

Data Science 2

Toetsen – deel 1

Wim De Keyser

Geert De Paepe

Jan Van Overveld

KdG Karel de Grote
Hogeschool

Quote van de week

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write"

Herbert George Wells (1866-1946)

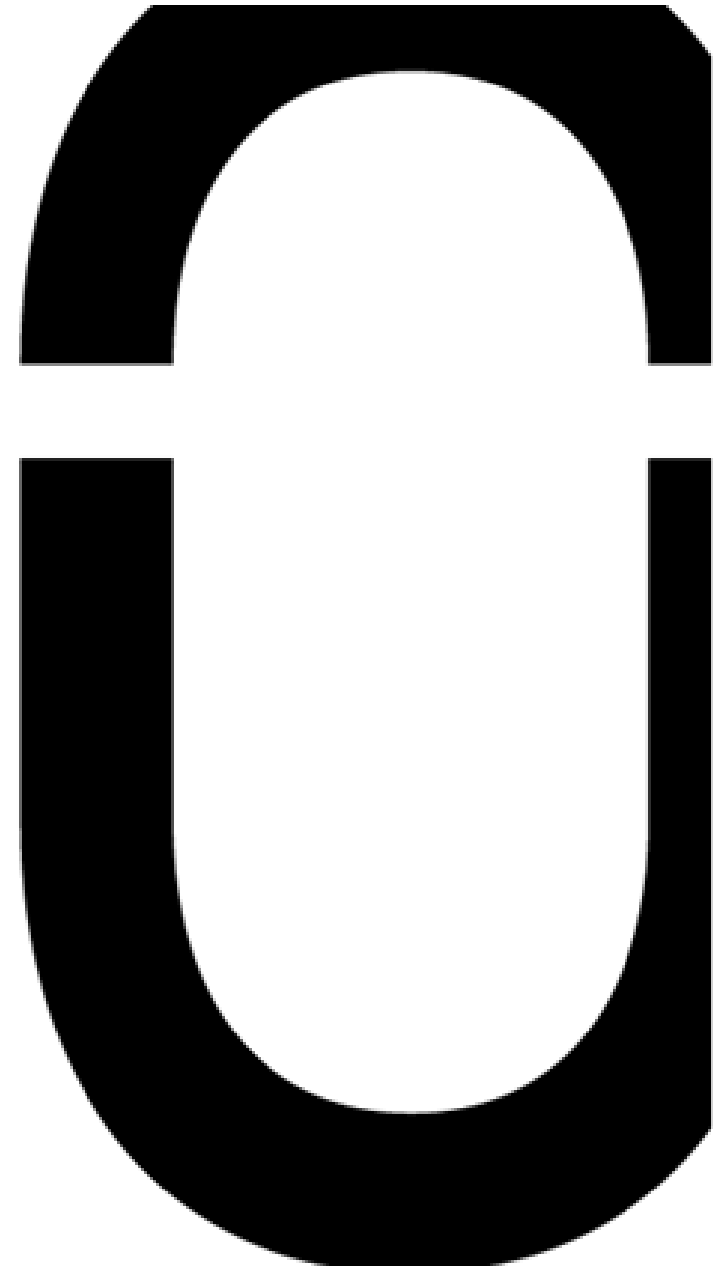


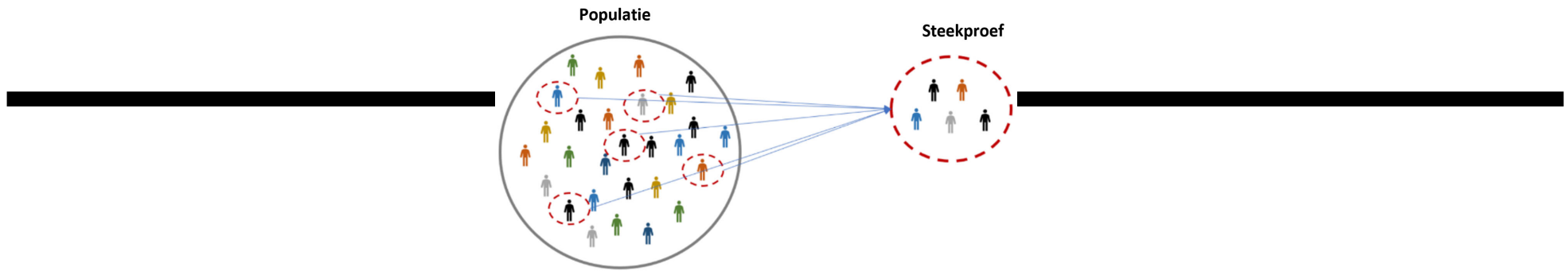
Agenda



1. Voorbeeld
2. Betrouwbaarheidsintervallen
 - Normaalverdeling
 - Student-verdeling
3. Toetsen van hypothesen
 - Hypothese
 - t-toets
 - Z-toets
4. In de media

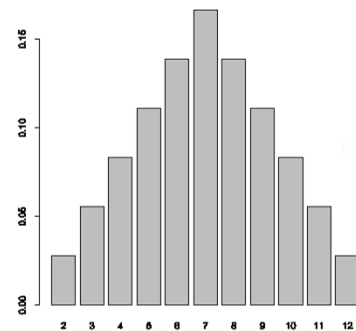
Herhaling: Kansverdelingen





Theoretisch oneindig experiment, **Populatie**

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



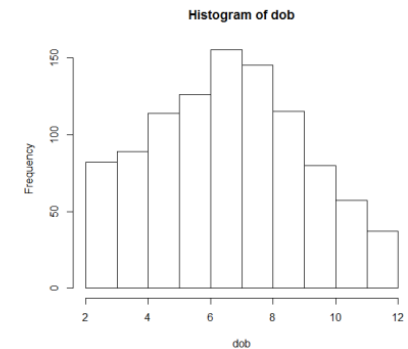
kansverdeling (kansen)

⇒ Verwachte waarde: μ

⇒ Standaardafwijking: σ

Experiment, **Steekproef**

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



Frequentietabel (relatieve frequenties)

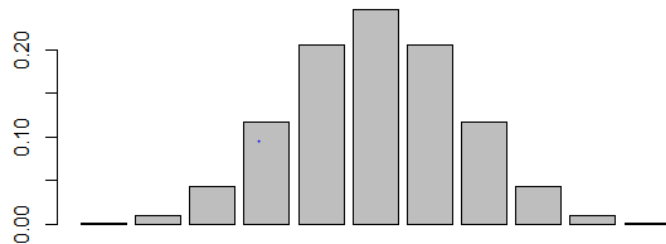
⇒ Gemiddelde: \bar{x}

⇒ Standaardafwijking: s

De binomiale verdeling

"Kans op x maal succes op n experimenten (en p is de kans op succes bij 1 experiment)"

- Discrete verdeling
- Parameters: n en p



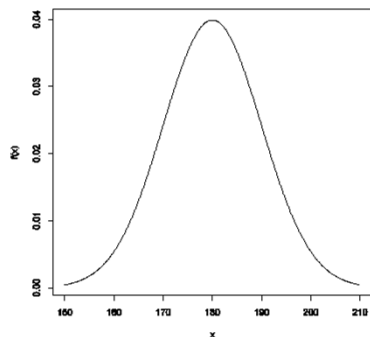
```
>>> from scipy.stats import binom
>>> binom.pmf(x,n,p) // P(x)
>>> binom.cdf(x,n,p) // P(0)+..+P(x)
```

• Voorbeelden:

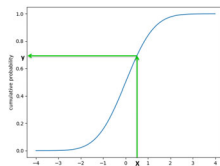
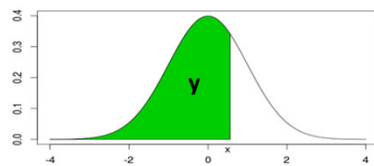
- "Wat is de kans op 3 keer kop wanneer je 5 keer met een muntstuk gooit?"
- "Wat is de kans om te slagen op een meerkeuze examen (20 vragen, telkens 4 keuzemogelijkheden, geen giscorrectie) wanneer je niets van de leerstof kent?"
- Wat is de kans dat 10 van de 11 voetballers op het veld in de eerste 6 maanden van het jaar jarig zijn?

