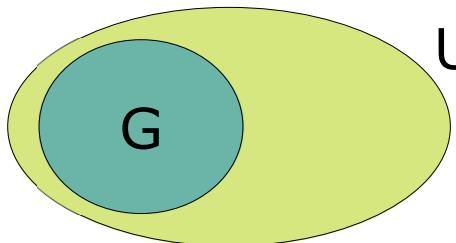
Oefeningen

Herhaling: kans



Kans : Een experiment dat verschillende uitkomsten produceert ondanks dezelfde beginsituatie.

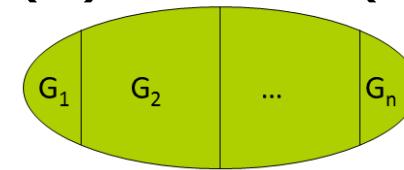
Kansen met verzamelingenleer:



- **U - Verzameling** U (mogelijke uitkomsten)
- **Gebeurtenis** G
- Kans dat gebeurtenis optreedt: $P(G) = \#G / \#U$

Tegengestelde gebeurtenis : $P(G) = 1 - P(\bar{G})$

Uitsluitende gebeurtenissen :



$$U = \{(1,1), (1,2), \dots, (6,6)\}$$



G = een even aantal ogen

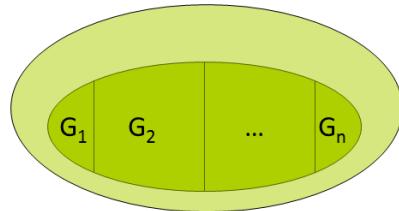
Tegengestelde gebeurtenis van G :
 \bar{G} = een oneven aantal ogen

Uitsluitende gebeurtenis:
 G_1 : # ogen = 2, ... G_{11} : # ogen = 12

De somregel:

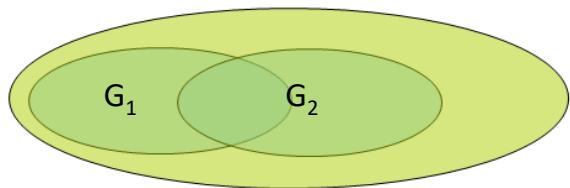
- gebeurtenis G bestaat uit **uitsluitende deelgebeurtenissen**

$$\begin{aligned} P(G) &= P(G_1 \text{ OR } G_2 \text{ OR } \dots \text{ OR } G_n) \\ &= P(G_1 \cup G_2 \cup \dots \cup G_n) \\ &= P(G_1) + P(G_2) + \dots + P(G_n) \end{aligned}$$

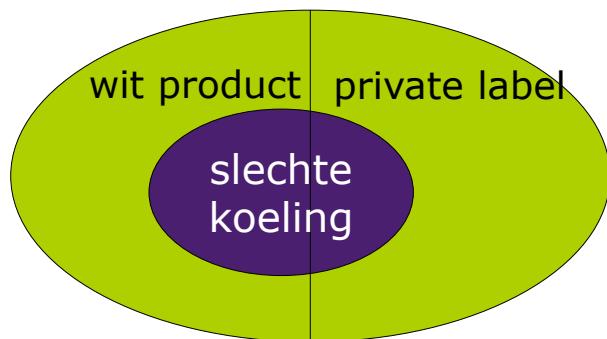


- gebeurtenis G bestaat uit **overlappende deelgebeurtenissen**

$$\begin{aligned} P(G) &= P(G_1 \text{ OR } G_2) = P(G_1 \cup G_2) \\ &= P(G_1) + P(G_2) - P(G_1 \cap G_2) \end{aligned}$$



Voorwaardelijke kans: de kans dat het een slechte koeling is, gegeven dat het een wit product is



$$P(G_{\text{slechte koeling}} \mid G_{\text{wit product}})$$

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

Afhangelijke gebeurtenissen: de uitkomst van de ene gebeurtenis heeft een invloed op de andere gebeurtenis

$$P(G_{\text{slechte koeling}} \mid G_{\text{wit product}}) \neq P(G_{\text{slechte koeling}})$$

Onafhangelijke gebeurtenissen: de uitkomst van de ene gebeurtenis heeft geen invloed op de andere gebeurtenis

$$P(G_{\text{harten}} \mid G_{\text{aas}}) = P(G_{\text{harten}})$$

$$P(G_{\text{harten}}) \times P(G_{\text{aas}}) = P(G_{\text{harten}} \cap G_{\text{aas}})$$

De productregel:

- gebeurtenis G bestaat uit **onafhankelijke deelgebeurtenissen**

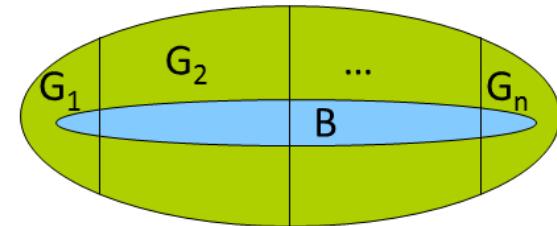
$$\begin{aligned} P(G) &= P(G_1 \text{ EN } G_2 \text{ EN } \dots \text{ EN } G_n) \\ &= P(G_1 \cap G_2 \cap \dots \cap G_n) \\ &= P(G_1) \times P(G_2) \times \dots \times P(G_n) \end{aligned}$$

- gebeurtenis G bestaat uit **afhankelijke deelgebeurtenissen**

$$\begin{aligned} P(G) &= P(G_1 \text{ EN } G_2) \\ &= P(G_1 \cap G_2) \\ &= P(G_1 | G_2) \times P(G_2) \end{aligned}$$

Wet van de totale kans:

- G_1, G_2, \dots, G_n zijn uitsluitende gebeurtenissen



$$\begin{aligned}P(B) &= P(B \cap G_1) + P(B \cap G_2) + \dots + P(B \cap G_n) \\&= P(B | G_1) \cdot P(G_1) + P(B | G_2) \cdot P(G_2) + \dots + P(B | G_n) \cdot P(G_n) \\&= \sum_{i=1}^n P(B | G_i) \cdot P(G_i)\end{aligned}$$

Wet van Bayes:

$\{G_1, G_2, \dots, G_n\}$ vormen een partitie is van G

We weten dat: $P(G_k | B) \cdot P(B) = P(G_k \cap B) = P(B \cap G_k) = P(B | G_k) \cdot P(G_k)$

$$\text{Dus: } P(G_k | B) = \frac{P(B | G_k) \cdot P(G_k)}{P(B)}$$

Wat is een kansverdeling?

Kansverdelingen

- Wat is een kansverdeling?
 - doe een theoretische oneindige steekproef/experiment
 - bepaal de relatieve frequenties van iedere waarde
 - deze tabel noemen we "kansverdeling"

Kansverdelingen

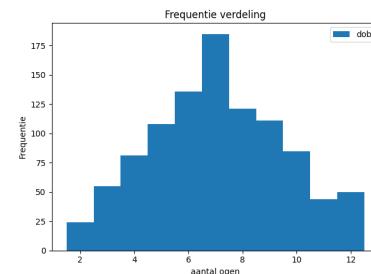
Experiment: gooij met 2 dobbelstenen

- Wat gaat de kansverdeling zijn voor het totaal aantal ogen?
- Experiment in Python:

```
>>> dob1 = np.random.randint(1,7,1000)
>>> dob2 = np.random.randint(1,7,1000)
>>> dob = pd.DataFrame({'dob':dob1 + dob2})
>>> plt.figure()
>>> dob.plot.hist(title='Frequentie verdeling',
bins = [1.5,2.5,3.5,4.5,5.5,6.5,7.5,8.5,9.5,10.5,11.5,12.5],
stacked=True)
>>> plt.xlabel('Aantal ogen')
>>> plt.ylabel('Frequentie')
>>> plt.show()
```



x_i	#	F_i
2		3
3		2
4		7
5		10
6		4
7		5
8		5
9		7
10		3
11		4
12		0
		50

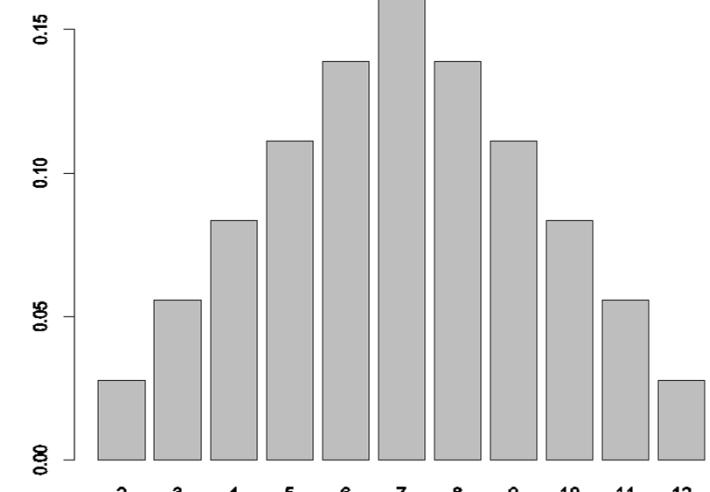


Kansverdelingen

Theoretisch:

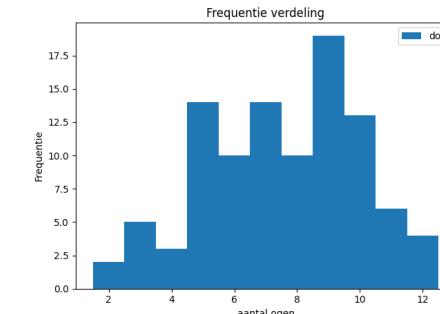
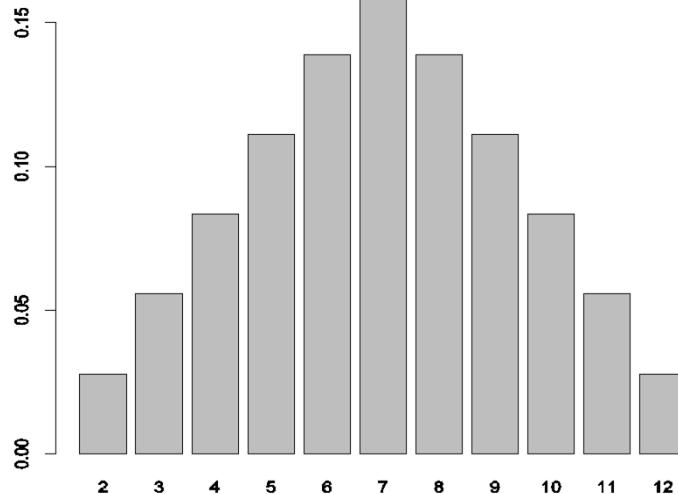


$G = x_i$	Mogelijke uitkomsten	#G	$\#G/\#U = P(x_i)$
2	(1,1)	1	1/36
3	(1,2); (2,1)	2	2/36
4	(1,3); (2,2); (3,1);	3	3/36
5	(1,4); (2,3), (3,2); (4,1)	4	4/36
6	(1,5); (2,4); (3,3); (4,2); (5,1)	5	5/36
7	(1,6); (2,5); (3,4); (4,3); (5,2); (6,1)	6	6/36
8	(2,6); (3,5); (4,4); (5,3); (6,2)	5	5/36
9	(3,6); (4,5); (5,4); (6,3)	4	4/36
10	(4,6); (5,5); (6,4)	3	3/36
11	(5,6); (6,5)	2	2/36
12	(6,6)	1	1/36
	#U (= TOTAAL)	36	36/36
			1,0000

