Data Science 2
Toetsen – deel 2

Wim De Keyser
Geert De Paepe
Jan Van Overveld

KdG Karel de Grote Hogeschool

Quote van de week

"As Confucius might have said, if the difference isn't different enough to make a difference, what's the difference?"

Victor Chew (1923-?)

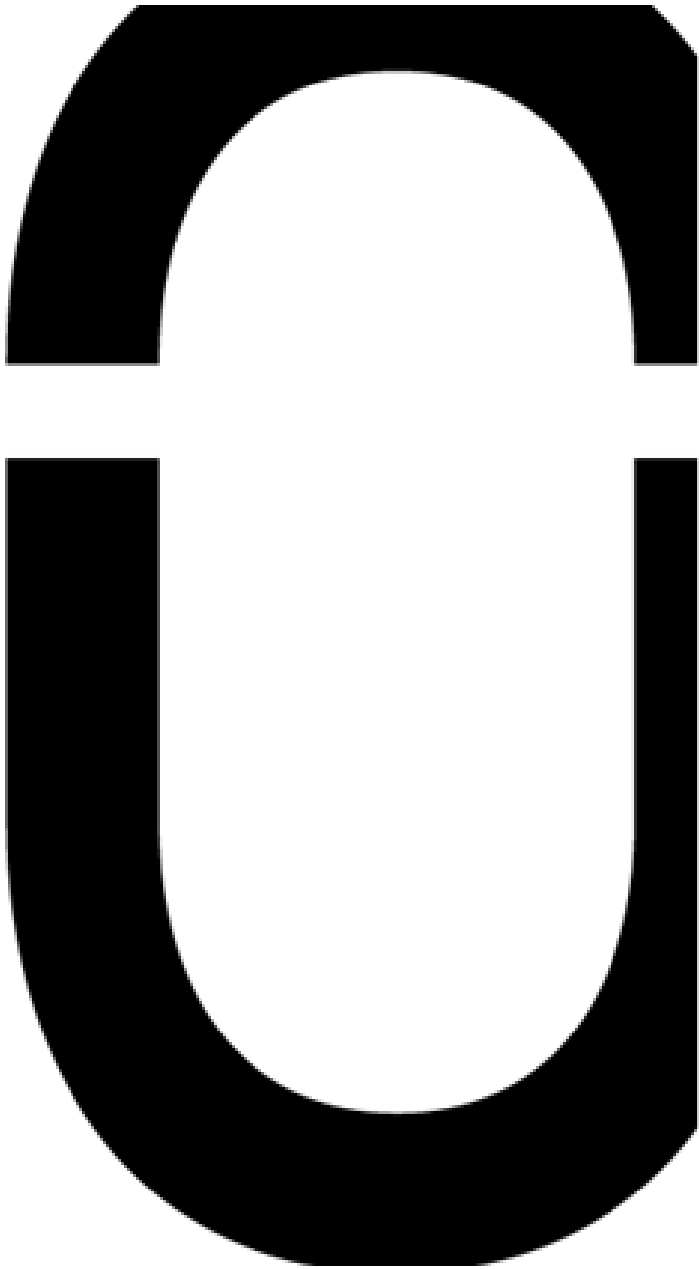
Agenda

1. Grenswaarde α bepalen ⇨ p-waarde
2. Hypothese toetsen gemiddelde
3. Andere toetsen
4. Chi-kwadraat toets
5. In de media

Cursus/ Boek *Tekst op Canvas*

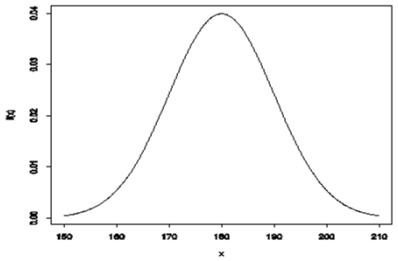
canvas Oefeningen

Herhaling

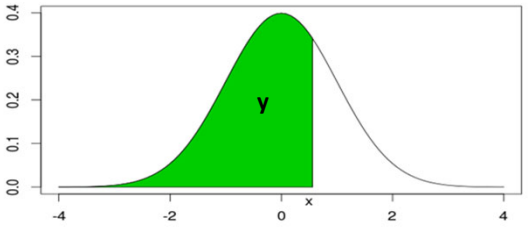


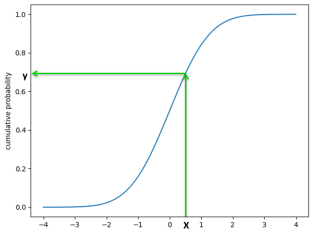
De normale verdeling

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

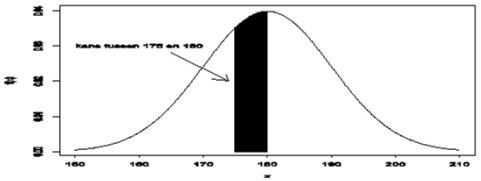


```
>>> from scipy.stats import norm
>>> norm.cdf(x,loc=μ,scale= σ)
```