De normale verdeling

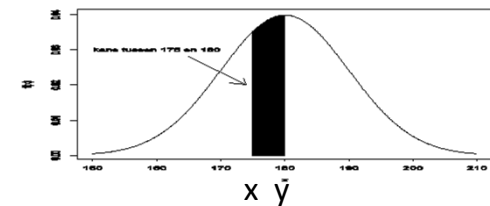
- Kansen symmetrisch verdeeld
- Continue verdeling
- Parameters: μ en σ



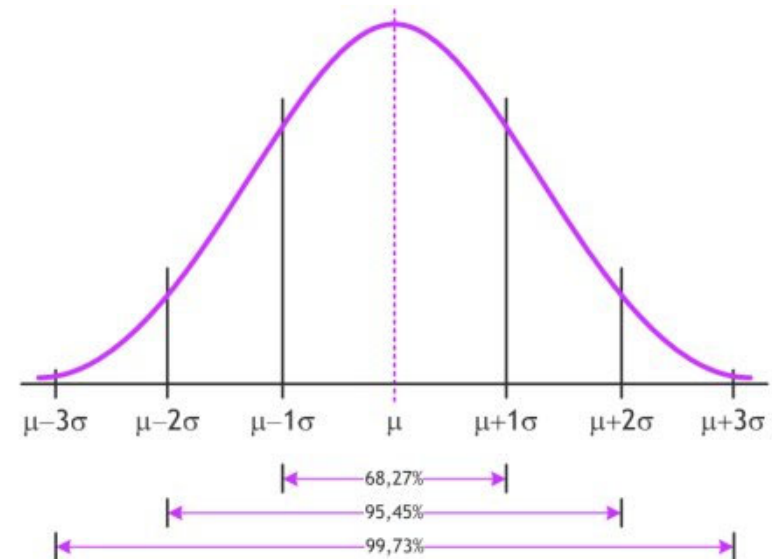
```
>>> from scipy.stats import norm
>>> norm.cdf(x, loc= $\mu$ , scale= $\sigma$ )
```



```
>>> norm.cdf(y, loc= $\mu$ , scale= $\sigma$ ) -
norm.cdf(x, loc= $\mu$ , scale= $\sigma$ )
```



- Eigenschappen:

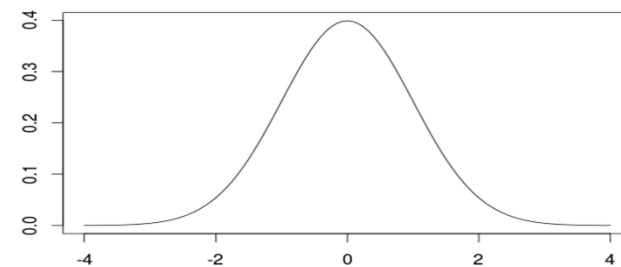


- Voorbeelden normale verdeling:

↳ "Wat is de kans dat een persoon tussen de 175 en 185 cm meet wanneer je weet dat de populatie normaal verdeeld is met $\mu = 170$ en $\sigma = 5,3$?"

De standaard normale verdeling

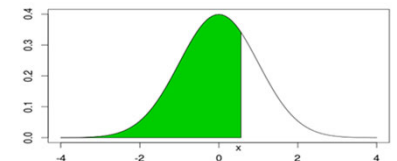
- Normale verdeling met als parameters: $\mu = 0$ en $\sigma = 1$



- normale verdeling omzetten in de standaardnormale:

$$Z = (X - \mu) / \sigma$$

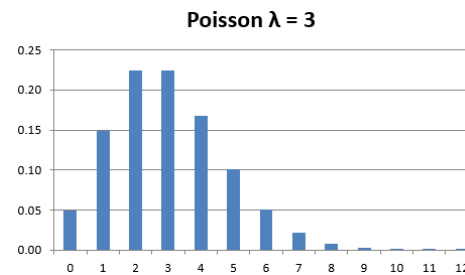
>>> `norm.cdf(x)`



De Poisson verdeling

"Kans op x voorvallen gedurende een gegeven tijdsinterval (of afstand, of volume, of...) wanneer er zich gemiddeld λ voorvallen voordoen"

- Discrete verdeling
- Parameters: λ



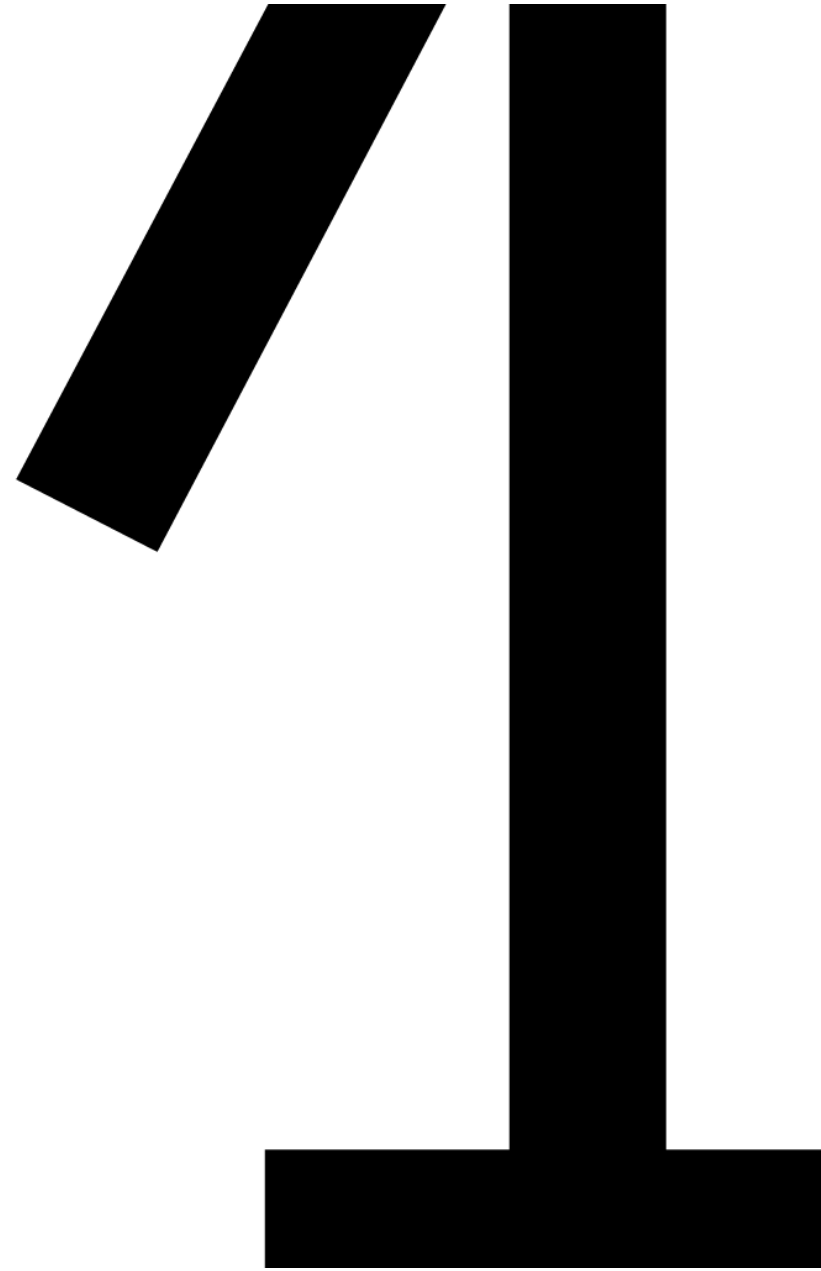
```
>>> from scipy.stats import poisson
>>> poisson.pmf(x,  $\lambda$ ) // P(x)
>>> poisson.cdf(x,  $\lambda$ ) // P(0)+..+P(x)
```

- Voorbeelden:

⇒ "Wat is de kans op 400 mails op een dag wanneer je gemiddeld 300 mails per dag ontvangt?"

⇒ "Wat is de kans op meer dan 25 branden in een gemeente met 9.000 gebouwen wanneer er 10.000 branden per jaar zijn in België met 4.500.000 gebouwen?"

Voorbeeld



Voorbeeld

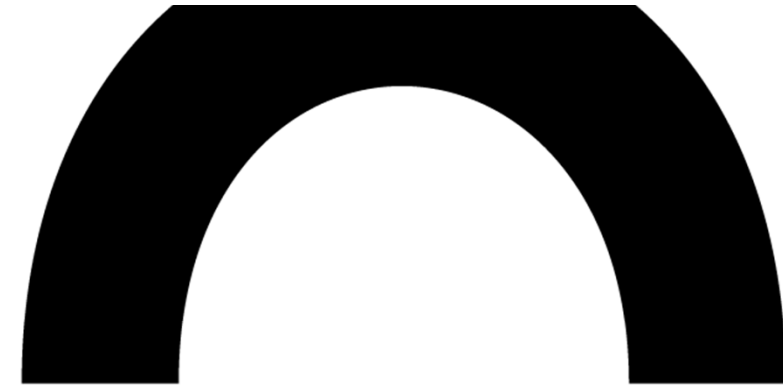
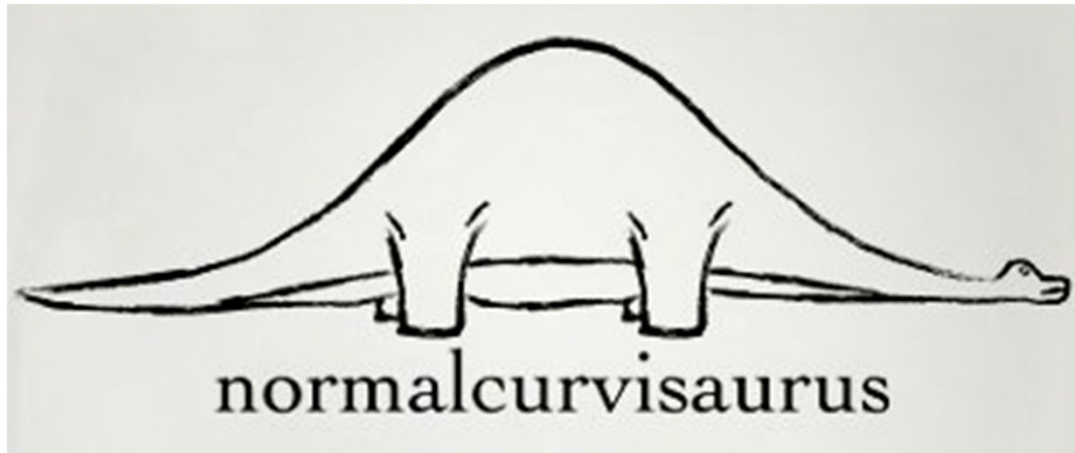
We voeren een kwaliteitscontrole uit in een fabriek van laptops. We meten (o.a.) het verbruik van laptops:

- aantal laptops aselekt getest: $n = 30$
- gemiddeld: $\bar{x} = 40$ Watt
- standaardafwijking: $s = 20$ Watt

Vraag: Wat zegt dit over het verbruik van alle geproduceerde laptops?