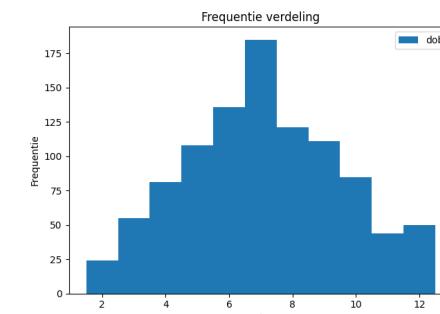


Kansverdelingen

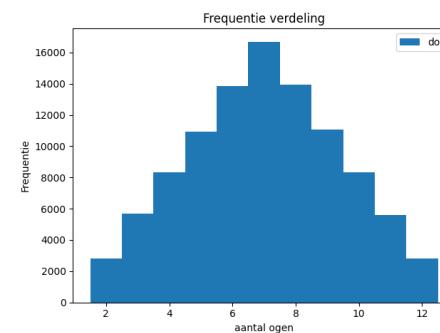
Theoretisch vs experiment:



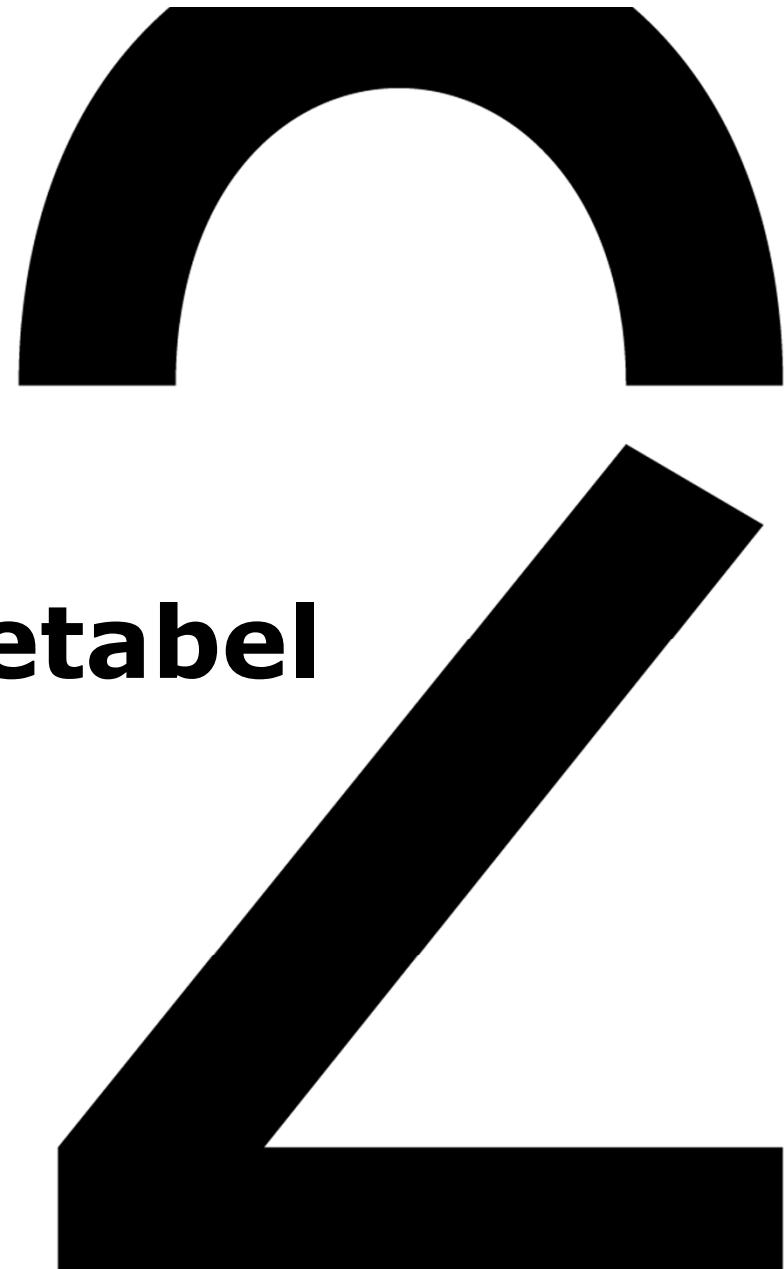
100 x gooien



1000 x gooien



10000 x gooien



Verband met frequentietabel

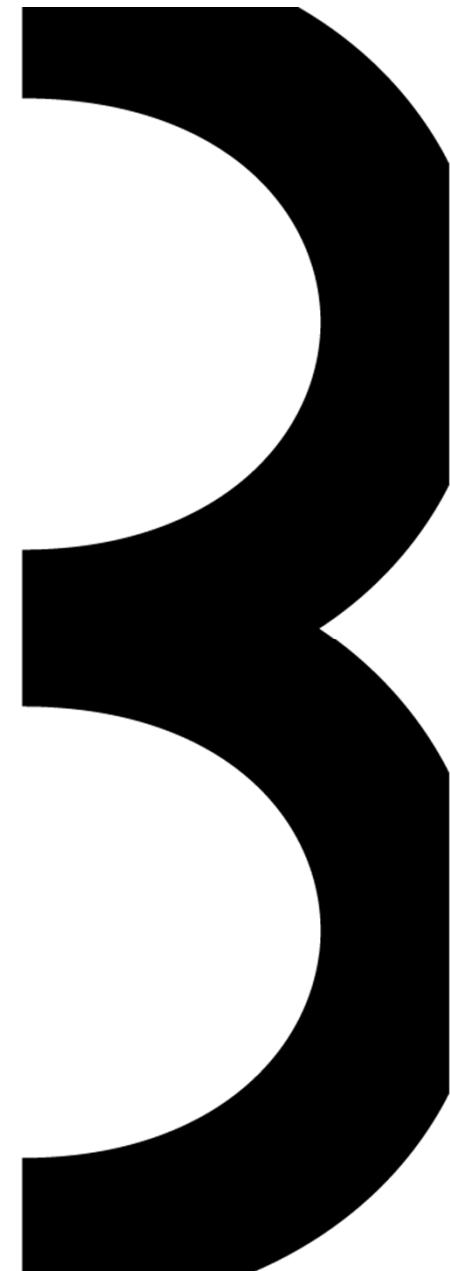
Kansverdelingen en frequentietabellen

- Kansverdeling
 - theoretisch bepaald
 - steekproef wordt oneindig groot verondersteld
- Frequentietabellen
 - experimenteel bepaald
 - eindige steekproef
 - benadert kansverdeling
- Allebei: relatieve frequenties

X_i	$P(X_i)$
2	0,0278
3	0,0556
4	0,0833
5	0,1111
6	0,1389
7	0,1667
8	0,1389
9	0,1111
10	0,0833
11	0,0556
12	0,0278
	1,0000

X_i	#	F_i	f_i
2		3	3/50 0,06
3		2	2/50 0,04
4		7	7/50 0,14
5		10	10/50 0,20
6		4	4/50 0,08
7		5	5/50 0,10
8		5	5/50 0,10
9		7	7/50 0,14
10		3	3/50 0,06
11		4	4/50 0,08
12		0	0/50 0,00
		50	50/50 1,0000

Centrum- en spreidingsmaten van een kansverdeling



Centrummaten en spreidingsmaten

- Kansen en relatieve frequenties zijn gelijkaardig
- Is er dan een gemiddelde, mediaan, modus (centrummaten kansverdeling)?
- Is er een IQR, variantie, standaardafwijking (spreidingsmaten kansverdeling)?

Centrummaten kansverdeling

- Gemiddelde noemen we "**verwachte waarde**"
- Dit is namelijk de waarde die we verwachten voor de populatie
- Symbool: μ
- $\mu = \text{som van (elke waarde} * \text{kans op waarde})$
- Voor het gooien met 2 dobbelstenen is dat:

$$\mu = \sum_{i=2}^{12} x_i \cdot P(x_i) = 2 \cdot \frac{1}{36} + \dots + 7 \cdot \frac{6}{36} + \dots + 12 \cdot \frac{1}{36} = 7$$



Centrummaten kansverdeling

- Voorbeeld: multiple choice examen met giscorrectie

- 1 vraag
- keuze uit 4 antwoorden
- juist antwoord = 1 punt
- fout antwoord = -1/4 punt
- kans op een goed antwoord is p

Score	Kans
1	p
-1/4	1-p

- Stel dat je er niks van kent. Wat is p dan? Wat is dan de verwachte score?
$$p = 1/4 \quad \mu = 1 \cdot 1/4 - 1/4 \cdot 3/4 = 1/16$$
- Stel dat je twijfelt tussen 3 antwoorden. Wat is p ? Wat is dan de verwachte score?
$$p = 1/3 \quad \mu = 1 \cdot 1/3 - 1/4 \cdot 2/3 = 1/6$$
- Stel dat je twijfelt tussen 2 antwoorden. Wat is p ? Wat is dan de verwachte score?
$$p = 1/2 \quad \mu = 1 \cdot 1/2 - 1/4 \cdot 1/2 = 3/8$$
- Is het nu beter om te gokken of niks in te vullen?

Spreidingsmaten kansverdeling

- Variantie en standaardafwijking zijn ook verondersteld op de hele populatie

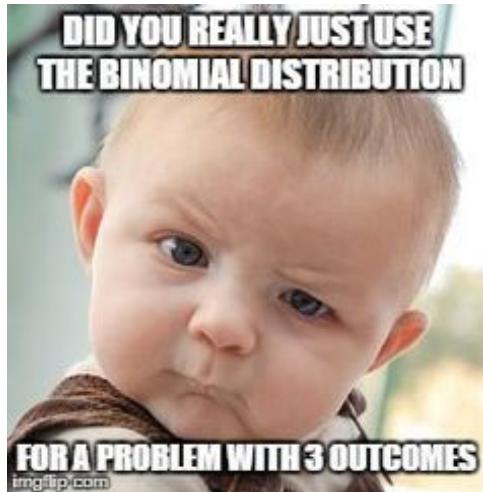
$$\sigma^2 = \sum_{i=1}^n (x_i - \mu)^2 \cdot P(x_i) \quad \sigma = \sqrt{\sigma^2}$$

- We gebruiken dus symbolen σ^2 en σ
- Voor het gooien met 2 dobbelstenen is dat:

$$\sigma^2 = (2 - 7)^2 \cdot \frac{1}{36} + \dots + (7 - 7)^2 \cdot \frac{6}{36} + \dots + (12 - 7)^2 \cdot \frac{1}{36} = 5,8333$$

$$\sigma = \sqrt{5,8333} = 2,415$$





Veel voorkomende kansverdelingen

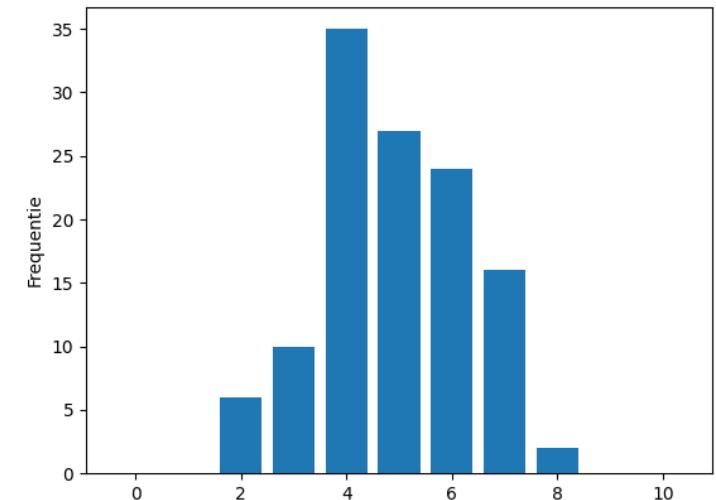
- ↳ **De binomiale verdeling**
- ↳ De normale verdeling
- ↳ Verband tussen de binomiale en de normale verdeling
- ↳ De Poisson verdeling
- ↳ Andere verdelingen