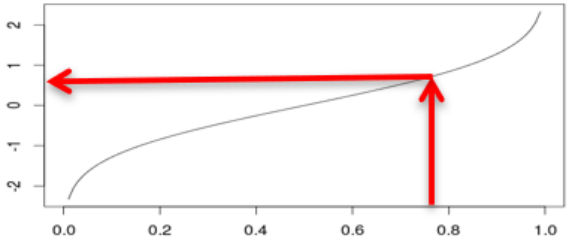




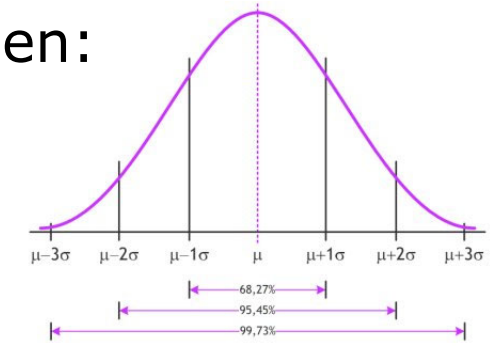

```
>>> norm.cdf(y,loc=μ,scale= σ) –
    norm.cdf(x,loc=μ,scale= σ)
```



```
>>> norm.ppf(x)
```



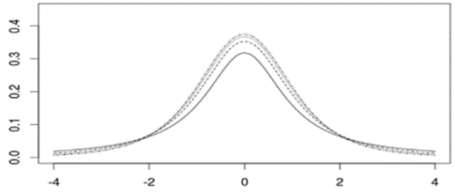
- Eigenschappen:



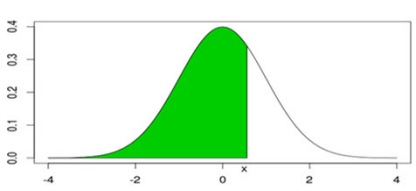
De student verdeling

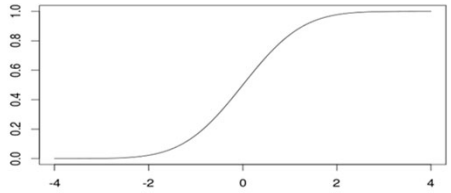
- Kansen symmetrisch verdeeld
- Continue verdeling
- Parameter: df

```
>>> from scipy.stats import t
>>> t.pdf(x, df)
```

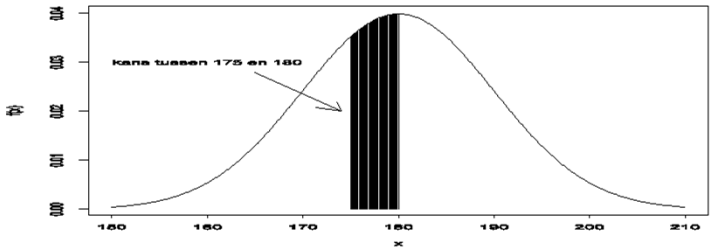


```
>>> t.cdf(x, df)
```

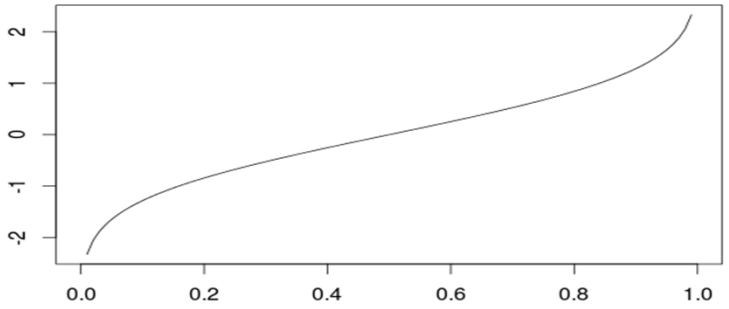




```
>>> t.cdf(y, df) – t.cdf(x,df)
```

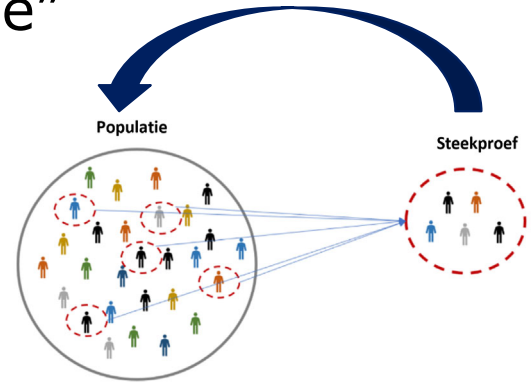



```
>>> t.ppf(x, df)
```



Betrouwbaarheidsinterval

"We willen op basis van een steekproef iets zeggen over de populatie"

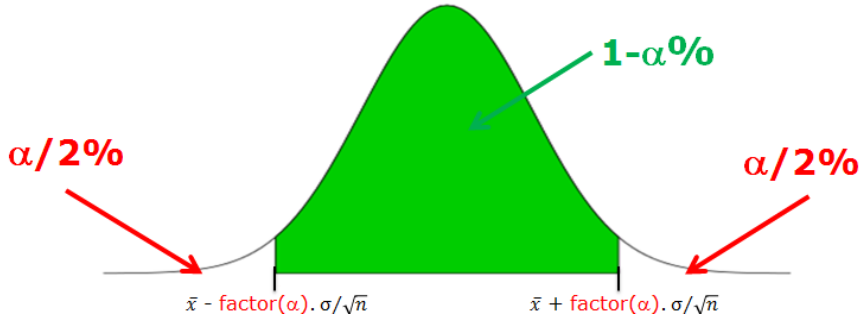


Uitspraak niet 100% wel 1-α = P

Steekproef: n,$\bar{x}$,s ⇨ Populatie: μ?