Eigenschappen

- Hoe meer laptops we testen, hoe zekerder we worden
- Hoe groter de standaardafwijking van onze steekproef, hoe onzekerder we worden
- We weten 100% zeker dat het gemiddelde verbruik van de geproduceerde laptops tussen $-\infty$ en $+\infty$ Watt ligt
- We zijn minder zeker dat het verbruik tussen 30 en 50 Watt ligt
- We zijn redelijk zeker dat het verbruik van laptops niet boven de 1000 Watt ligt
- We zijn 100% zeker dat het gemiddelde verbruik niet exact 40,000000... Watt is



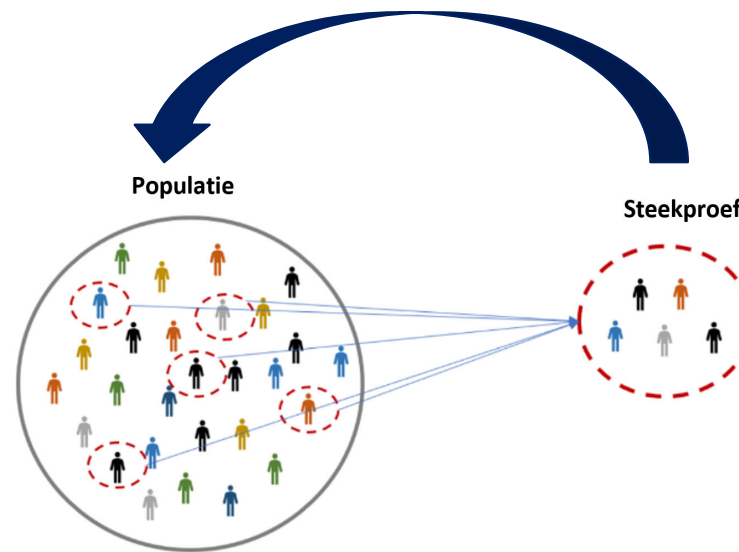
Betrouwbaarheidsintervallen

- **Normaalverdeling**
- Student-verdeling



Probleemstelling

“We willen op basis van een **steekproef** iets zeggen over de **populatie**”



⇒ **Betrouwbaarheidsinterval**

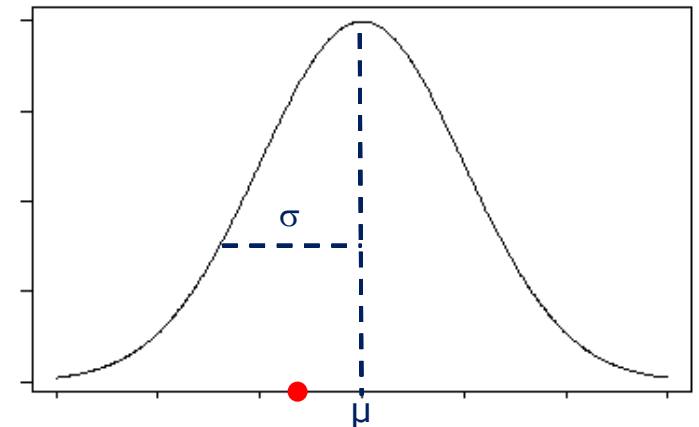
Tussen welke grenzen, bekomen adhv de parameters van een steekproef, ligt een bepaalde parameter van de populatie? Hoe 'zeker' is deze uitspraak?

Betrouwbaarheidsinterval

- Stel dat het verbruik van laptops normaal verdeeld is

- gemiddelde is μ
- standaardafwijking is σ

- Kies een willekeurige 1 laptop



- Wat is de kans dat het verbruik tussen x en y ligt?

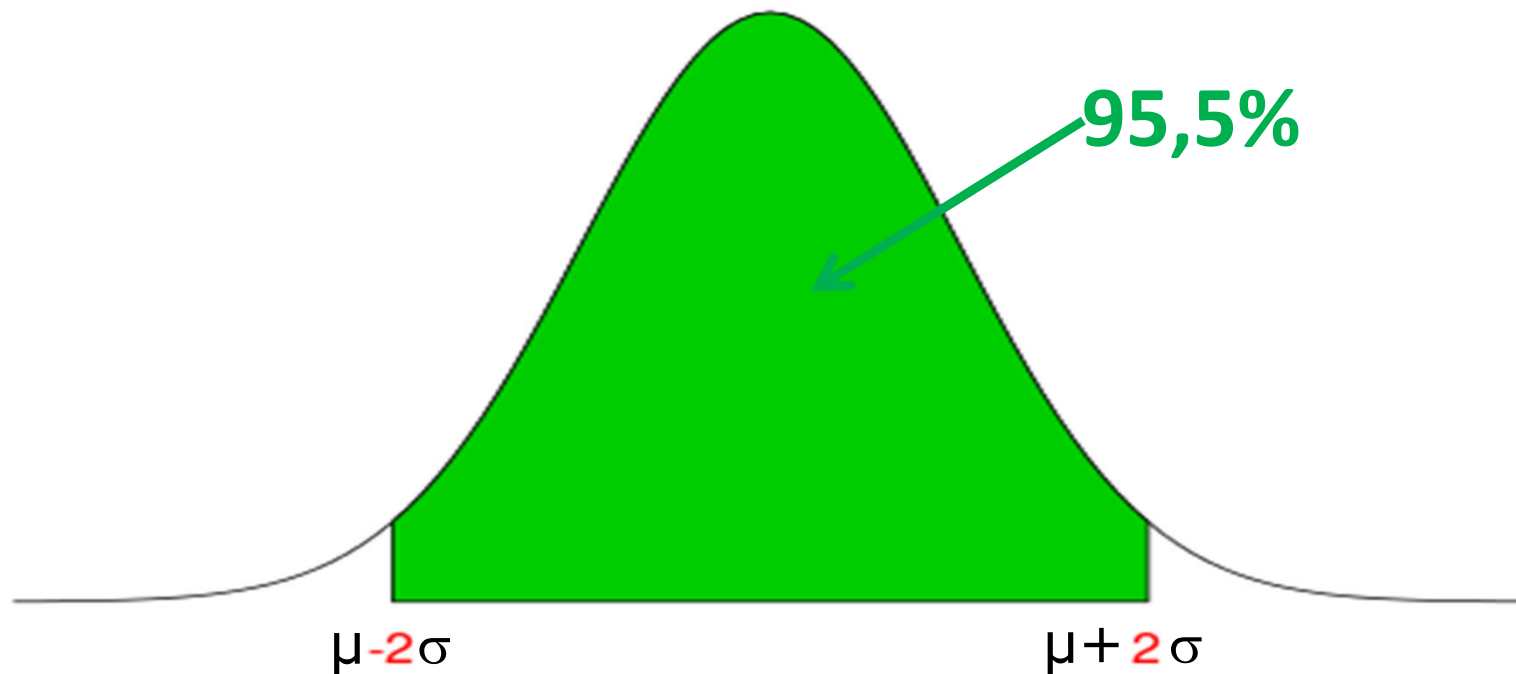
- Hoe zou je dit berekenen?

```
>>> norm.cdf(y, loc=  $\mu$ , scale=  $\sigma$ )  
- norm.cdf(x, loc=  $\mu$ , scale=  $\sigma$ )
```

- Wat is de kans dat het verbruik tussen $\mu - 2\sigma$ en $\mu + 2\sigma$ ligt?

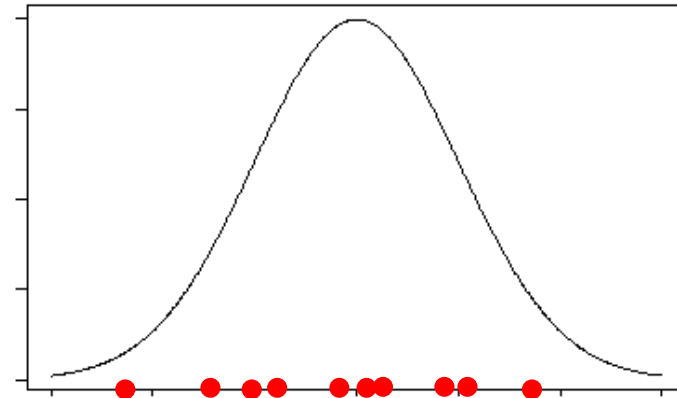
Betrouwbaarheidsinterval

- Wat is de kans dat het verbruik tussen $\mu - 2\sigma$ en $\mu + 2\sigma$ ligt?