De binomiale verdeling



- Experiment: gooien van een muntstuk
 - Neem elk een muntstuk
 - Gooi elk 10 keer het muntstuk op en noteer telkens het resultaat (kop of munt)
 - Hoeveel studenten hadden x-keer kop?
 - Resultaten (vorige jaren):

```
>>> x=range(0,11)
>>> y= (0,0,6,10,35,27,24,16,2,0,0)
>>> plt.figure()
>>> plt.bar(x, y)
>>> plt.ylabel('Frequentie')
>>> plt.show()
```

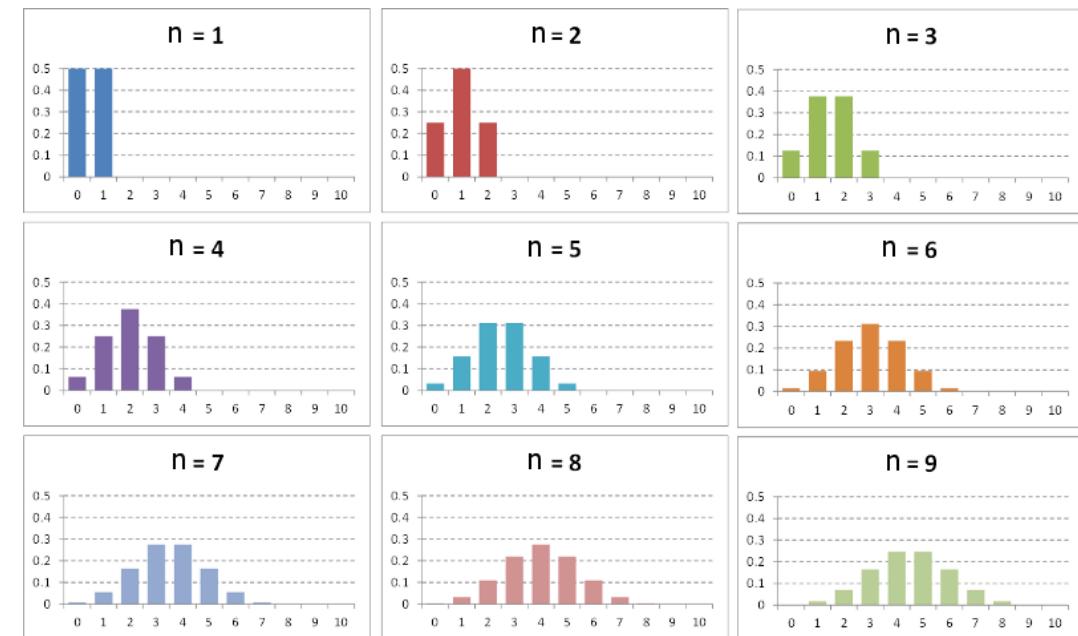
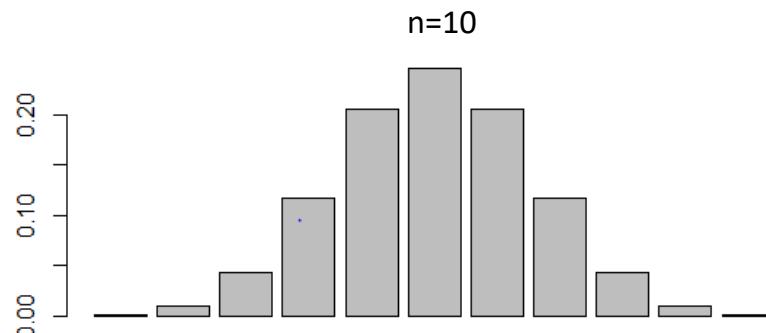


De binomiale verdeling

- Is een model voor een experiment waarbij:
 - er maar **2** uitkomsten mogelijk zijn ('succes' en 'geen succes')
 - iedere uitkomst een bepaalde waarschijnlijkheid heeft:
 - eerste mogelijke uitkomst (succes) : p
 - tweede mogelijke uitkomst (geen succes): $1-p$
 - experiment wordt verschillende achtereenvolgende keren (n) uitgevoerd en de resultaten worden gemeten
 - de vraag is: "Hoeveel kans is er dat bij n experimenten er x gelijk zijn aan de eerste mogelijke uitkomst?"
(maw kans dat x experimenten op n succesvol zijn?)

De binomiale verdeling

- Experiment: gooien van een muntstuk
 - 2 uitkomsten mogelijk: kop of munt
 - iedere uitkomst heeft een bepaalde waarschijnlijkheid:
 - kop: $0,5 (= p)$
 - munt: $0,5 (= 1 - p)$
 - Experiment n keren uitvoeren



De binomiale verdeling - Voorbeeld

- Examen met 5 meerkeuzevragen
- Iedere vraag heeft 4 mogelijke antwoorden, waarvan 1 juist
- Wat is de kans dat we exact 2 vragen juist beantwoorden?
 - we hebben niet gestudeerd en gokken dus...
 - er is geen giscorrectie



De binomiale verdeling - Voorbeeld

- n = aantal vragen
- x = aantal juist
- p = kans op succes
- formule:

$$P(x) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x}$$

- Voorbeeld: 2 van de 5 vragen juist:

$$P(2) = \binom{5}{2} \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right)^3 = 10 \cdot 0,0625 \cdot 0,4219 = 0,2637$$



De binomiale verdeling - Praktisch

- Wat is dus de kans om 0, 1, 2, 3, 4 en 5 vragen juist te hebben?

in Python:

```
>>> from scipy.stats import binom  
>>> x = range(0,6)  
>>> binom.pmf(x,5,1/4) # probability mass function
```

- Wat is dus de kans om te slagen op dit examen?

$$\begin{aligned}P(x \geq 3) &= P(x=3) + P(x=4) + P(x=5) \\&= 1 - P(x \leq 2)\end{aligned}$$

in Python:

```
>>> binom.pmf(3,5,1/4) + binom.pmf(4,5,1/4) + binom.pmf(5,5,1/4)  
>>> 1 - (binom.pmf(0,5,1/4) + binom.pmf(1,5,1/4) + binom.pmf(2,5,1/4))  
>>> 1 - binom.cdf(2,5,1/4) # cumulative density function
```



De binomiale verdeling - Praktisch

- Uiteindelijke kansverdeling voor dit voorbeeld:

aantal vragen juist	kans	cum. kans
0	0,2373	0,2373
1	0,3955	0,6328
2	0,2637	0,8965
3	0,0879	0,9844
4	0,0146	0,9990
5	0,0010	1,0000

```
>>> x = range(0,6)
```

```
>>> binom.pmf(x,5,1/4) # probability mass function => kans
```

```
>>> binom.cdf(x,5,1/4) # cumulative density function => cum. kans
```



Verwachte waarde

- Verwachte waarde = hoeveel vragen zullen gemiddeld goed zijn?
- $\mu = n * p$
- In dit geval: $5 * 1/4 = 1,25$
- In Python: >>> `binom.mean(5, 1/4)`



Standaardafwijking

- De standaardafwijking is ook gemakkelijk te berekenen:

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)}$$

- De standaardafwijking is hier dus:
 $\text{sqrt}(5 * 1/4 * 3/4) = 0,968$
- In Python: >>> `binom.std(5, 1/4)`



Oefening

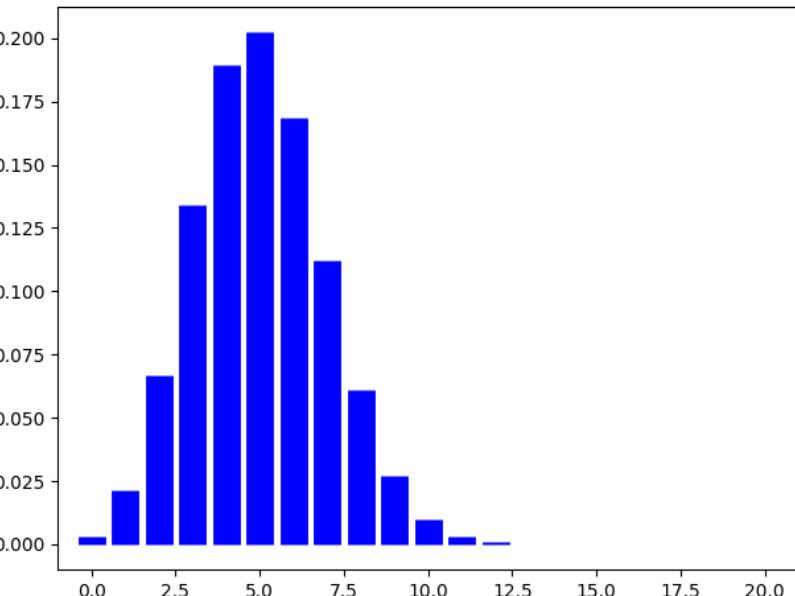
- Examen computersystemen:
 - 20 multiple choice vragen
 - 4 keuzemogelijkheden per vraag
- Wat is de verwachte score als iedereen gokt?
 - het gemiddelde was 7/20... $\mu = n \cdot p = 20 \cdot 1/4 = 5$
- Wat is de standaardafwijking?

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)} = 1,9365$$

Grafisch

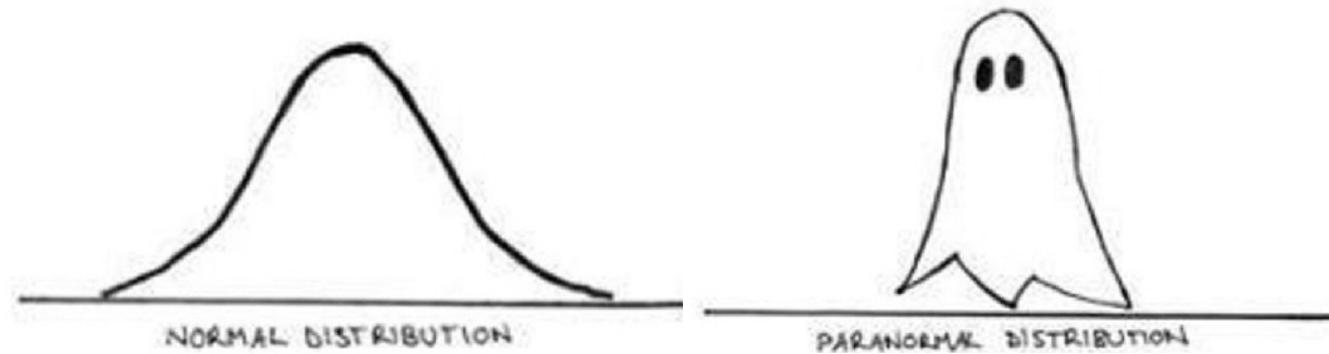


- Je kan de kansverdeling ook grafisch weergeven:



```
>>> n = 20
>>> p = 1/ 4
>>> x = range(0, n+1)
>>> plt.figure()
>>> fig, ax = plt.subplots(1, 1)
>>> ax.vlines(x, 0, binom.pmf(x, n, p),
              colors='b', lw=15)
>>> plt.show()
```

- Is deze links of rechts scheef?



Veel voorkomende kansverdelingen

- ↳ De binomiale verdeling
- ↳ **De normale verdeling**
- ↳ Verband tussen de binomiale en de normale verdeling
- ↳ De Poisson verdeling
- ↳ Andere verdelingen

De normale verdeling

- Is een model voor een experiment waarbij:
 - de variabele continu is
 - de kansen symmetrisch verdeeld zijn (evenveel kans om meer of minder dan het gemiddelde uit te komen)
 - de verwachtingswaarde en verwachte standaardafwijking gekend zijn

Voorbeelden normale verdeling

- Gemeten lengtes van jongens van 20 jaar
- We verwachten een gemiddelde lengte van 180 cm met een standaardafwijking van 10 cm
- Wat is nu de kans dat iemand kleiner is dan 140 cm?
- Wat is nu de kans dat iemands lengte tussen 170 cm en 190 cm ligt?
- Wat is de kans dat iemand exact 180 cm groot is?
(continue variabele...)