- Centrale limietstelling: $\bar{x}$ is normaal verdeeld met parameters μ en $\sigma/\sqrt{n}$
- $\mu - \text{factor}(\alpha)\,\sigma/\sqrt{n} < \bar{x} < \mu + \text{factor}(\alpha)\,\sigma/\sqrt{n}$
- $\bar{x} - \text{factor}(\alpha)\,\sigma/\sqrt{n} < \mu < \bar{x} + \text{factor}(\alpha)\,\sigma/\sqrt{n}$

Bepaal de factor(α):



Opmerking:
- indien **σ gekend**, gebruik de **normale verdeling**.
- indien **σ onbekend**, neem **s ipv σ** en gebruik de **student verdeling**

```
>>> t.ppf(1-α/2,n-1)
>>> t.ppf((1+P)/2,n-1)
```

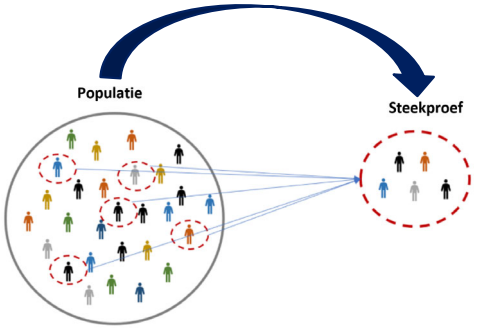
Bepaal het betrouwbaarheidsinterval:

```
>>> interval = (x_bar-factor*s/math.sqrt(n),
                x_bar+factor*s/math.sqrt(n))

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

Aanvaardingssinterval

"Kunnen we een bewering over de populatie weerleggen aan de hand van een steekproef?"



Bewering als hypothese formuleren:

$H_0 : \mu = \mu_0$ en $H_1 : \mu \neq \mu_0$

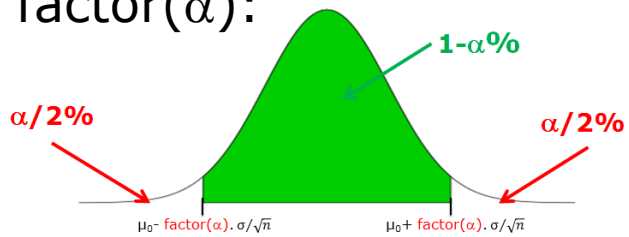
Besluit niet 100% wel $1-\alpha = P$

- Type I-fout: H_0 verwerpen terwijl toch waar
- Type II-fout: H_0 aanvaarden terwijl niet waar

Steekproef: $n, \bar{x}, s \Rightarrow H_0$ weerleggen?

- Indien H_0 waar, dan is $\bar{x}$ normaal verdeeld met parameters μ_0 en $\sigma/\sqrt{n}$
- $\mu_0 - \text{factor}(\alpha)\,\sigma/\sqrt{n} < \bar{x} < \mu_0 + \text{factor}(\alpha)\,\sigma/\sqrt{n}$

Bepaal de factor(α):



Opmerkingen:
- indien **σ gekend**, gebruik de **normale verdeling**.
- indien **σ onbekend**, neem **s ipv σ** en gebruik de **student verdeling**