Betrouwbaarheidsinterval

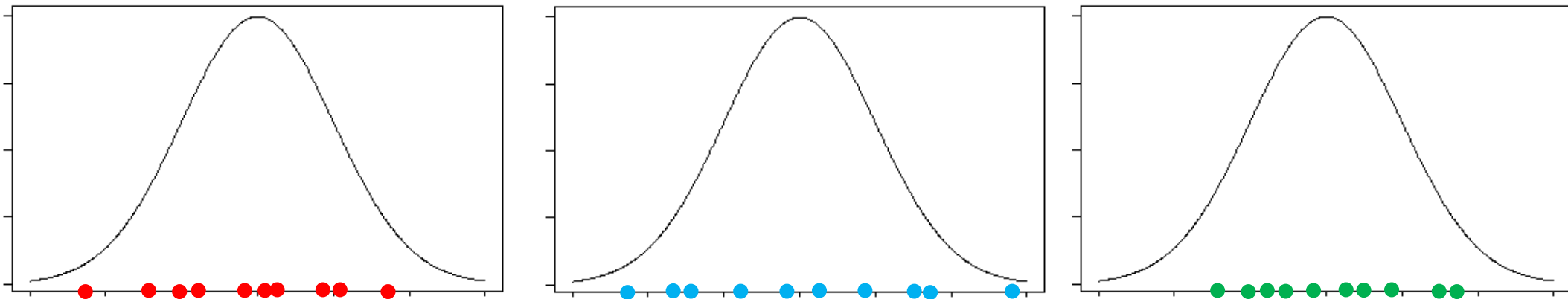
Stel dat je meer dan 1 laptop willekeurig kiest (je kiest er n) maw je neemt een steekproef:



- Neem het gemiddelde verbruik van die n laptops
- Is er meer of minder kans om het echte gemiddelde μ uit te komen?
- Wat zegt dit over de standaardafwijking?

Betrouwbaarheidsinterval

Stel dat je meerdere steekproeven neemt (telkens kies je aselect n laptops):



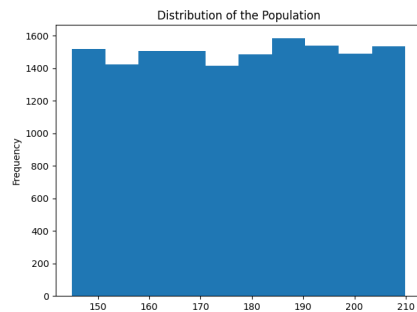
- Levert elke steekproef hetzelfde gemiddelde verbruik van n laptops op? Er werd toch gestart vanuit dezelfde populatie?
- Wat is de verdelingsfunctie voor het gemiddelde bekomen aan de hand van een steekproef?

Betrouwbaarheidsinterval

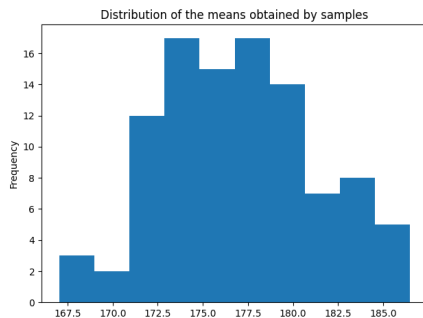
- **"Centrale limietstelling"** –zie simulatie-
 - steekproef met grootte n (> 30)
 - de kans om een bepaald gemiddelde uit te komen is bepaald door een normaalverdeling met als
 - gemiddelde μ
 - standaardafwijking σ/\sqrt{n}
- dus 95,5% kans dat: $\mu - 2 \sigma / \sqrt{n} < \bar{x} < \mu + 2 \sigma / \sqrt{n}$
- en dus is er 95,5% kans dat: $\bar{x} - 2 \sigma / \sqrt{n} < \mu < \bar{x} + 2 \sigma / \sqrt{n}$
- we benaderen σ door s : $\bar{x} - 2 s / \sqrt{n} < \mu < \bar{x} + 2 s / \sqrt{n}$

Betrouwbaarheidsinterval – “Centrale limietstelling” simulatie

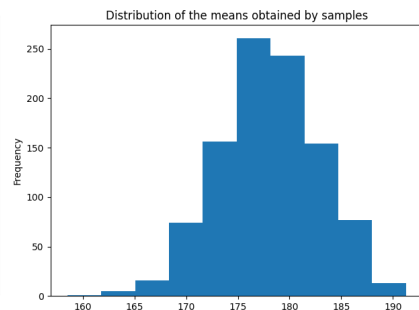
Uniforme verdeling
(= random binnen een interval van waarden)



100 keer een
steekproef
van 15 waarden



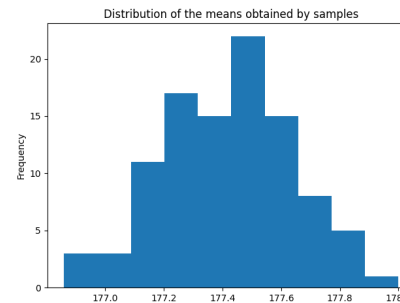
1000 keer een
steekproef
van 15 waarden



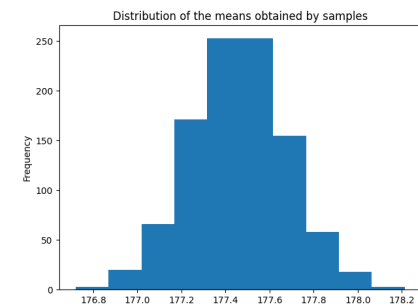
Normale verdeling
(= random binnen een interval van waarden)



100 keer een
steekproef
van 15 waarden

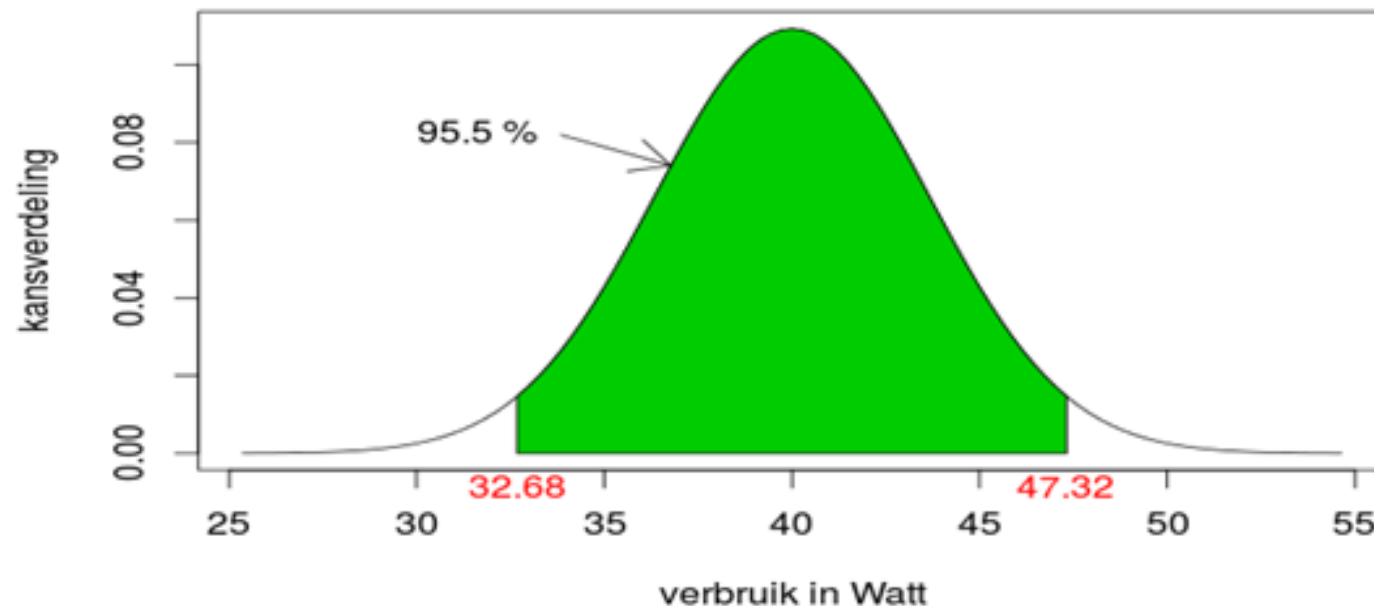


1000 keer een
steekproef
van 15 waarden



Betrouwbaarheidsinterval


- in ons voorbeeld: 30 laptops, $\bar{x} = 40$, $s = 20$
- 95,5% kans dat het gemiddelde verbruik van alle laptops ligt tussen:
 - $40 - 2 \cdot 20 / \sqrt{30} = 32,68$
 - $40 + 2 \cdot 20 / \sqrt{30} = 47,32$