De normale verdeling

- formule voor de normale verdeling:

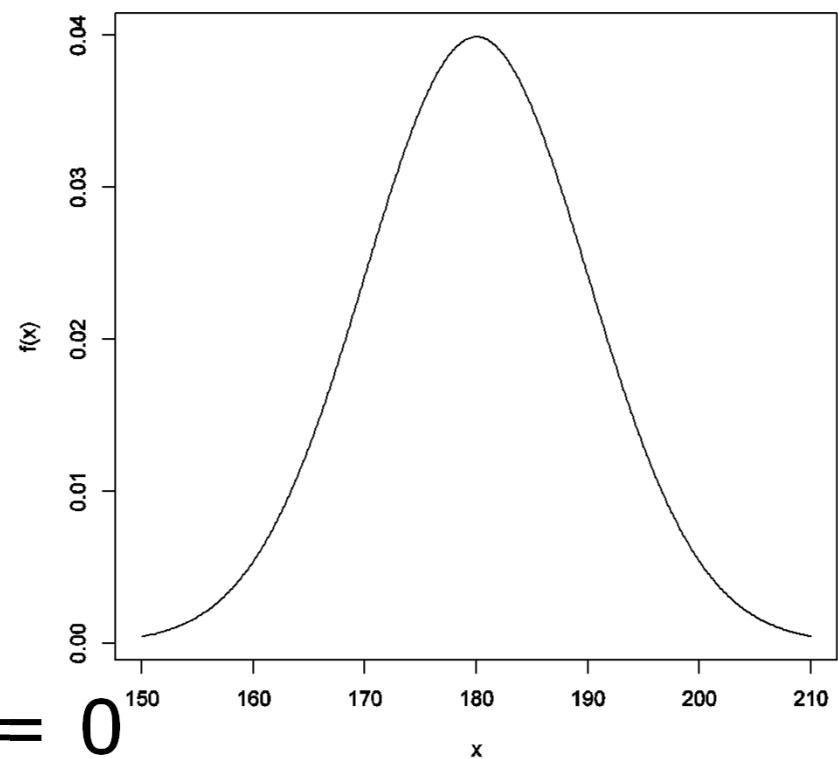
$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} e^{\left(\frac{-(x-\mu)^2}{2 \cdot \sigma^2} \right)}$$

- in Python:

```
>>> from scipy.stats import norm  
>>> norm.pdf(x, loc=μ, scale=σ)  
    # x -> [x-0.5, x+0.5[
```

- $f(x)$ is niet de kans dat x gemeten wordt!

Continue verdeling $\Rightarrow P(X=x) = 0$

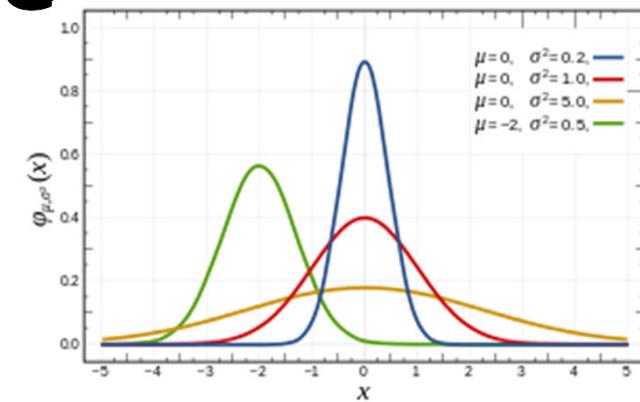


Berekenen van de oppervlakte

- Grafiek is bepaald door μ en σ
- Hoe berekenen we de oppervlakte?
 - integraal uitrekenen...
 - tabel (tabel voor iedere combinatie van μ en σ ?)
 - gebruik de cumulatieve kansverdeling

in Python:

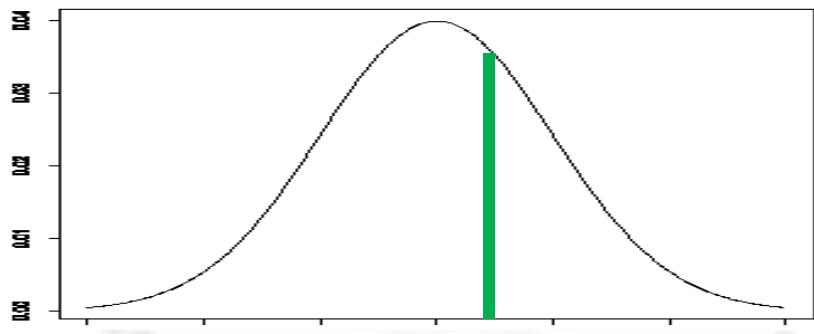
```
>>> norm.cdf(190, loc=180, scale=10)  
- norm.cdf(170, loc=180, scale=10)
```



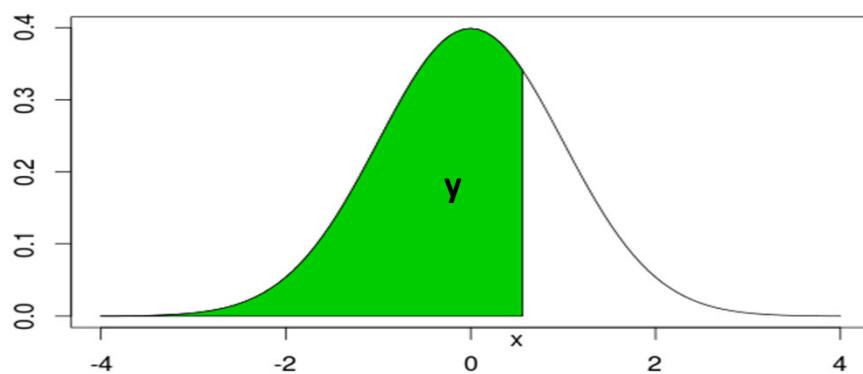
Wat is dit?
Hoe ziet die
eruit?

Berekenen van de oppervlakte

Probability density function

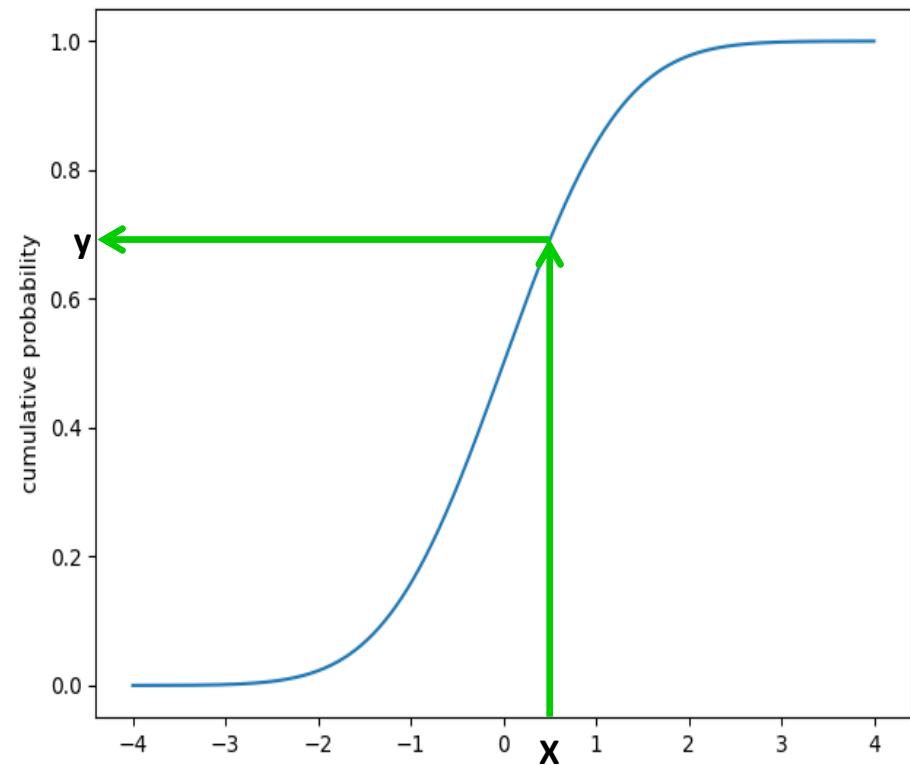


norm.pdf



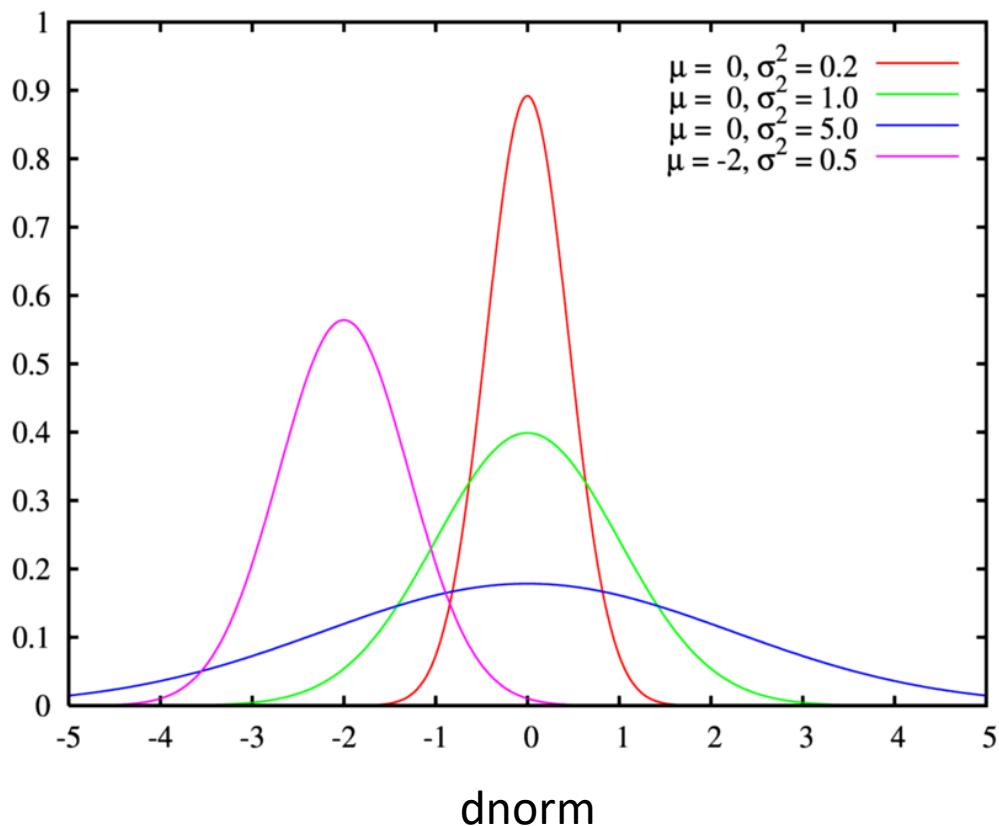
norm.cdf

Cumulative distribution function.