```
>>> t.ppf(1-α/2,n-1)
>>> t.ppf((1+P)/2,n-1)
```

Bepaal het aanvaardigsinterval:

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

Korte herhaling - Voorbeeld

- Voorbeeld
  - Iemand beweert dat de gemiddelde scherm-grootte van alle verkochte televisies 40 inch is
  - We kijken naar de volgende 50 verkochte televisies. We meten een gemiddelde van 43 inch en een standaardafwijking van 10 inch
  - We willen 95% zeker zijn van het resultaat

$\mu_0 = 40$

$n = 50$

$\bar{x} = 43$

$s = 10$

$P = 0,95$

- wat is
  - H_0, H_1?
  - Wat is het significantieniveau? Wat is dit?
  - Welke stappen doorloop je?
  - Welke verdeling gebruik je?
  - Wat besluit je?

$H_0 : \mu = 40$

$H_1 : \mu \neq 40$

$\alpha = 0,05$

t-verdeling

Korte herhaling - Voorbeeld

- Aanvaardingsinterval:

  μ_0 - factor(α)*s/sqrt(n) en μ_0 + factor(α)*s/sqrt(n)

  P = 0.95 = 1-α en α = 0.05

  => t.ppf((1+P)/2,n-1) = t.ppf(0.975,49) = 2.010

  40 – 2.010*10/sqrt(50) en 40 + 2.010*10/sqrt(50)

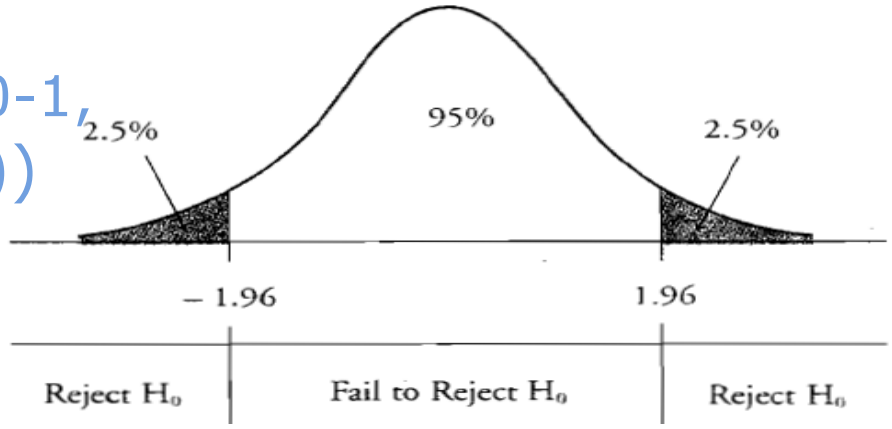
  40 – 2.843 en 40 + 2.843

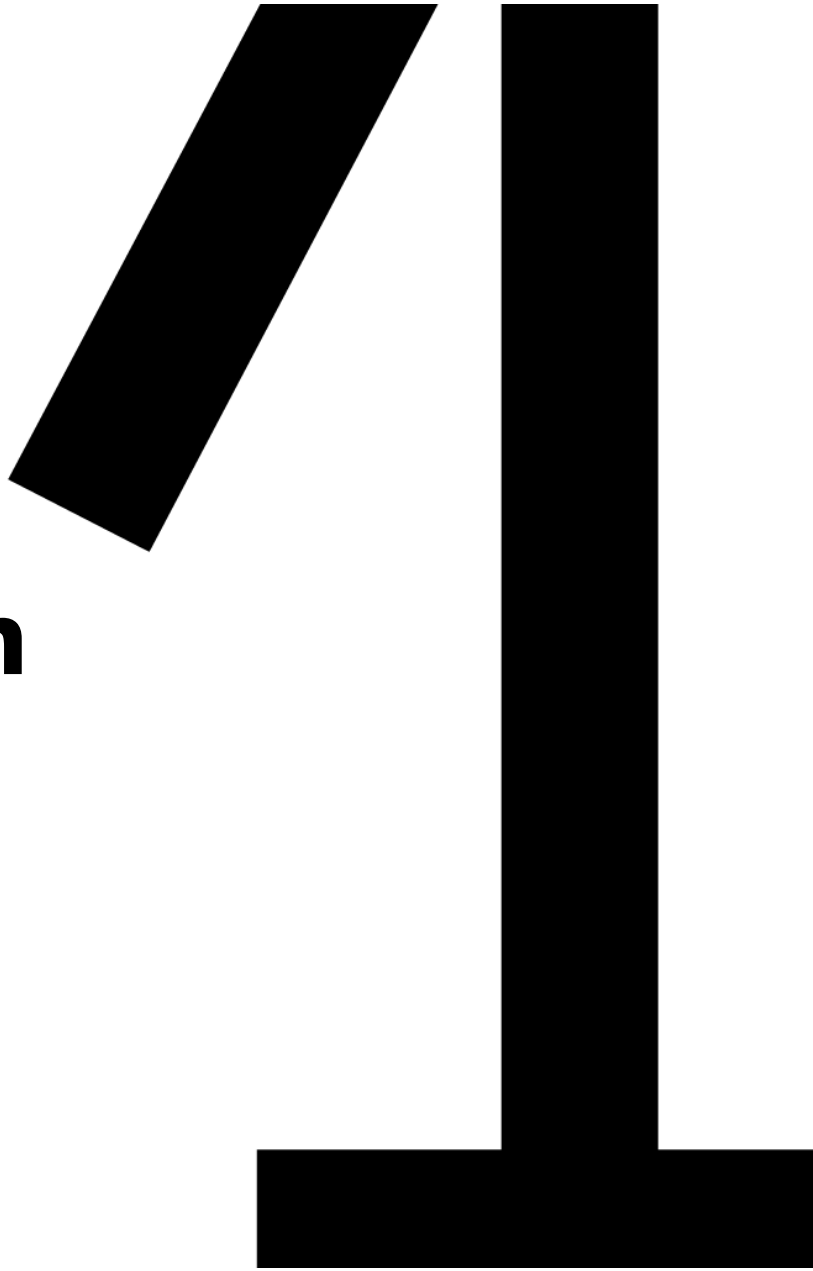
  37.157 en 42.843

  >>> t.interval(confidence=0.95, df=50-1,
  loc=40, scale=10/math.sqrt(50))

  43 = $\bar{x} \notin$ [37.157,42.843]

  H_0 kan worden verworpen.





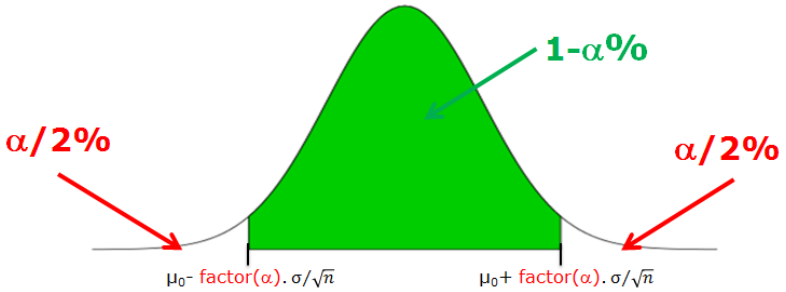
Grenswaarde α bepalen
⇨ p-waarde

Grenswaarde α bepalen

$H_0 : \mu = 40$ $\quad n = 50$

$H_1 : \mu \neq 40$ $\quad \bar{x} = 43$

$\quad s = 10$



α	P=(1-α)	$t_{\alpha/2}$ = factor(α)	Aanvaardingsinterval	Besluit
0.10	0.90	1.676551	[37.629,42.371]	H_0 verwerpen
0.05	0.95	2.009575	[37.157,42.843]	H_0 verwerpen
0.01	0.99	2.679952	[36.210,43.790]	kunnen H_0 niet verwerpen

Vaststelling: **α kleiner $\Leftrightarrow$ aanvaardingsinterval groter**

P groter $\Leftrightarrow$ aanvaardingsinterval groter

Voor welke α valt $\bar{x}$ nog net binnen het aanvaardings-interval?

$\bar{x} \in [\mu_0 - t_{\alpha/2} * s/\text{sqrt}(n) \; , \; \mu_0 + t_{\alpha/2} * s/\text{sqrt}(n)]$

$(\bar{x} - \mu_0)/ (s/\text{sqrt}(n)) \in [-t_{\alpha/2} \; , \; +t_{\alpha/2}]$

Opmerking: $t_{\alpha/2}$ = factor(α)

Grenswaarde α bepalen

$\bar{x} \in [\mu_0 - t_{\alpha/2} * s/\sqrt{n} \; , \; \mu_0 + t_{\alpha/2} * s/\sqrt{n}]$

$(\bar{x} - \mu_0) / (s/\sqrt{n}) \in [-t_{\alpha/2} \; , \; +t_{\alpha/2}]$

$(43 - 40) / (10/\sqrt{50}) \in [-t_{\alpha/2} \; , \; +t_{\alpha/2}]$

$2.121 \in [-t_{\alpha/2} \; , \; +t_{\alpha/2}]$

>>> t.ppf(?,49) = 2.121

>>> t.cdf(2.121,49) = ?

= 0.9805

$(1+p)/2 = 0.9805$

OF

$1 - \alpha/2 = 0.9805$

$\Rightarrow 2 - \alpha = 2 * 0.9805$

$\Rightarrow \alpha = 2 - 1.961$

$\Rightarrow \alpha = 0.039$

α	p	$t_{\alpha/2}$	Aanvaardings-interval	Besluit
0.039	0.961	2.121	[37.0,43.0]	kunnen H_0 niet verwerpen

Grenswaarde α bepalen

in Python, voer een 'one sample t-test' uit:

```
>>> from scipy.stats import ttest_1samp
>>> screens = pd.read_csv('screens.csv',delimiter=';', decimal='.')
>>> mu=40
>>> ttest_1samp(screens["New_size"],mu)
```

Output:

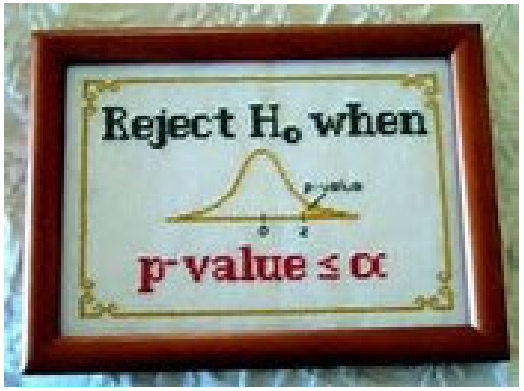
```
Ttest_1sampResult(statistic=2.1213203435596424, pvalue=0.0389780552443494 1)
```

Opmerking: Voor welke α valt $\bar{x}$ net **<u>buiten</u>** het aanvaardingsinterval?

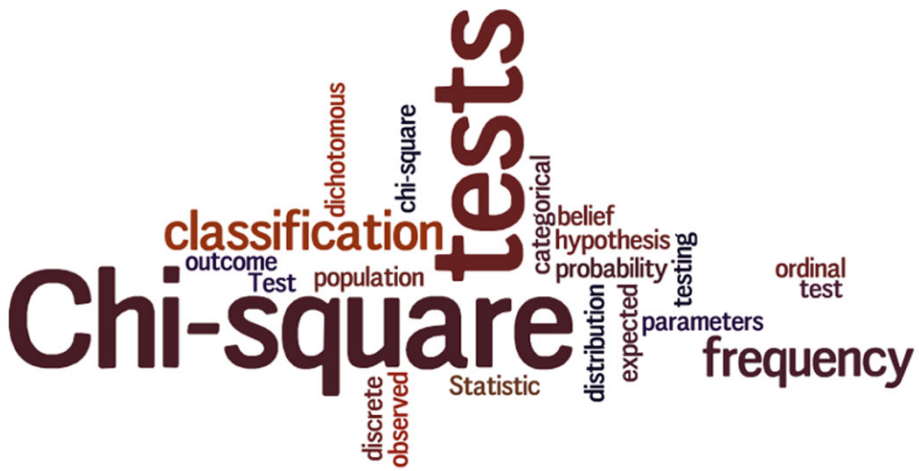
p-waarde

- *p*-waarde
  - Stel dat H_0 waar is, wat is dan de kans om de gevonden waarde of iets extremer (groter of kleiner) te vinden?
  - *p*-waarde moet kleiner zijn dan α om H_0 te kunnen verwerpen