Andere Betrouwbaarheidsintervallen

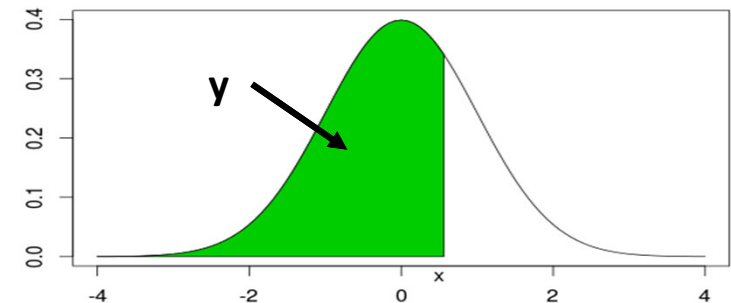
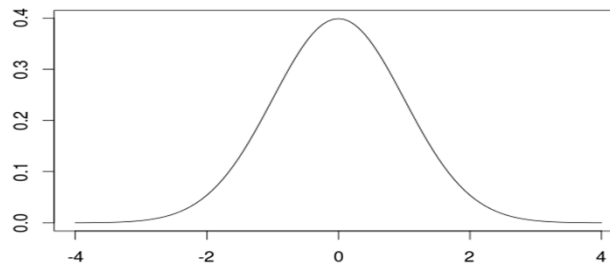
- Veel voorkomende betrouwbaarheidsintervallen:
 - 90% tussen $\bar{x} - 1,645 * s / \sqrt{n}$ en $\bar{x} + 1,645 * s / \sqrt{n}$
 - 95% tussen $\bar{x} - 1,96 * s / \sqrt{n}$ en $\bar{x} + 1,96 * s / \sqrt{n}$
 - 95,5% tussen $\bar{x} - 2 * s / \sqrt{n}$ en $\bar{x} + 2 * s / \sqrt{n}$
 - 99% tussen $\bar{x} - 2,576 * s / \sqrt{n}$ en $\bar{x} + 2,576 * s / \sqrt{n}$
 - 99,7% tussen $\bar{x} - 3 * s / \sqrt{n}$ en $\bar{x} + 3 * s / \sqrt{n}$
- Meestal gebruikt men het 95% betrouwbaarheidsinterval
- Je kan in Python de **factor** berekenen met:
 - >>> `norm.ppf((1+p)/2)`
 - of: >>> `norm.ppf(1-α/2)`

p = 0,95
of
α = 0,05



Samenvatting functies (scipy.stats)

- `norm.pdf`

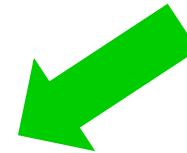
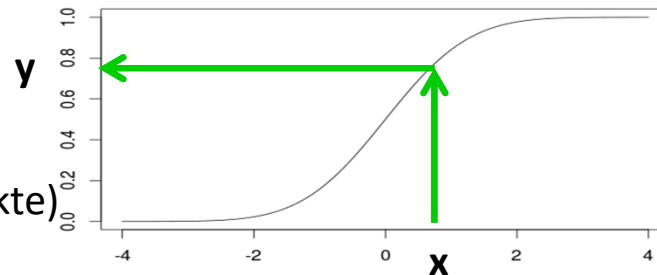


- `norm.cdf`

Parameter: z-score

Resultaat: kans (= oppervlakte)

E.g.: `>>> norm.cdf(1.645)`

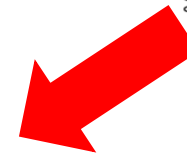
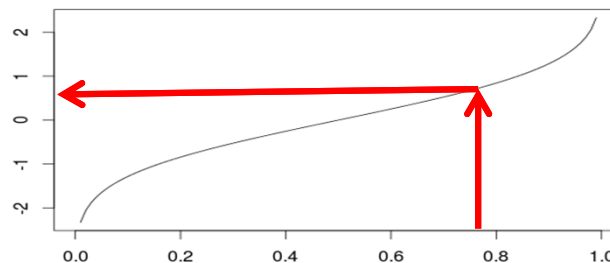


- `norm.ppf`

Parameter: kans (= oppervlakte)

Resultaat: z-score

E.g.: `>>> norm.ppf(0.95)`



Betrouwbaarheidsinterval

Met Python:

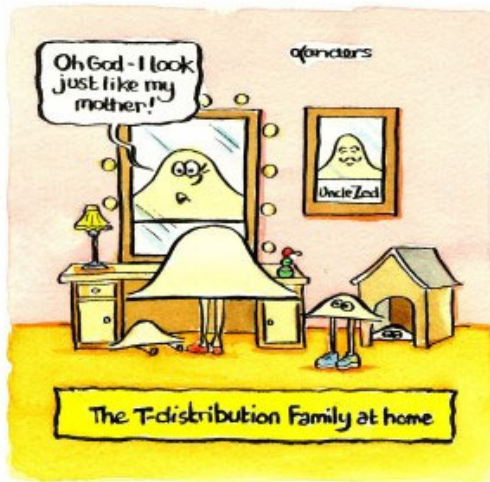
```
>>> df_c = pd.read_csv('consumptionLaptops.csv', decimal='.')

>>> x_bar = df_c.consumptionLaptops.mean()
>>> s = df_c.consumptionLaptops.std()
>>> n = len(df_c)

>>> factor = norm.ppf((1+0.955)/2)
>>> interval = (x_bar-factor*s/math.sqrt(n), x_bar+factor*s/math.sqrt(n))

>>> #OR

>>> norm.interval(confidence=0.955, loc=x_bar, scale=s/math.sqrt(n) )
```

Betrouwbaarheidsintervallen

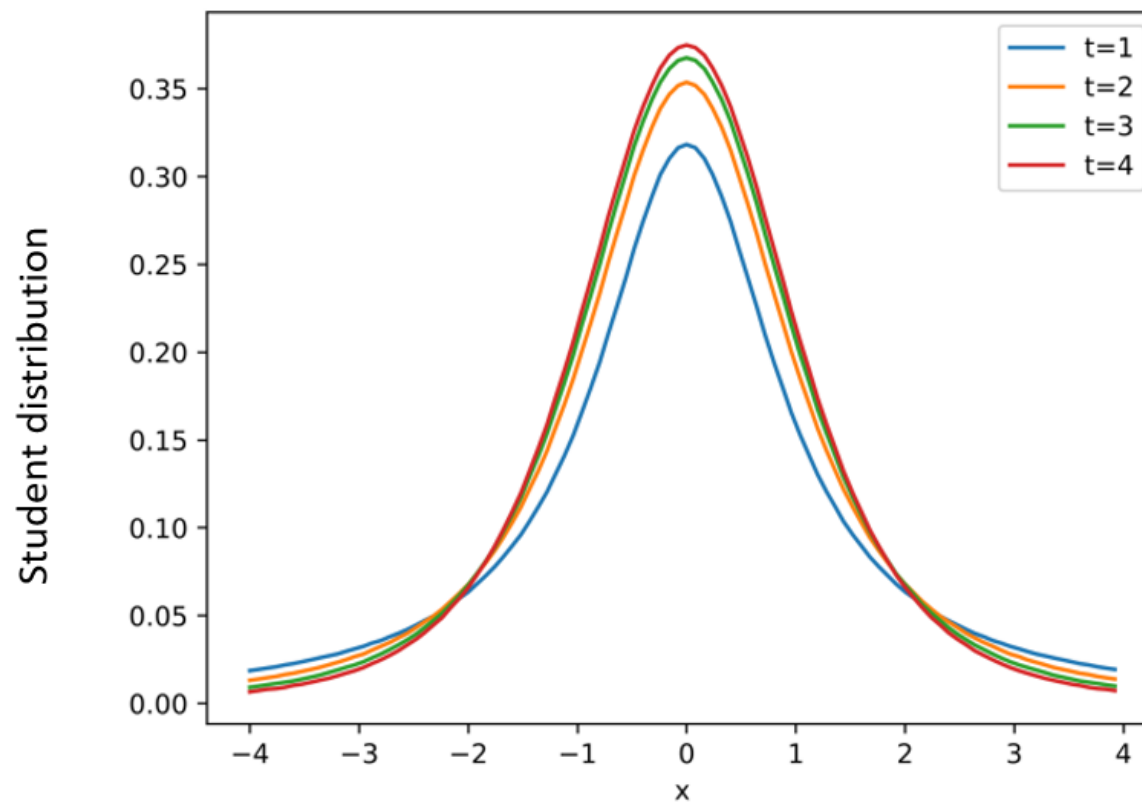
- Normaalverdeling
- **Student-verdeling**

De student-verdeling

- Probleem: we benaderen σ door s
- Als n klein is (<30) is s een slechte benadering
⇒ normaalverdeling is niet goed
- Gebruik t-verdeling (student-verdeling)
- Een t-verdeling heeft een extra parameter: steeds gelijk aan $n-1$
- In Python:

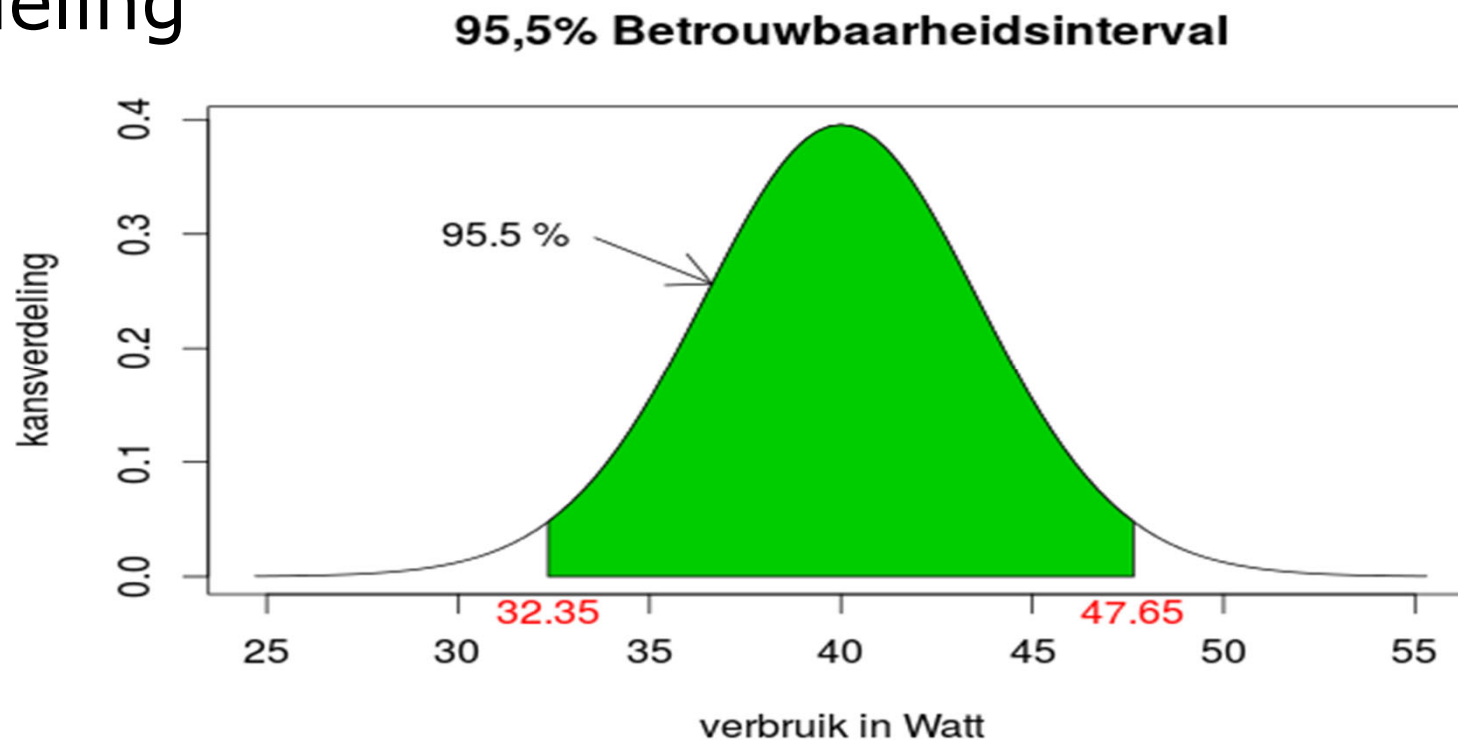
```
>>> from scipy.stats import t  
>>> t.cdf(x,df, loc= $\mu$ , scale= $s$ ) #s as approx. of  $\sigma$   
>>> t.ppf(q,df, loc= $\mu$ , scale= $s$ ) #s as approx. of  $\sigma$ 
```

De student-verdeling



De student-verdeling

- berekeningen zijn gelijkaardig met de normale verdeling



De student verdeling - Betrouwbaarheidsinterval

Met Python:

```
>>> df_c = pd.read_csv('consumptionLaptops.csv', decimal='.')

>>> x_bar = df_c.consumptionLaptops.mean()
>>> s = df_c.consumptionLaptops.std()
>>> n = len(df_c)

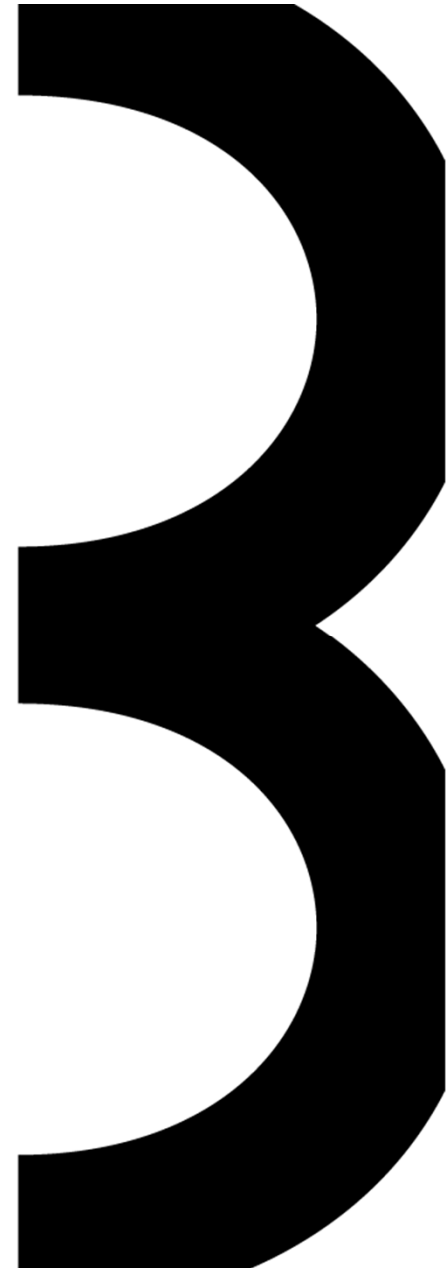
>>> factor = t.ppf((1+0.955)/2, n-1, loc= x_bar, scale=s/math.sqrt(n))
>>> interval = (x_bar-factor*s/math.sqrt(n), x_bar+factor*s/math.sqrt(n))

>>> #OF:
>>> t.interval(confidence=0.955, df=n-1, loc=x_bar, scale=s/math.sqrt(n))
```



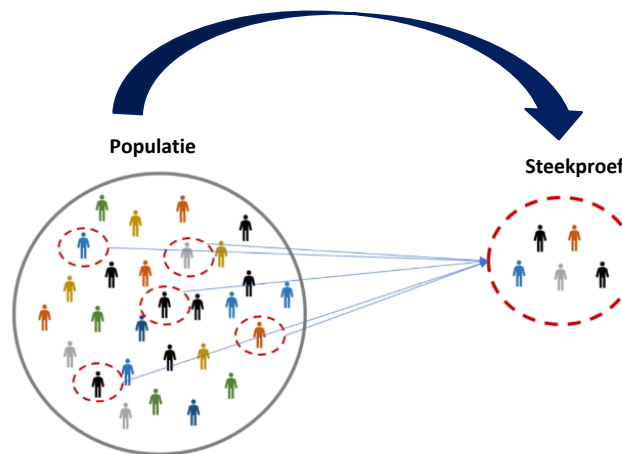
Toetsen van hypothesen

- **Hypothese**
- t-toets
- Z-toets



Probleemstelling

"Kunnen we een bewering over de **populatie** (= hypothese) weerleggen aan de hand van een **steekproef**?"



⇒ **Aanvaardingsinterval**

Valt een steekproefparameter binnen bepaalde grenzen, bekomen adhv de bewering over de populatie? Hoe 'zeker' is deze uitspraak?

Hypothese

- Bij toetsen gaan we uit van een veronderstelling (mbt de populatie) en proberen we na te gaan (adhv een steekproef) of we die kunnen bevestigen of verwerpen
- **Hypothese** = Veronderstelling
- **We proberen de hypothese tegen te spreken**
 - Als we een meting doen die in de lijn ligt van de veronderstelling, hebben we niks bewezen
 - Als we een meting doen die de hypothese tegenspreekt, is er veel kans dat deze niet waar is.

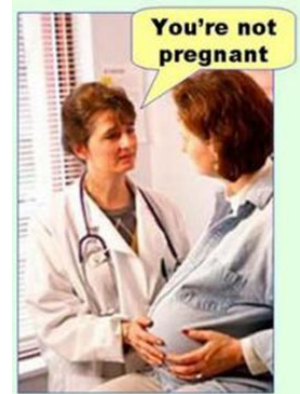
Hypothese

- 2 hypothesen
 - H_0 : **nul-hypothese**
 - H_1 : **alternatieve hypothese**
- Een besluit is **nooit 100% zeker**
- Er is steeds een “**significantieniveau**” : α
 - de kans dat ons besluit niet juist is
 - meestal is $\alpha=0,05$

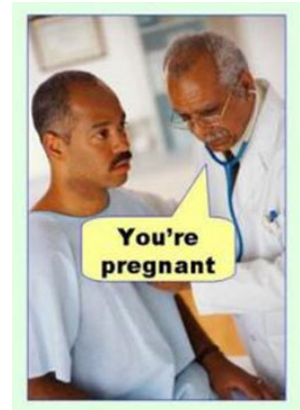
Hypothese

- Als $\alpha=0,05$, dan is er 5% kans dat we een fout maken:

Fout bij het verwerpen van H_0 terwijl die toch waar is (**type I-fout**): α



- Maar er is ook een fout bij het aanvaarden van H_0 terwijl die niet waar is (**type II-fout**): β



Hypothese

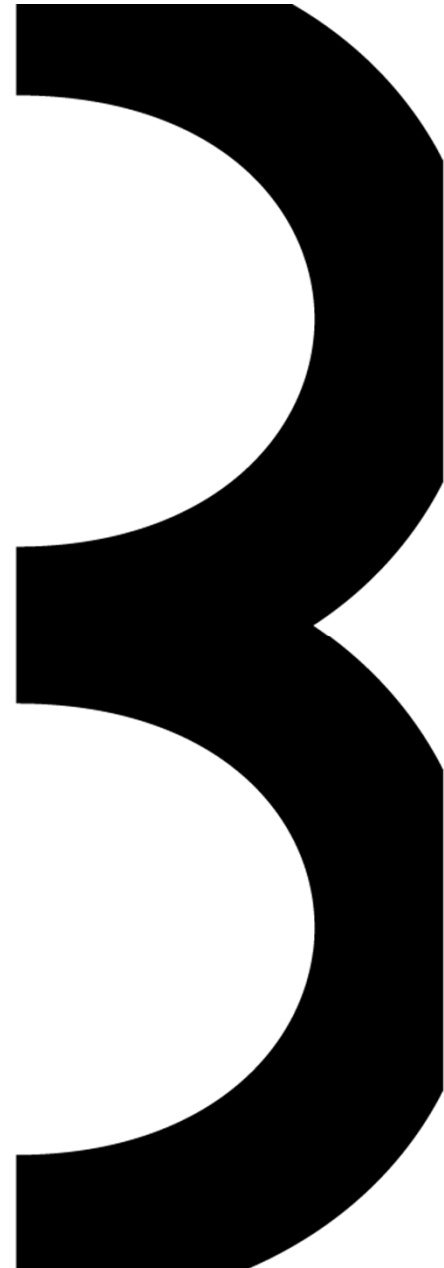
Table of error types		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	Correct inference (true positive) (probability = $1 - \alpha$)	Type II error (false positive)) (probability = β)
	Reject	Type I error (false negative) (probability = α)	Correct inference (true negative) (probability = $1 - \beta$)