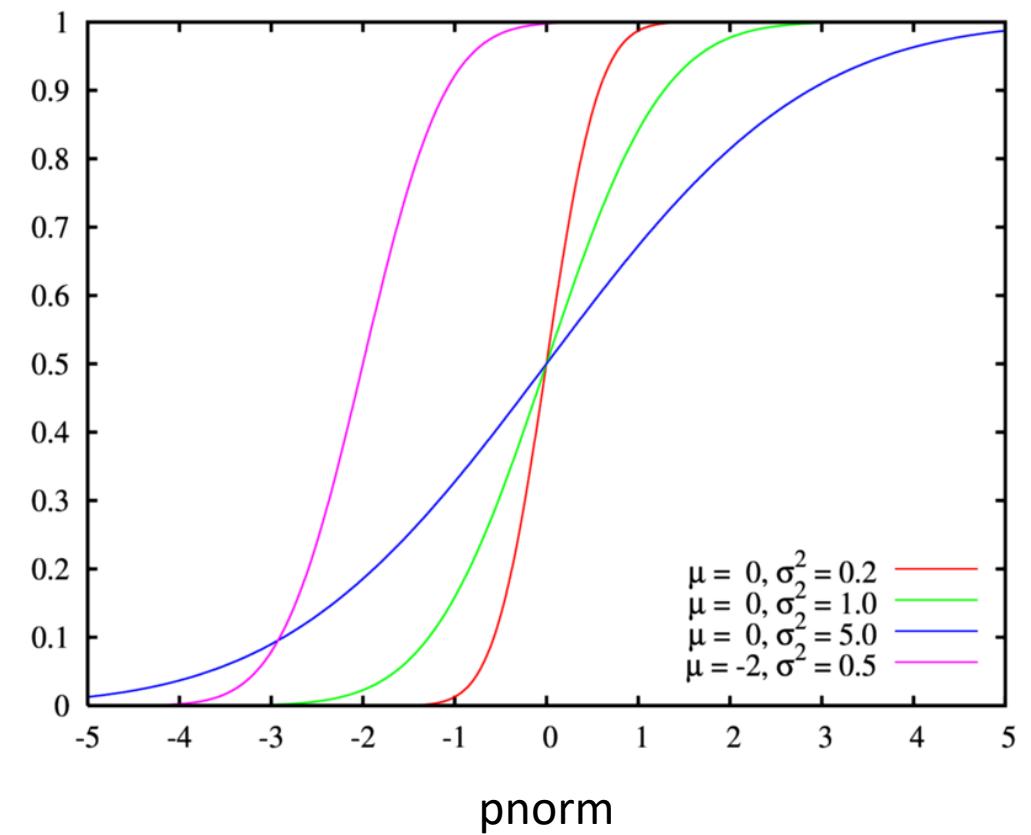


Berekenen van de oppervlakte

kansverdeling



cumulatieve kansverdeling

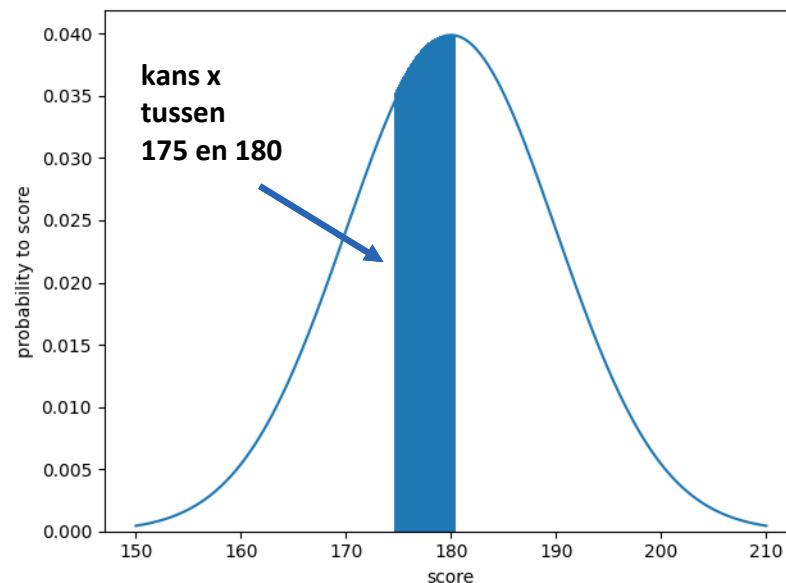


De normale verdeling

Kans dat x tussen x_1 en x_2 ligt

= oppervlakte tussen x_1 en x_2

voorbeeld: kans dat lengte tussen 175 en 180 ligt:

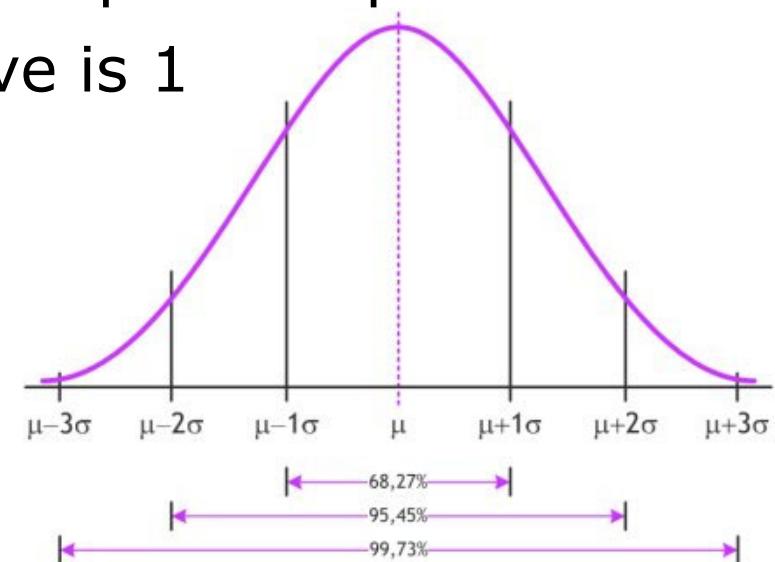


In Python:

```
>>> norm.cdf(180, loc=180, scale=10)  
- norm.cdf(175, loc=180, scale=10)
```

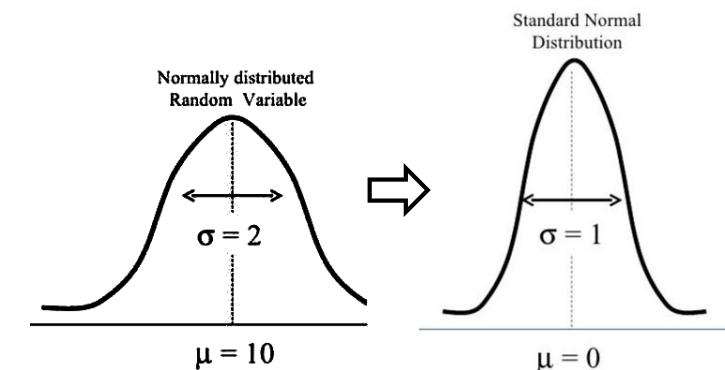
De normale verdeling

- Eigenschappen (onafhankelijk van μ en σ)
 - 68,3% van de oppervlakte ligt tussen $\mu-\sigma$ en $\mu+\sigma$
 - 95,5% van de oppervlakte ligt tussen $\mu-2\sigma$ en $\mu+2\sigma$
 - 99,7% van de oppervlakte ligt tussen $\mu-3\sigma$ en $\mu+3\sigma$
 - de totale oppervlakte onder de curve is 1
- Dus:
 - 68,3% van de jongens is tussen 170 en 190 cm



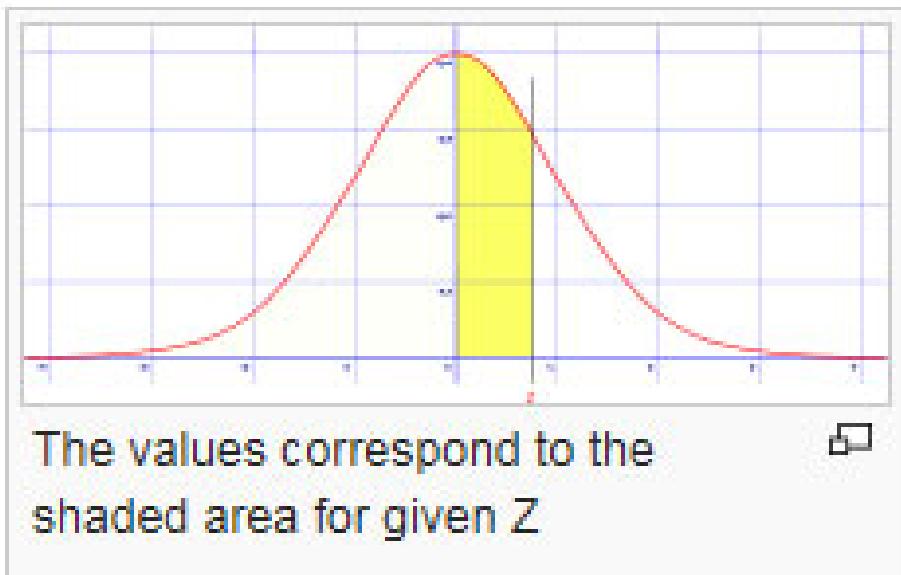
De standaardnormale verdeling

- Verwachte waarde = 0
- Standaardafwijking = 1
- Je kan iedere normale verdeling omzetten in de standaardnormale via een "Z-transformatie" (cfr Z-scores): $Z = (X-\mu)/\sigma$
- Kan gebruikt worden om oppervlaktes te bepalen adhv de tabel van de standaardnormale verdeling (zie volgende slide)



De standaardnormale verdeling

Tabel: cummulatieve verdelingsfunctie vanaf het gemiddelde (0 tot Z)



Overbodig
Sinds studenten een
laptop gebruiken!

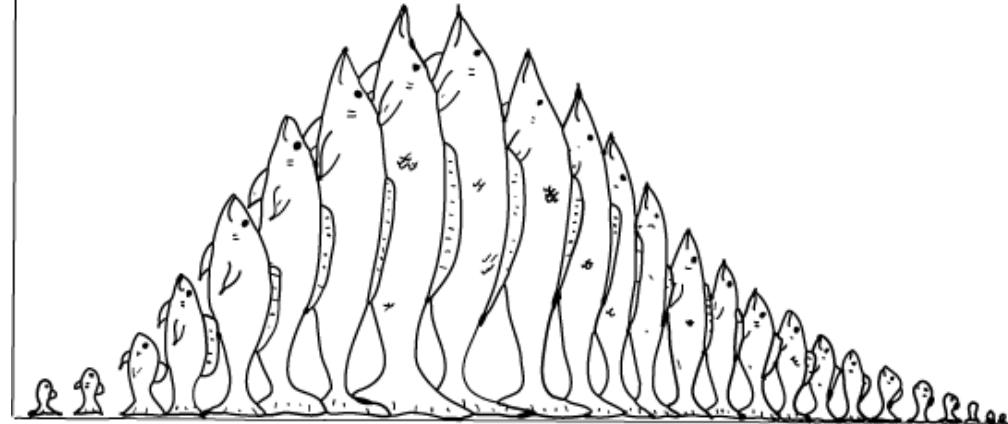
z	+0.00	+0.01	+0.02	+0.03	+0.04	+0.05	+0.06	+0.07	+0.08	+0.09
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03980	0.04380	0.04778	0.05172	0.05567	0.05966	0.06360	0.06749	0.07142	0.07535
0.2	0.07930	0.08317	0.08708	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16278	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21568	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23897	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29951	0.30224	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31814	0.32121	0.32338	0.32549	0.32894	0.33147	0.33398	0.33646	0.33891
1.0	0.34134	0.34455	0.34811	0.35149	0.35503	0.35814	0.36133	0.36578	0.36993	0.36214
1.1	0.36440	0.36750	0.37076	0.37386	0.37693	0.37998	0.38300	0.38700	0.39100	0.38298
1.2	0.38747	0.38866	0.38877	0.39065	0.39251	0.39435	0.39617	0.39798	0.39973	0.40147
1.3	0.40326	0.40490	0.40660	0.40834	0.40988	0.41149	0.41303	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42641	0.42775	0.42922	0.43056	0.43189
1.5	0.43319	0.43449	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44650	0.44738	0.44824	0.44910	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45613	0.45637	0.45728	0.45813	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327
1.8	0.46407	0.46485	0.46553	0.46638	0.46712	0.46784	0.46856	0.46926	0.46995	0.47062
1.9	0.47128	0.47133	0.47257	0.47320	0.47381	0.47441	0.47500	0.47558	0.47615	0.47670
2.0	0.47724	0.47778	0.47831	0.47882	0.47932	0.47982	0.48030	0.48077	0.48124	0.48169
2.1	0.48214	0.48257	0.48300	0.48341	0.48382	0.48422	0.48461	0.48500	0.48537	0.48574
2.2	0.48610	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.48840	0.48870	0.48899
2.3	0.48928	0.48956	0.48983	0.49010	0.49036	0.49061	0.49086	0.49111	0.49134	0.49158
2.4	0.49180	0.49202	0.49224	0.49245	0.49266	0.49286	0.49305	0.49324	0.49343	0.49361
2.5	0.49379	0.49396	0.49413	0.49430	0.49446	0.49461	0.49477	0.49492	0.49508	0.49520
2.6	0.49534	0.49547	0.49560	0.49573	0.49585	0.49598	0.49609	0.49621	0.49632	0.49643
2.7	0.49653	0.49664	0.49674	0.49683	0.49693	0.49702	0.49711	0.49720	0.49728	0.49736
2.8	0.49744	0.49752	0.49760	0.49767	0.49774	0.49781	0.49788	0.49795	0.49801	0.49807
2.9	0.49813	0.49819	0.49825	0.49831	0.49836	0.49841	0.49846	0.49851	0.49856	0.49861
3.0	0.49885	0.49889	0.49874	0.49878	0.49882	0.49886	0.49889	0.49893	0.49896	0.49900