```
Ttest_1sampResult(statistic=2.1213203435596424, pvalue=0.0389780552443494 1)
```



α	P=(1-α)	$t_{\alpha/2}$ = factor(α)	Aanvaardingsinterval	Besluit
0.10	0.90	1.676551	[37.629,42.371]	H_0 verwerpen
0.05	0.95	2.009575	[37.157,42.843]	H_0 verwerpen
0.01	0.99	2.679952	[36.210,43.790]	kunnen H_0 niet verwerpen



Chi-kwadraat toets

4

Chi-kwadraat toets

- Stel dat je wil nagaan of een variabele een bepaalde verdeling heeft
- Voorbeeld:
  - Eigenaar van een winkel verkoopt computers, films, spelletjes en televisies
  - Hij denkt dat de verkoop als volgt verdeeld is
    - computers: 30%
    - films: 10%
    - spelletjes: 30%
    - televisies: 30%
  - Vraag: klopt dit?

Chi-kwadraat toets

- We doen een steekproef (30 klanten):
  - computers: 15
  - films: 3
  - spelletjes: 7
  - televisies: 5
- We hadden verwacht:
  - computers: 9
  - films: 3
  - spelletjes: 9
  - televisies: 9
- Vraag: hoe “ver” liggen deze van elkaar?

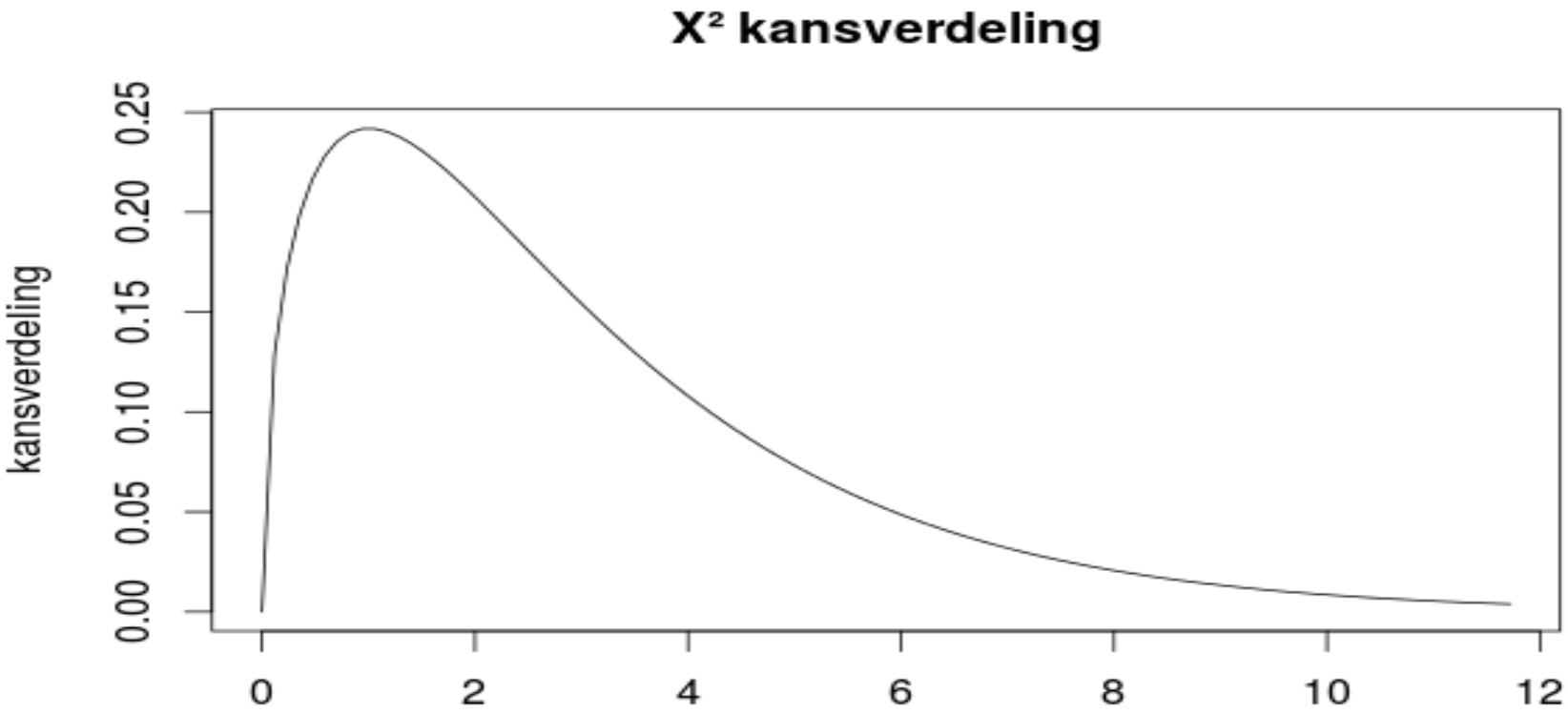
Chi-kwadraat toets

- Oplossing: bereken χ^2:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- Dit is een soort afstand tussen de geobserveerde frequenties en de verwachtte frequenties
- Frequenties zijn altijd absolute frequenties!
- Als H_0 waar is, dan is er een grote kans om een kleine χ^2 te vinden
- als χ^2 groot is, dan is H_0 dus waarschijnlijk fout
- in ons geval: $\chi^2 = \frac{(15-9)^2}{9} + \frac{(3-3)^2}{3} + \frac{(7-9)^2}{9} + \frac{(5-9)^2}{9} = 6.22$

Chi-kwadraat toets



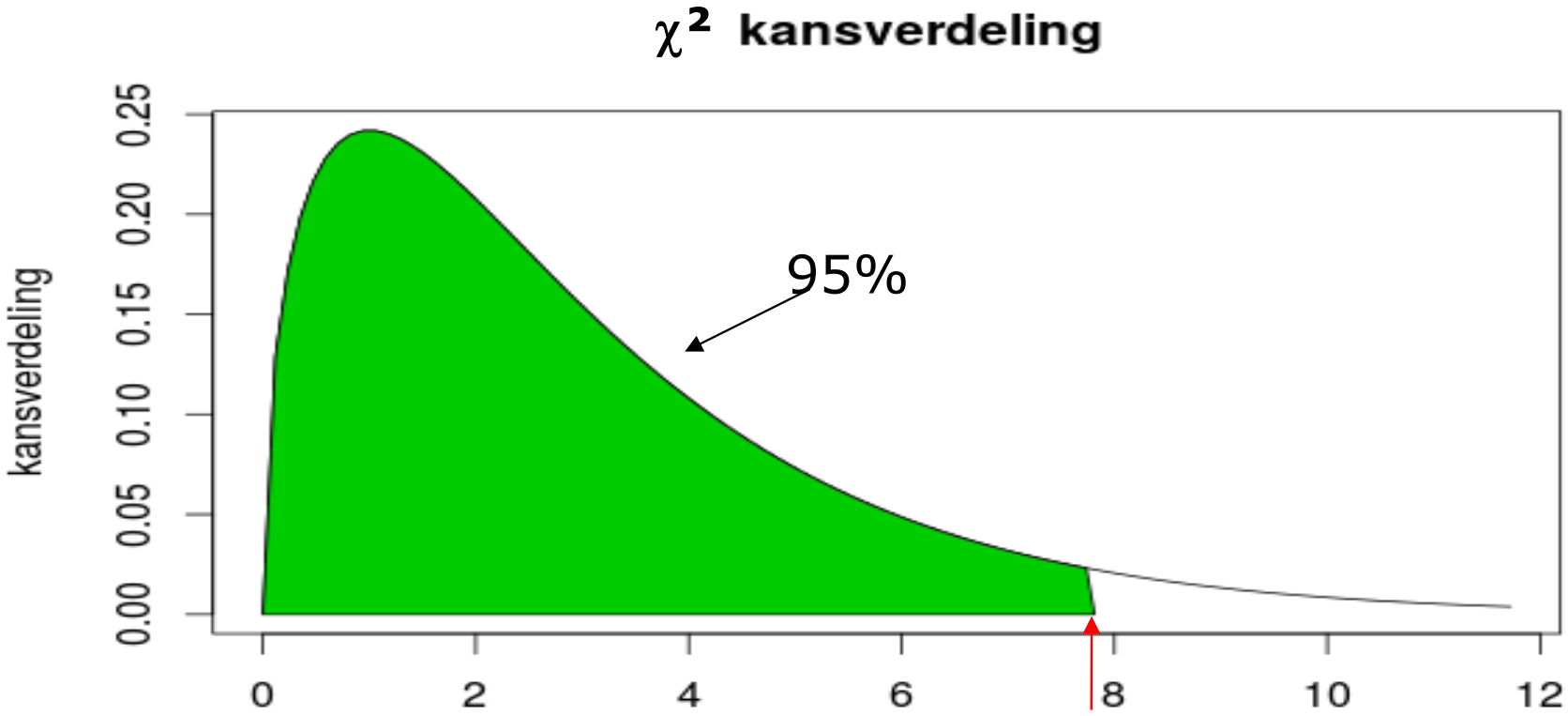