The t-test:



Toetsen van hypothesen

- Hypothese
- **t-toets**
- Z-toets



t-toets

➤ Voorbeeld

- H_0 : laptops verbruiken gemiddeld 31 Watt

$$H_0 : \mu = 31$$

- H_1 : laptops verbruiken gemiddeld meer of minder dan 31 Watt

$$H_1 : \mu \neq 31$$

➤ Werkwijze

- bepaal significantieniveau α
- doe een steekproef
- stel een aanvaardingsinterval op ($p=1-\alpha$)
- trek een besluit

t-toets

- In ons voorbeeld
 - $\alpha=0,05$
 - 30 laptops, σ onbekend, dus gebruik t-verdeling
 - aanvaardingsinterval (0,95)
 - >>> `t.ppf(0.975,29)` = 2,045
 - $\mu_0 - 2,045 \sigma / \sqrt{n} < \bar{x} < \mu_0 + 2,045 \sigma / \sqrt{n}$
 - $23,53 = 31 - 2,045 \cdot 20 / \text{sqrt}(30)$
 - $38,46 = 31 + 2,045 \cdot 20 / \text{sqrt}(30)$
- We weten dus 95% zeker dat het gemiddelde verbruik tussen deze grenzen zou moeten liggen indien H_0 waar is
- Dus...? (H_0 stelde dat μ gelijk is aan 31 Watt en $\bar{x} = 40$ Watt)

t-toets

Met Python:

```
>>> df_c = pd.read_csv('consumptionLaptops.csv', decimal='.')

>>> x_bar = df_c.consumptionLaptops.mean()
>>> s = df_c.consumptionLaptops.std()
>>> n = len(df_c)
>>> mu0=31

>>> factor = t.ppf((1+0.95)/2, n-1)
>>> interval = (mu0-factor*s/math.sqrt(n), mu0+factor*s/math.sqrt(n))
>>> #OF:

>>> t.interval(confidence=0.95, df=n-1, loc=mu0, scale=s/math.sqrt(n))
```

t-toets

➤ Stel dat H_0 zei dat het gemiddelde verbruik 45 Watt is

➤ Welk besluit trek je nu?

$$\alpha = 0,05$$

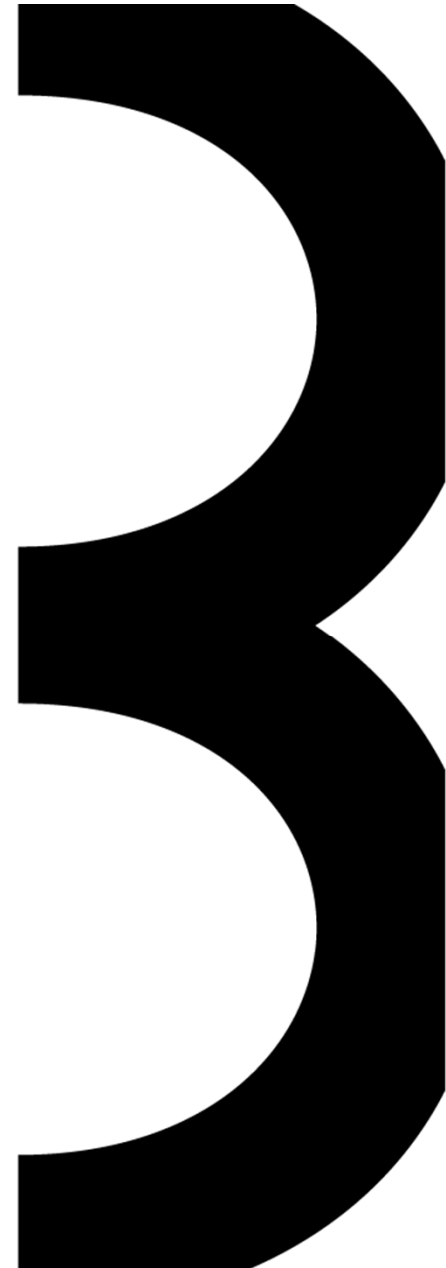
➤ $35,53 = 45 - 2,045 \cdot 20 / \sqrt{30}$

➤ $52,46 = 45 + 2,045 \cdot 20 / \sqrt{30}$

```
>>> t.interval(confidence=0.95, df=30-1, loc=45,  
               scale=20/math.sqrt(30))
```

Toetsen van hypothesen

- Hypothese
- t-toets
- **Z-toets**



Z-toets

- Soms weet je de waarde van σ wel.
 - ⇒ Gebruik dan de normaalverdeling voor het aanvaardingsinterval

OF

- Soms is n heel groot (> 100)

Old habits die hard!
Benaderingen niet meer nodig sinds beschikbaarheid
van statistische software packages op computers



In de Media

4

In de Media

<https://www.demorgen.be/nieuws/grote-meerderheid-canadezen-wil-niet-betalen-voor-veiligheid-prins-harry-en-meghan-markle~bae82a56/>

Grote meerderheid Canadezen wil niet betalen voor veiligheid prins Harry en Meghan Markle



Harry en Meghan. Archiefbeeld. Beeld Photo News

Zo'n 77 procent van de Canadezen vindt dat de belastingbetaler in Canada niet moet betalen voor de veiligheid van de Britse prins Harry en zijn echtgenote Meghan Markle. Dat blijkt uit onderzoek van Nanos op vraag van de televisiezender CTV. Het echtpaar woont sinds ze zich terug terugtrokken uit de koninklijke familie in een buitenwijk van Victoria, in de provincie Brits-Columbia.

IB en BELGA 3 februari 2020, 3:50

De 77 procent wil niet betalen omdat Harry en Meghan niet in Canada zijn als vertegenwoordigers van de Britse koningin. Koningin Elisabeth II is het officiële staatshoofd van Canada. Slechts 19 procent van de Canadezen ziet er geen graten in.

Er zijn nog geen officiële aankondigingen geweest over de beveiliging van het paar en de kosten die daarbij komen kijken. De autoriteiten hebben enkel laten weten dat het onderwerp besproken zal worden.

PRIVACY

Uit de bevraging blijkt ook dat meer dan twee derde van de ondervraagden gelooft dat de privacy van het echtpaar en hun zoon beter zal worden gerespecteerd door de Canadezen dan in Groot-Brittannië. Zo'n 71 procent denkt dat er in hun land minder media-aandacht zal zijn voor Harry en Meghan.

Het echtpaar dreigde eerder al juridische stappen te zetten nadat een fotograaf, die in de struiken verstopt zat op Vancouver Island, foto's had genomen toen de voormalige Amerikaanse actrice op wandel was met haar 8 maanden oude zoon in een draaaszak en met haar twee honden.

35 PROCENT WIL BANDEN MET BRITSE MONARCHIE DOORKNIPPEN

Verder blijkt uit de bevraging dat 35 procent van de ondervraagden de banden met Britse monarchie wil doorknippen. Amper 32 procent is onvoorwaardelijk voorstander van het behoud van de banden met de koninklijke familie en de status van de constitutionele monarchie in hun land. Zo'n 28 procent is voorstander, maar toch eerder onder voorbehoud.

Voor de enquête werden 1.003 Canadezen bevestigd via telefoon of online. Er is een foutenmarge van 3,1 procent.

In de Media

DE
REDACTIE.BE

Extra inspanningen nodig om Vlaamse natuurdoelen te halen



vr 02/12/2016 - 20:52 Frank Segers

Er zijn nog extra inspanningen nodig om de doelstellingen uit het Vlaams Milieubeleidsplan 2011-2015 te halen. Dat blijkt uit een evaluatie van die doelstellingen door het Instituut voor Natuur en Bosonderzoek (INBO).

<http://deredactie.be/cm/vrtnieuws/binnenland/1.2834220>

Vlaanderen streeft naar een duurzaam gebruik van de open ruimte. Het Ruimtelijk Structuurplan Vlaanderen (RSV) voorziet hiervoor onder meer in de afbakening van een natuurlijke structuur. De kern van deze natuurlijke structuur zal bestaan uit 125.000 ha Vlaams Ecologische netwerk (VEN) waarin de functie natuur primeert.

Zeven jaar na het verstrijken van het streefjaar binnen het RSV is ongeveer driekwart van dat Vlaams Ecologisch Netwerk afgebakend, goed voor 92.000 ha.

Wat de afbakening van natuurverwervingsgebieden betreft in datzelfde RSV, is amper zes procent afgebakend. Met een toename van ongeveer 1.100 ha is er de afgelopen jaren weinig vooruitgang geboekt, is te lezen.

Maar er is wel een ander belangrijk doel bereikt. Met een oppervlakte van 79.521 ha effectief natuurbeheer is het plan van 70.000 ha ruim gehaald. Aandachtspunt is wel dat de terreinbeherende verenigingen (zoals bijvoorbeeld Natuurpunt) al jaren minder terreinen verwerven. Ook heeft die stijging niet geleid tot een afdoende bescherming van bedreigde en beschermde soortengroepen, zoals bijvoorbeeld akkervogels.