De standaardnormale verdeling

- Voorbeeld
 - lengtemetingen: mean=180, sd=10
 - wat is de Z-score van iemand met lengte 190?
 - wat is de Z-score van iemand met lengte 200?
 - stel dat je enkel de standaardnormale verdeling kan berekenen. Hoe bereken je hoeveel kans er is om iemand te meten tussen 190 en 200?
 - In Python: >>> `norm.cdf(...)` – `norm.cdf(...)`

$$(200-180)/10$$

$$(190-180)/10$$



Veel voorkomende kansverdelingen

- ↳ De binomiale verdeling
- ↳ De normale verdeling
- ↳ Verband tussen de binomiale en de normale verdeling
- ↳ **De Poisson verdeling**
- ↳ Andere verdelingen

De Poisson verdeling

- Is een model voor een experiment waarbij:
 - de variabele discreet is
 - de kansen weergeven hoeveel keer bepaalde voorvallen gedurende een gegeven tijdsinterval, afstand, gebied, volume,... voorkomen
 - De parameter λ (een positief getal) het verwachte aantal voorvallen in het tijdsinterval weergeeft. λ is eveneens het gemiddelde van de Poisson verdeling.

Voorbeelden Poisson verdeling

- het aantal mails die iemand op een dag krijgt
- het aantal keren in een minuut dat een webserver wordt benaderd
- het aantal dode dieren op een kilometer weg
- het aantal naaldbomen op een hectare
- het aantal keren dat er een brand is in een huis op 20 jaar tijd

Wordt frequent gehanteerd door verzekерingsmaatschappijen

De Poisson verdeling

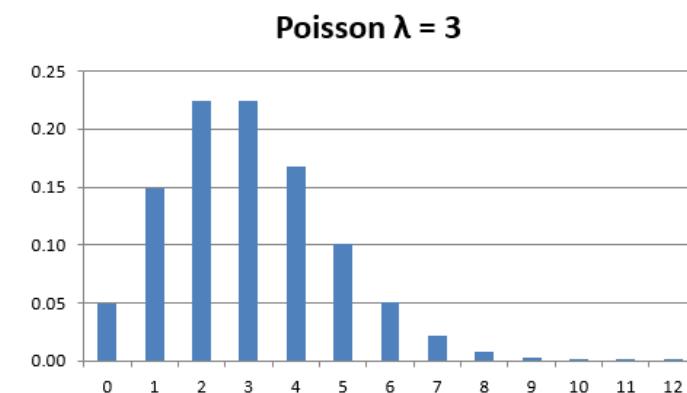
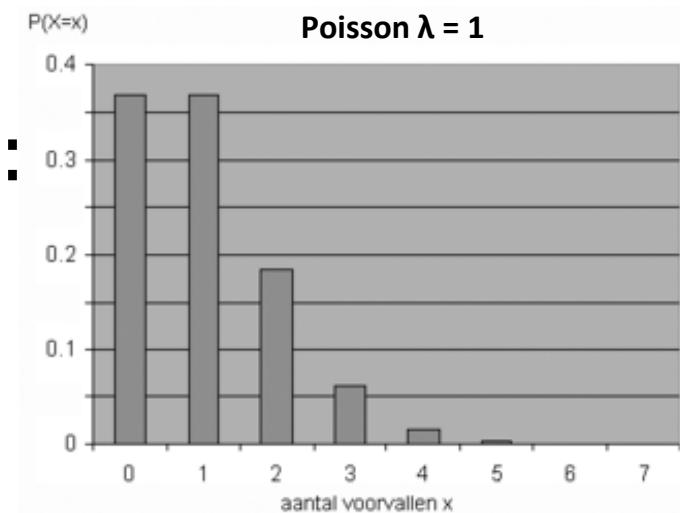
- formule voor de Poisson verdeling:

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

- in Python:

```
>>> from scipy.stats import poisson  
>>> poisson.pmf(k, λ)  
>>> k=range(0,13)  
>>> poisson.pmf(k, 3) # λ = 3
```

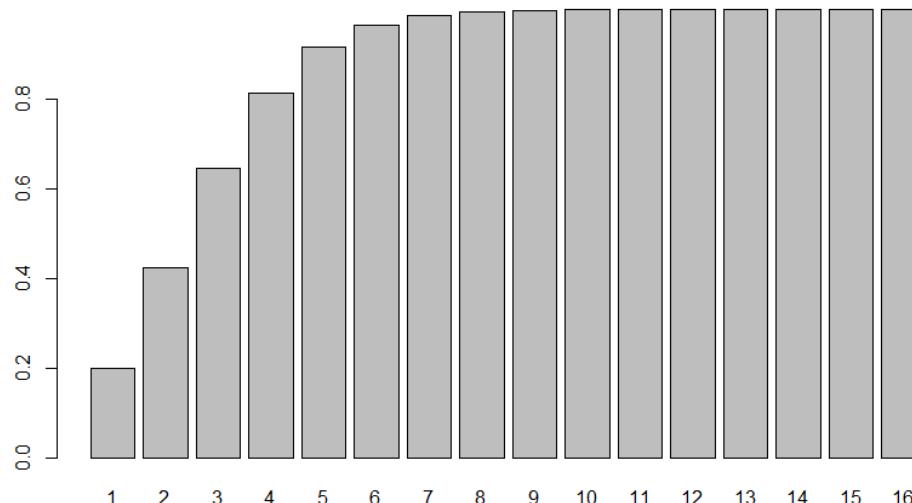
- De som van alle 'staafjes' is gelijk aan 1



De Poisson verdeling

- Cummulatieve verdelingsfunctie:
- in Python:

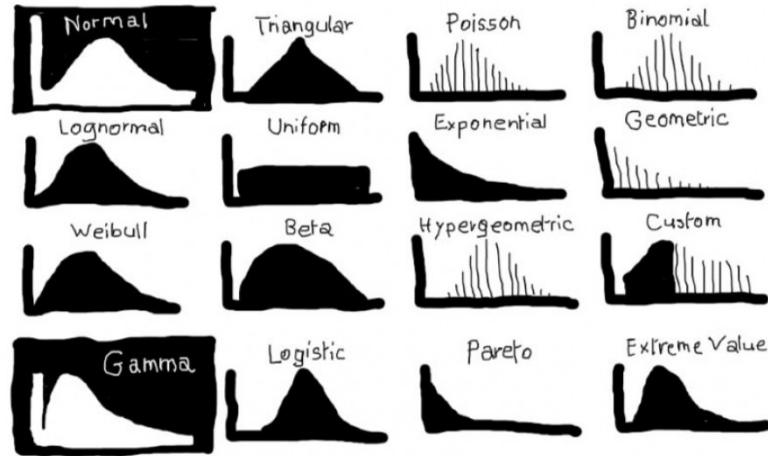
```
>>> poisson.cdf(2, 3) # poisson.pmf(0, 3) + poisson.pmf(1, 3)  
                  + poisson.pmf(2, 3)
```



De Poisson verdeling - Toepassing

In België zijn er ongeveer 10.000 branden in gebouwen per jaar. Er zijn ongeveer 4.500.000 gebouwen. In een kleine gemeente zijn er 9.000 gebouwen.

- a. Bereken de kans dat er 25 branden in 2023 in de gemeente zullen plaats vinden.
 - b. Bereken ook de kans op minstens 25 branden in 2023.
-
- 1 brand op 450, 20 branden op 9.000
 - in Python:
`>>> poisson.pmf(25, 20) # 0,044588`
`>>> 1-poisson.cdf(24, 20) # 0,15678`



Veel voorkomende kansverdelingen

- ↳ De binomiale verdeling
- ↳ De normale verdeling
- ↳ Verband tussen de binomiale en de normale verdeling
- ↳ De Poisson verdeling
- ↳ **Andere verdelingen**

Andere verdelingsfuncties

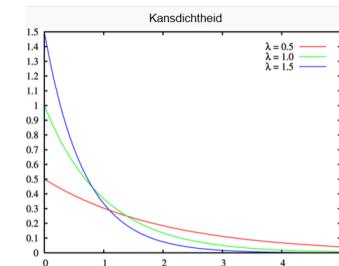
Er bestaan heel wat verdelingsfuncties met elk een specifiek toepassingsgebied:

beta
Cauchy
Chi-kwadraat
F
gamma
geometrische
hypergeometrische

Exponentiële ←
Log-normale
multinomiale
negatief binomiale
Student
uniforme
Weibull

...

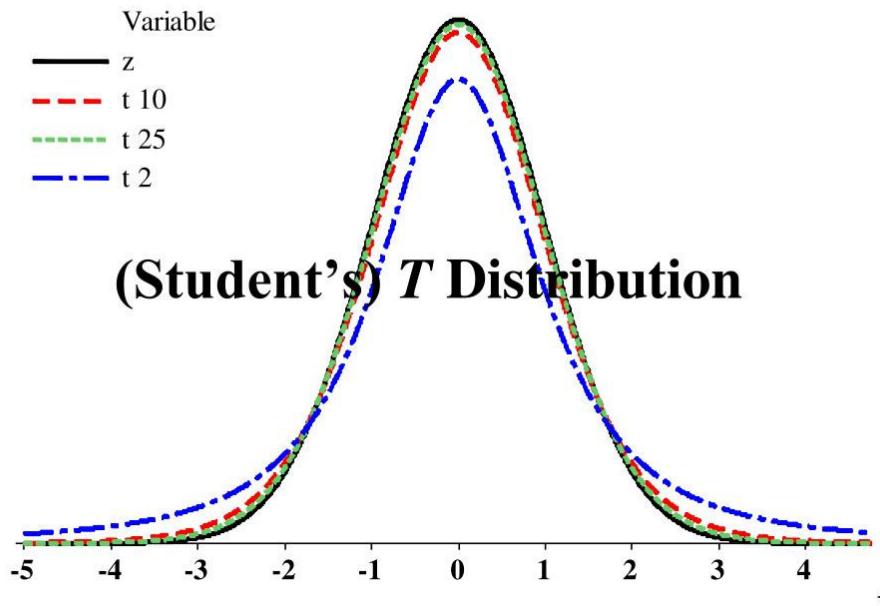
Wordt gebruikt om de tijd tussen twee gebeurtenissen te modelleren die met een constante gemiddelde intensiteit λ gebeuren.
Zie je het verband met de Poisson verdeling?