Bos

Wat de recente toename van bos betreft, spreekt het INBO zich niet uit, aangezien het verschil tussen de laatste Boswijzermeting (2013) en de nulmeting (2010) zich binnen het betrouwbaarheidsinterval bevindt.

Er is wel een belangrijke stijging van de totale oppervlakte toegankelijke bossen, maar onder andere het vaak lange inspraaktraject dat nodig is bij de opmaak van een toegankelijkheidsregeling maakt dat slechts 28 procent van het doel werd bereikt. Wel zijn er voldoende stadsbosprojecten opgestart.

Achttien jaar na de vaststelling van het RSV is slechts 39 procent van de vooropgestelde oppervlakte natuur-, reservaat- en bosgebied en overig groengebied gerealiseerd, maar Vlaams minister Joke Schauvliege (CD&V) wijst op de nieuwe principes voor de ruimtelijke ordening (BRV) die deze week zijn goedgekeurd, en een stevige vergroening van de ruimte inhouden.

Impact klimaatverandering

Het INBO merkt ook op dat in de natuur in Vlaanderen steeds meer aanwijzingen te vinden zijn van de impact van klimaatverandering. "Bij een aantal bomen, waaronder berk, en grassoorten, komt de stuifmeelproductie vroeger op gang", schrijven de wetenschappers. Ook de bladontwikkeling bij eik en beuk vertoont wijzigingen. Het uitlopen van beide soorten verloopt vroeger in warme jaren dan in koude, maar de langetermijneffecten daarvan zijn nog onduidelijk.

Naast verschuivingen in tijd zijn er ook ruimtelijke wijzigingen. Zo breiden zuidelijke en zuidoostelijke soorten zich uit naar het noorden. Dat is onder meer het geval voor libellen.

Omdat invasies van exotische soorten internationaal als één van de grootste bedreigingen voor de biodiversiteit wordt beschouwd, maakte Europa een signaallijst op. In Vlaanderen komen minstens 89 soorten uit die lijst voor, waarvan 41 zich ook echt invasief gedragen.

In de Media

DE
REDACTIE.BE

15 procent Vlamingen ontevreden over hygiëne partner



<http://deredactie.be/cm/vrtnieuws/binnenland/1.1713282>

🕒 ma 26/08/2013 - 13:51 🇧🇪 Belga

Meer dan één op de vier Vlamingen stoort zich vaak aan de lichaamsgeur van anderen. Ook vindt 15 procent dat de persoonlijke hygiëne van de partner beter kan. Nochtans oordeelt 94 procent van de Vlamingen dat het goed is gesteld met zijn hygiëne, wisselt 83 procent elke dag van onderbroek en poetst 40,5 procent de tanden een keer per dag. Dat blijkt uit een onderzoek bij meer dan 3.000 Vlamingen door de Christelijke Mutualiteit (CM).



Gemiddeld spenderen we 28 minuten per dag in de badkamer. Dagelijks brengt bovendien 55 procent van de vrouwen meer dan dertig minuten in de badkamer door, tegen 30 procent van de mannen. Ook neemt 38 procent van de Vlamingen minstens elke dag een douche of een bad.

Meer dan de helft van de respondenten beschouwt baden of douchen niet alleen als iets functioneels, zegt Marjolein Cuvelier van het CM-tijdschrift "çava?". "Vaak is het een moment om tot rust te komen en te ontspannen." Na het vrijen is 31 procent geneigd zich te douchen, 19 procent doet dit na elke vrijpartij.

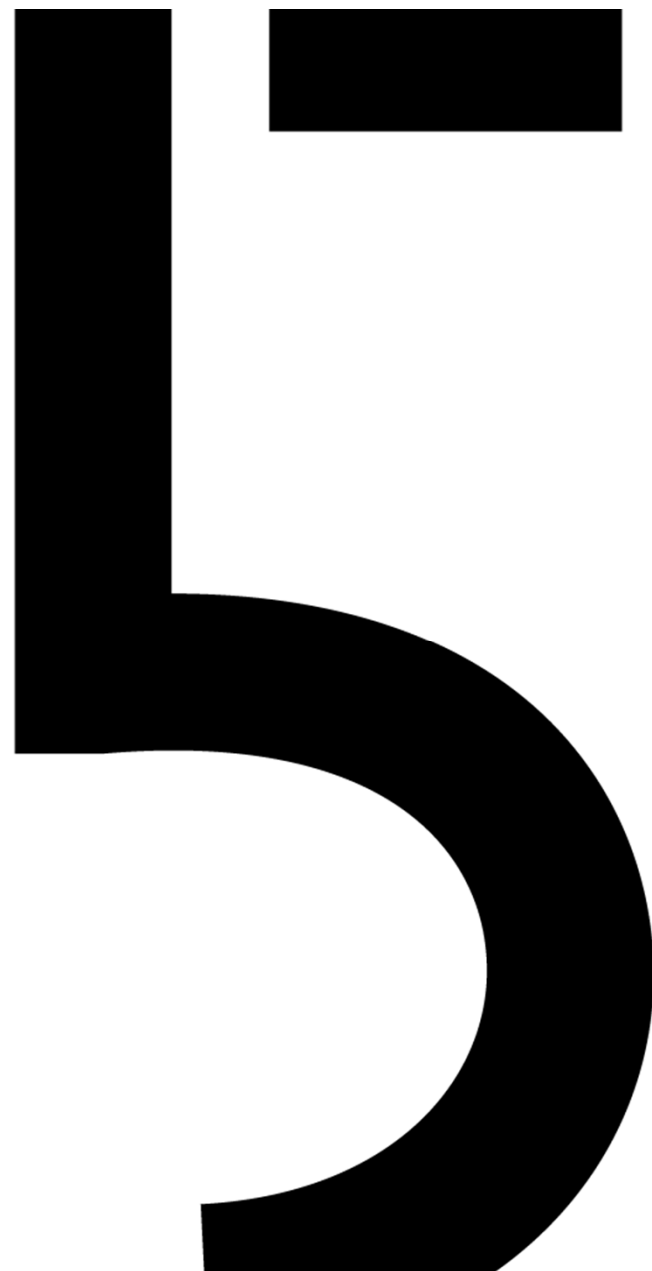
Voorts wast 54 procent de handen na elk toiletbezoek, terwijl 35 procent dit "meestal" doet. Ook gaat één op de drie Vlamingen (35 procent) nooit op de bril van een openbaar toilet zitten. Negentien procent legt eerst een laagje toiletpapier. Acht procent vermijdt zelfs openbare toiletten en wacht tot thuis.

Het onderzoek wijst verder nog uit dat 87 procent van de vrouwen en 3 procent van de mannen de benen scheert. Voor de oksels is dat 23 procent van de mannen, tegenover 91 procent van de vrouwen. 15 procent van de mannen werkt dan weer de wenkbrauwen bij, tegen 43 procent van de vrouwen.

De foutenmarge van het onderzoek bedraagt 1,65 procent, bij een betrouwbaarheidsinterval van 95 procent.



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

studentq

```
>>> import pandas as pd
```

```
>>> studentq = pd.read_csv('vragenlijst 21-22.csv', delimiter=';',  
decimal='.')
```

Vragenlijst



1.a Voeg een kolom toe aan het dataframe en plaats daarin de gestalte van een persoon uitgedrukt in zijn schoenmaat (maw lengte gedeeld door schoenmaat)

1.b Bepaal het gemiddelde en de standaardafwijking

Vragenlijst



2. Geef het betrouwbaarheidsinterval ($\alpha = 5\%$) voor de gemiddelde verhouding gestalte-schoenmaat

Vragenlijst



3. Iemand beweert dat de verhouding gestalte-schoenmaat van een mens gelijk is aan 4,2 met als standaardafwijking 0,05. Kan je op basis van de gegevens uit de vragenlijst dit ($\alpha = 5\%$) bijtreden? En indien $\alpha = 2.5\%$?

Oefeningen



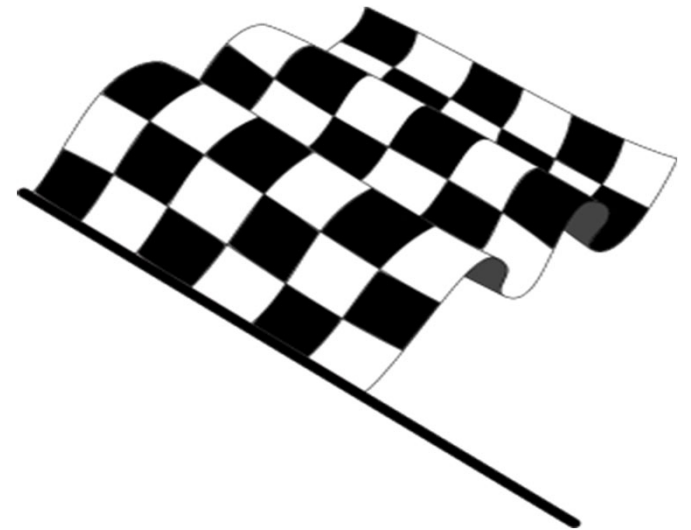
canvas



Oefening

- vraag 3
- vraag 4
- Vraag 5
- vraag 6

Ook: extra cursusmateriaal



KdG Karel de Grote
Hogeschool