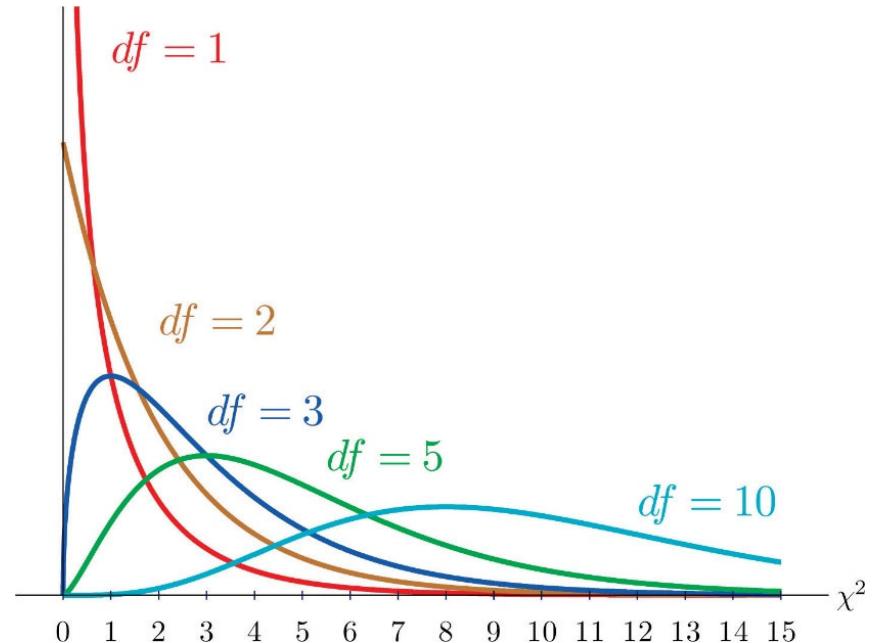
<https://docs.scipy.org/doc/scipy/reference/stats.html>

Verdelingsfunctie-'families'

Student

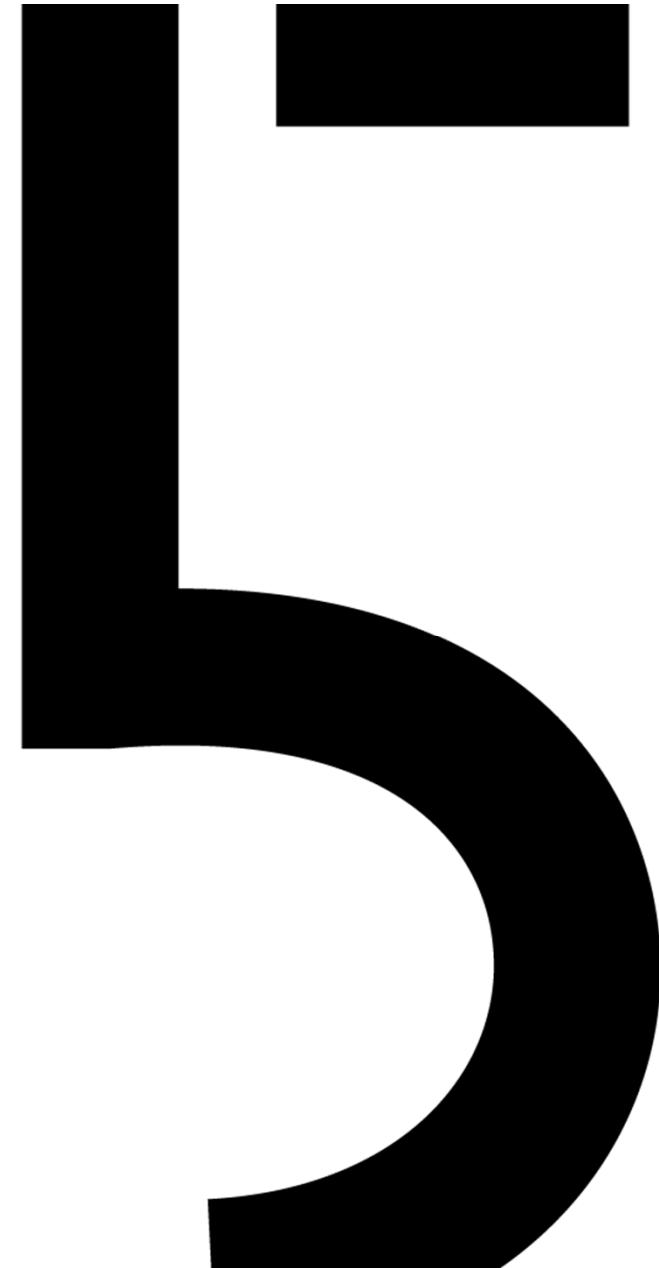


Chi-kwadraat



Opmerking: Gaan we in de komende lessen gebruiken

In de praktijk



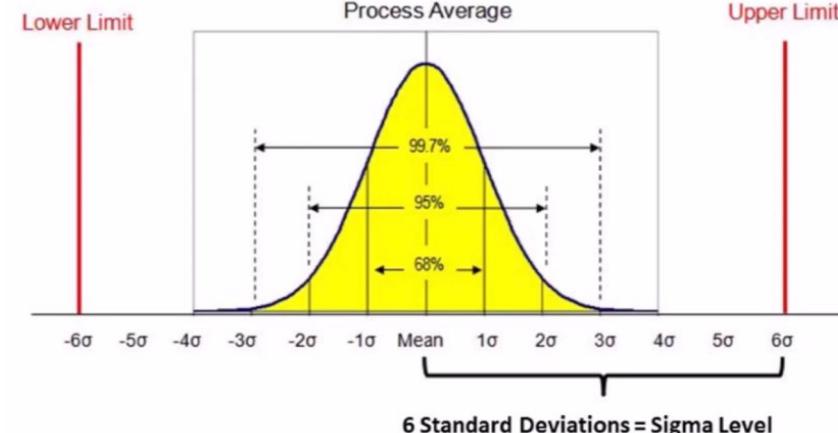
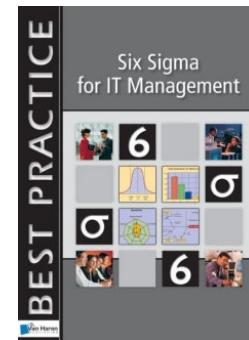
In de praktijk



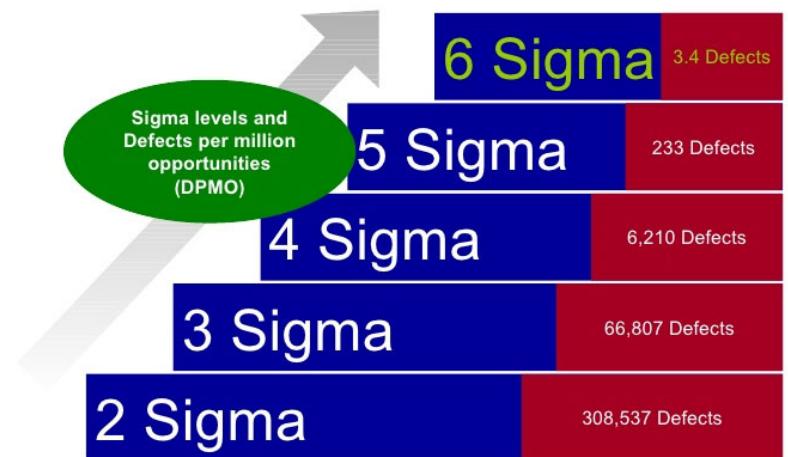
Six Sigma (6σ): een managementstrategie bestaande uit een verzameling kwaliteitsmanagementmethoden, voornamelijk van statistische aard waarmee men de kwaliteit van productie- en administratieve processen stap voor stap kan verbeteren. Deze methode werd in 1986 door Motorola in de VS ontwikkeld en deze wordt tegenwoordig in vele sectoren van het bedrijfsleven toegepast.

... moet de productie dusdanig worden ingericht dat idealiter 99.99966% van de geleverde onderdelen binnen deze marges valt, wat overeenkomt met zes standaardafwijkingen (= sigma) binnen een normale verdeling ...

Bron: https://nl.wikipedia.org/wiki/Six_Sigma



Path to Six Sigma





In de media



In de media

<http://deredactie.be/cm/vrtnieuws/wetenschap/1.2558287>

Formule toont aan dat samenzweringen de neiging hebben te mislukken

vr 29/01/2016 - 00:51 Luc De Roy

Het is uiterst moeilijk om een samenzwering verborgen te houden, omdat vroeg of laat een van de samenzweerders het geheim verraat. Een natuurkundige heeft nu een vergelijking opgesteld waaruit afgeleid kan worden hoelang een samenzwering kan overleven voor ze, opzettelijk of per ongeluk, onthuld wordt aan het grote publiek.

De vergelijking van doctor David Grimes, een natuurkundige aan de Oxford University, is gebaseerd op drie factoren: het aantal samenzweerders, hoeveel tijd er al verstrekken is sinds het begin van de samenzwering en de intrinsieke waarschijnlijkheid dat de samenzwering zal mislukken.



Een inventing tegen griep.

Grimes heeft vervolgens zijn formule toegepast op vier beroemde samenzweringstheorieën: het idee dat de maanlanding niet echt heeft plaatsgevonden, de theorie dat de opwarming van de aarde bedrog is, het geloof dat vaccinaties autisme veroorzaken, en de theorie dat de farmaceutische bedrijven een remedie tegen kanker hebben maar die achterhouden.

De analyse van Grimes toont aan dat als het in die vier gevallen inderdaad om samenzweringen zou gaan, die nu dan zeer waarschijnlijk al ontmaskerd zouden zijn als zodanig.

De "samenzwering" rond de maanlanding zou meer bepaald al na 3,7 jaar aan het licht zijn gekomen, de "fraude" rond de klimaatverandering na 3,7 tot 26,8 jaar, de "samenzwering" rond de inventingen en autisme na 3,2 tot 34,8 jaar, en de "kancersamenzwering" na 3,2 jaar.

"De mathematische methoden in deze studie zijn in grote mate gelijk aan de wiskunde die ik gebruik heb in mijn academisch onderzoek naar de fysische eigenschappen van straling", zei Grimes aan de BBC.

**DE
REDACTIE.BE**

Bestaande samenzweringen

Om tot zijn formule te komen, begon Grimes met de zogenoemde Poissonverdeling, een kansverdeling die een veel gebruikt statistisch instrument is dat de waarschijnlijkheid berekent dat een bepaalde gebeurtenis zal plaatsvinden binnen een bepaalde periode.

Aan de hand van een handvol veronderstellingen, in combinatie met mathematische deducties, kwam Grimes tot een algemene, maar onvolledige formule. Wat ontbrak was meer bepaald een goede schatting voor de intrinsieke waarschijnlijkheid dat een samenzwering zou mislukken. Om die vast te stellen, analyseerde Grimes de gegevens van drie echte samenzweringen.



Het NSA-hoofdkwartier in Maryland.

De eerste was het afluisterprogramma van de Amerikaanse National Security Agency (NSA), dat bekendstond als PRISM. Bij het programma waren maximaal 36.000 mensen betrokken en na zo'n zes jaar werd het aan het licht gebracht door klokkenluider Edward Snowden.

De tweede samenzwering was het Tuskegee-syfilisonderzoek, waarbij bijna 400 zwarte mannen die besmet waren met de geslachtsziekte syfilis, opzettelijk geen behandeling kregen. In totaal kunnen er 6.700 mensen betrokken geweest zijn bij het onderzoek, en na 25 jaar bracht dokter Peter Buxton het aan het licht.

De derde samenzwering was een schandaal bij de FBI, waarbij bleek dat de forensische onderzoeken van de dienst onwetenschappelijk en misleidend waren, wat resulteerde in het gevangenzetten en zelfs executeren van onschuldige mensen. Volgens Grimes waren er ten hoogste 500 mensen op de hoogte, en het duurde zo'n zes jaar voor het schandaal uiteklie.

"Best case scenario"

Grimes wijst erop dat de formule die hij opgesteld heeft, uitgaat van een "best case scenario", het best mogelijke scenario voor de samenzweerders. Het vertrekt namelijk van de veronderstelling dat de samenzweerders geen kletskousen zijn, maar integendeel goed zijn in het bewaren van geheimen, en dat de buitenwereld geen vermoeden heeft van de samenzwering en er dus geen onderzoek naar gevoerd wordt.

Op basis van de gegevens van de drie bekende samenzweringen, berekende Grimes dat de intrinsieke waarschijnlijkheid dat een samenzwering mislukt, vier op een miljoen is. Dat is een laag getal, maar de kans dat een samenzwering bekend raakt, wordt toch erg groot naarmate de tijd verstrijkt en het aantal samenzweerders toeneemt.



Zo begon de "maanlandingsamenzwering" in 1965, en er zouden 411.000 werknemers van de NASA bij betrokken zijn. Met deze parameters leidt de vergelijking van Grimes tot de conclusie dat de samenzwering na 3,7 jaar aan het licht gekomen zou zijn.

Bovendien volgt uit de formule dat, aangezien de samenzwering nu al meer dan 50 jaar geheim is gebleven, er niet meer dan 261 samenzweerders bij betrokken zouden kunnen zijn. Daarom is het redelijker om aan te nemen dat de maanlanding wel degelijk plaatsgevonden heeft.

Professor Monty McGovern, een wiskundige aan de University of Washington, zei aan de BBC dat de methode die in de studie gehanteerd wordt "zeer redelijk lijkt, en de waarschijnlijkheden die berekend worden, zijn zeer geloofwaardig".

Grimes zei dat hij hoopt dat zijn studie een aantal aanhangers van samenzweringstheorieën zal kunnen overtuigen, hoewel hij niet al te optimistisch is. "Hoewel ik denk dat het moeilijk, zo niet onmogelijk is, om mensen met een stellige overtuiging tot andere gedachten te brengen, hoop ik dat deze studie nuttig zal zijn voor mensen die minder overtuigd zijn, en die zich afvragen of geleerde bedrog zouden kunnen plegen of niet."

De studie van David Grimes is gepubliceerd in het online tijdschrift Plos One.

In de media

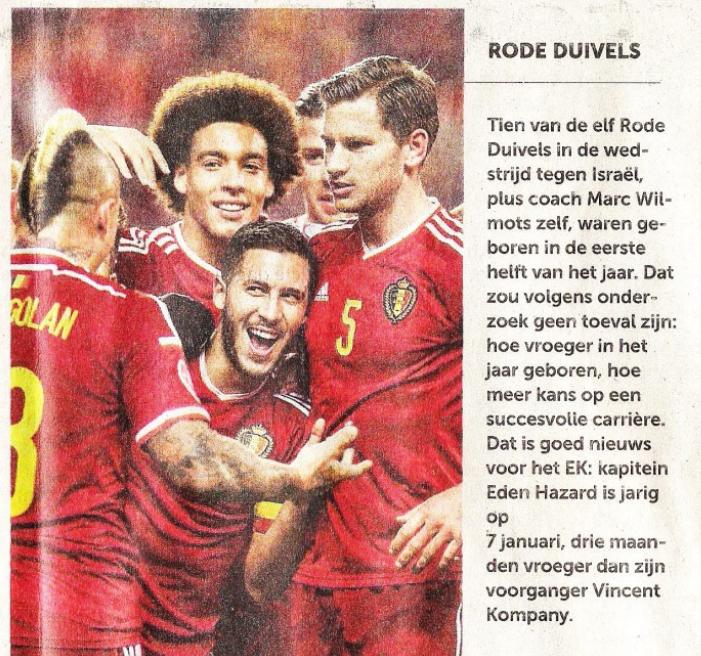
Onderzoek toont aan dat toeval een doorslaggevende rol speelt

Succes is iets voor GELUKZAKKEN

Wie een carrière als Rode Duivel ambieert, is het best in het begin van het jaar op de wereld gekomen. En Bill Gates was wellicht niet zo succesvol geweest als hij vijftig jaar eerder was geboren. Econoom Robert Frank is duidelijk: zonder een ferme dosis geluk kom je er niet.

ANN VAN DEN BROEK

Volledig artikel zie: De Morgen 23 mei 2016



Oefening: Wat is de kans dat 11 van de 12 personen in de eerste 6 maanden van het jaar jarig zijn?

Antwoord: We gebruiken de binomiale verdeling met $n=12$, $x=11$ en $p=0,5$:

>>> [binom.pmf\(11,12,1/2\)](#)

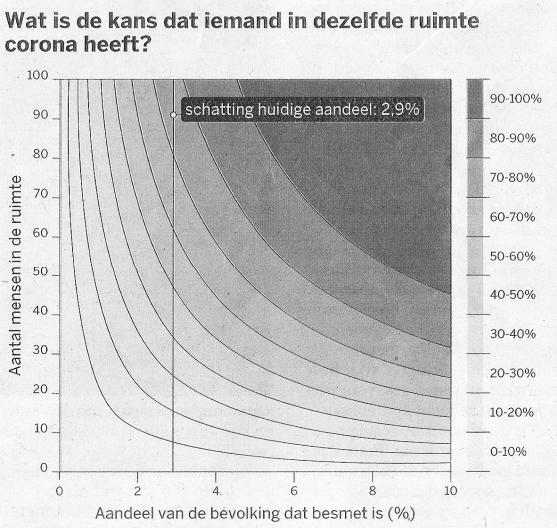
= 0,0029

In de media

Extra Oefening:

- Lees het artikel volledig.
- Welke kansverdeling wordt er gebruikt?
- Maak zelf de berekeningen en vergelijk jouw bekomen resultaten met de resultaten in het artikel.

Hoe groot is de kans dat iemand in dezelfde ruimte corona heeft?



De Standaard – 6 juli 2022

RISICOBEPERKING Het aantal coronabesmettingen neemt fors toe. Maar hoe groot is de kans dat er in dezelfde ruimte iemand zit die besmet is? We rekenden het voor u na.

Er raast een zomergolf door het land, maar grote averij veroorzaakt die niet. Op dit moment zijn er dagelijks 119 ziekenhuisopnames en 8 overlijdens door covid. Er liggen 73 patiënten met covid op intensieve zorg. Maar toch: ook als u de ergste scenario's ontloopt, loopt u liever geen besmetting op. Niemand is graag ziek, zeker niet met een vakantie of festival in het vooruitzicht.

Maar hoe groot is eigenlijk de kans dat iemand die zich in dezelfde ruimte bevindt, besmet is? Dat hangt af van drie factoren, stipte de Britse wiskundige bioloog Kit Yates (Universiteit van Bath) enkele dagen geleden op Twitter aan: hoeveel mensen er op dit moment besmettelijk zijn, hoe trouw mensen die besmet zijn in isolatie gaan en het aantal mensen dat in dezelfde ruimte zit.

Het aantal mensen dat op dit moment corona heeft, is niet zo eenvoudig te bepalen. Volgens de laatste gegevens van Sciensano testten in de voorbije week 39.398 mensen positief. Dat is een forse onderschatting, want lang niet iedereen laat zich nog officieel testen. Volgens onderzoekers van de Antwerpse en Hasseltse universiteit waren er begin mei ongeveer 8,5 keer meer besmettingen dan de statistieken aangaven. Als dat nog steeds klopt, waren er in de laatste week ruim 330.000 besmettingen. Dat is net geen 3 procent van de bevolking.

Hoeveel mensen gaan netjes in

isolatie bij een besmetting? Laten we even, net zoals Yates, uitgaan van de helft. Dat lijkt weinig, maar heel wat mensen weten niet dat ze besmet zijn.

Het aantal mensen waarmee u de ruimte deelt, kunt u zelf tellen (of schatten bij grotere groepen).

Grote groepen = bingo

Gebruiken we deze drie parameters, dan kunnen we zelf beginnen kansrekenen. Als 3 procent van de bevolking corona heeft en de helft zich isolateert, dan heb je 50 procent kans dat er minstens één besmette

Zit je in een vliegtuig met 200 mensen, dan heb je 95 procent kans dat iemand corona heeft

bij is als er 45 mensen in de ruimte zijn. Bij een groep van 19 mensen bedraagt de kans 25 procent. Bij grote groepen mag u er dus al snel van uitgaan dat er minstens één besmette aanwezig is. Zit je op een vliegtuig met 200 mensen bijvoorbeeld, dan heb je 95 procent kans dat iemand corona heeft.

Dat wil natuurlijk niet zeggen

dat u 95 procent kans loopt om in een vliegtuig besmet te worden. Het is niet omdat er minstens één besmette is, dat hij ook het virus aan u doorgeeft. Dat hangt onder meer af van hoe besmettelijk die persoon is, hoelang u in dezelfde ruimte zit, of die ruimte goed geventileerd is, hoe dicht u zit, of er een mondmasker gedragen werd, en zo ja: welk mondmasker. Dat zijn net iets te veel factoren om een snelle berekening te maken, maar in elk geval ligt de kans op besmetting veel lager dan de kans dat u met een besmet persoon in één ruimte zit.

Op de trein

Uit de berekeningen zou u kunnen afleiden dat het risico bij kleine groepen verwaarloosbaar laag is. Dat kan tegenvallen. Neem nu het eerste rijtuig van de trein die gisterochtend om negen uur van Gent naar Brussel reed. Daarin zaten twaalf mensen. Als 3 procent van de bevolking besmet is en de helft gaat (tijdig) in isolatie, dan is de kans 17 procent dat er minstens één besmette in die ruimte zat. Dat lijkt mee te vallen. Maar als u elke weekdag diezelfde trein neemt, heen en terug, dan stijgt de kans dat u op minstens één van uw treinritten het rijtuig deelt met een besmette tot 84 procent.

Reken daarbij de ontmoetingen op het werk, de supermarkt of op vakantie, en u begrijpt dat het zo goed als onmogelijk is om besmette medemensen te blijven ontlopen. Tenzij u thuisblijft.

Dries De Smet



Vragenlijst

Vragenlijst



- Download het bestand *vragenlijst 21-22.xlsx* van Canvas
- Exporteer het excel-bestand als een csv bestand
- Plaats *vragenlijst 21-22.csv* in je Python workspace
- Lees de data in en plaats het in het dataframe
studenq

```
>>> import pandas as pd  
>>> studenq = pd.read_csv('vragenlijst 21-22.csv', delimiter=';',  
decimal=',')
```

Vragenlijst



Veronderstel dat we met de vragenlijst **ALLE** studenten van INF1 hebben ondervraagd.

1. Wat is de kans dat van 10 willekeurig gekozen studenten, exact 3 studenten hun bloedgroep kennen?

Vragenlijst



Veronderstel dat we met de vragenlijst **ALLE** studenten van INF1 hebben ondervraagd.

2.a Wat is de kans dat een student zijn studies onderschat (m.a.w. inschat dat 1 studiepunt overeenkomt met 15 of minder uren) ?

2.b Je dient samen met vier medestudenten een groeps-werk te maken. Wat is de kans dat er minstens 1 van je groepsgenoten zijn studies onderschat?

Vragenlijst



Veronderstel dat we met de vragenlijst **ALLE** studenten van INF1 hebben ondervraagd.

3. Bereken op basis van de gegevens hoeveel % van de studenten maximaal anderhalve standaardafwijking groter of kleiner zijn dan het gemiddelde.

Hoeveel % verwacht je voor een normale verdeling?

Oefeningen





KdG Karel de Grote
Hogeschool