```
>>> from scipy.stats import chi2
>>> chi2.pdf (x, df=m-1)
```

Chi-kwadraat toets

- De chi-kwadraat verdeling wordt bepaald door een parameter m-1
- m is het **aantal waarden** die geteld werden (4 in ons geval)
- We kunnen nu weer een 95% aanvaardingsinterval opstellen (m.a.w. $\alpha = 0.05$)
  - begint altijd van 0
  - gaat tot een grenswaarde:

```
>>> from scipy.stats import chi2
>>> chi2.ppf ((1- α), df=m-1)
```

Chi-kwadraat toets



7.815 = chi2.ppf (0.95, df=4-1)
chi2.cdf (7.815, df=4-1) = 0.95

Chi-kwadraat toets

- chi2.ppf (0.95, df=4-1) = 7,8
- Er is dus 95% kans om een χ^2 te vinden die kleiner is dan 7,8 (chi2.cdf (7.815, df=4-1) = 0,95)
- Wij vonden een χ^2 = 6,22
- Dus?

Chi-kwadraat toets

- We kunnen ook *p*-waarde berekenen
  = Wat is de kans dat je χ^2 of hoger zou uitkomen als H_0 waar is?
  = oppervlakte van χ^2 tot oneindig
  = 1-chi2.cdf(χ^2, m-1)
- in ons geval:
  1-chi2.cdf(6.22, 4-1) = 0,1013
- dus 10,13% kans dat je 6,22 of hoger uitkomt als H_0 waar is
- dit is groter dan 5%, dus kunnen we H_0 niet verwerpen

Chi-kwadraat toets

- Je kan de test ook als volgt doen:

```
>>> from scipy.stats import chisquare
>>> measured_values = [15,3,7,5]
>>> expected_values = [9,3,9,9]
>>> chisquare(measured_values , expected_values)
```

- Output:

```
Power_divergenceResult(statistic=6.222222222222222, pvalue=0.10128520246221875)
```

- Kijk naar de p-waarde om een besluit te trekken

In de media

In de Media

De Standaard – 15 februari 2019

FACTCHECK

‘Mondiaal neemt veertig procent van de insectensoorten af’

AUSTRALISCHE ONDERZOEKERS DEZE WEEK IN HET TIJDSCHRIFT ‘BIOLOGICAL CONSERVATION’

- ● ONGEFUNDEERD
- ○ NIET WAAR
- ○ EERDER NIET WAAR
- ○ EERDER WEL WAAR
- ○ WAAR

GEMMA VENHUIZEN

Berichten over kelderende insectenaantallen halen vaak het nieuws: zo verdween in Duitsland de afgelopen dertig jaar ruim driekwart van de vliegende insecten. Ook deze week was er insectennieuws naar aanleiding van Australisch onderzoek dat meer dan 40 procent van de insectensoorten afneemt in aantal *(DS 12 februari)*.

Twee Australische biologen, Francisco Sánches-Bayo en Kris Wyckhuys, brachten de mondiale achteruitgang in kaart. Daarvoor analyseerden ze 73 artikelen over insectenbiodiversiteit van de voorbije 40 jaar. In het tijdschrift *Biological Conservation* schrijven ze over een ‘wereldwijde afname’. Ze noteren ‘dramatische afnamesnelheden’ die in enkele decennia kunnen leiden tot ‘het uitsterven van 40 procent van de insectensoorten wereldwijd’. **Belangrijkste oorzaak, volgens de onderzoekers: grootschalige landbouw.**

En, klopt het? Het artikel somt meerdere waarheden op: veel soorten gaan inderdaad in aantal en biomassa achteruit, en onder meer vliesvleugeligen, libellen en eendagsvliegen zijn daarvan de dupe. **‘De afname is wijdverbreid, misschien wel sterker voor insecten dan voor vogels en zoogdieren, en dat is zeer verontrustend’**, zegt hoogleraar plant ecology Hans de Kroon van de Radboud Universiteit (betrokken bij een Duitse studie naar vliegende insecten). Hij noemt het artikel ‘een eerste poging om de achteruitgang op mondiale schaal te duiden’.

Niet onbevooroordeeld

Maar die poging rammelt statistisch, zegt hoogleraar statistiek Casper Albers van de Rijksuniversiteit Groningen. ‘De onderzoekers hebben gekeken naar 73 artikelen, waarvan 60 uit Europa en de VS. Voor de rest van de wereld – en die is nogal groot – zijn maar 13 studies bekeken – te weinig om conclusies te trekken voor Afrika en Azië.’

Ook hebben de auteurs specifiek gezocht naar artikelen die de woorden decline/declining (afname/afnemend) bevatten. In die zin zijn ze niet onbevooroordeeld. Bovendien klopt de toetsingsprocedure niet die gebruikt is om de afnamepercentages te bepalen, zegt Albers. Daarin geeft een zogeheten p-waarde aan of verschillen tussen groepen zodanig groot zijn dat ze betekenisvol zijn.

Albers: ‘Heel kort door de bocht: als p kleiner is dan 0,05, is het significant, anders niet.’ Met niet-significante verschillen zouden de auteurs kunnen aantonen dat de achteruitgang overal op aarde op dezelfde manier verloopt. In het artikel wordt onder meer beweerd dat er geen significant verschil is tussen de afnamepercentages in het Verenigd Koninkrijk, Noord-Amerika en Europa: de p-waarde is daar 0,21. Maar bij narekening kwam Albers op 0,035 uit. Wél significant dus. Ook op enkele andere plekken blijken de p-waardes niet te kloppen.

‘Er zijn te weinig studies die de situatie in Afrika en Azië onderzoeken om wereldwijde conclusies te kunnen trekken’

CASPER ALBERS
Hoogleraar statistiek
Rijksuniversiteit Groningen

‘De conclusies kunnen niet zo getrokken worden, want de gerapporteerde statistieken zijn aantoonbaar fout.’ Daardoor is evenmin te zeggen wat de belangrijkste oorzaak van de afname is, aldus Albers. ‘Als de statistiek goed zou zijn, zou het wellicht mogelijk zijn om gecultiveerde landbouw als schuldige aan te wijzen in Europa en de VS, maar het is niet mogelijk dit te generaliseren naar de rest van de wereld.’

Conclusie: insectenaantallen in Europa en de VS nemen inderdaad af, maar er is **te weinig informatie om te spreken van een wereldwijde trend**. Ook is het afnamepercentage onduidelijk. We beoordelen deze uitspraak als **ongefundeerd**.

In de Media

Metro – 9 december 2015

Mannen zetten Ikea-meubel sneller in elkaar

OSLO Wie zet er het snelst een Ikea-meubel in elkaar: een man of een vrouw? Dat is de vraag die psychologen van de Arctic Univerisity of Norway wilden beantwoorden. Het resultaat? Mannen handelen die klus sneller af dan het andere geslacht. Voor het experiment moesten veertig vrouwen tegen veertig mannen strijden om zo snel mogelijk een keukentrolley van Ikea in elkaar te zetten, mét handleiding. De vrouwen deden er 23,65 minuten over, maar de heren haalden met 22,48 minuten een betere tijd. ■

Wees kritisch: Hier wordt enkel het ene gemeten gemiddelde vergeleken (<, = of >) met het andere gemeten gemiddeld!

⇨Het gaat om twee steekproeven

⇨Een tweezijdige vergelijkingstoets ($H_0 : \mu_1 = \mu_2$) had hier moeten plaats vinden: is het verschil (nl. 1,17 = 23,65 - 22,48) wel significant?

⇨Conclusie pas mogelijk na het vastleggen van een significantieniveau en het bepalen van het aanvaardingsinterval

Opmerking: Daar de standaardafwijkingen van de twee steekproeven niet in het artikel worden vermeld, kunnen we de vergelijkingstoets **niet** uitvoeren.

In de Media

Vrouwen excelleren in andere domeinen dan mannen, en dat is logisch

Obstakels wegwerken voor de emancipatie van meisjes en vrouwen is een uitstekend idee. Maar overal streven naar een vooraf vastgelegde sekseratio is dat niet, schrijft Griet Vandermassen.

Griet Vandermassen
Filosofe en auteur

Het thema van Internationale Vrouwendag is dit jaar #EmbraceEquity. Die -slogan laat zich moeilijk vertalen. Het concept van *equity* werd ontwikkeld binnen de filosofie van sociale rechtvaardigheid en is gekoppeld aan namen als John Rawls, Martha Nussbaum en Amartya Sen. Het omvat een andere benadering van sociale rechtvaardigheid dan het concept van gelijke kansen of *equality*. Bij dat laatste geef je alle mensen hetzelfde. *Equity* erkent dat niet iedereen dezelfde startpositie heeft in het leven. Het ijvert voor extra kansen voor wie dat nodig heeft, zodat iedereen de gelegenheid krijgt om een kwaliteitsvol bestaan uit te bouwen.

https://www.standaard.be/cnt/dmf20230307_98131075
8 maart 2023

Variability hypothesis

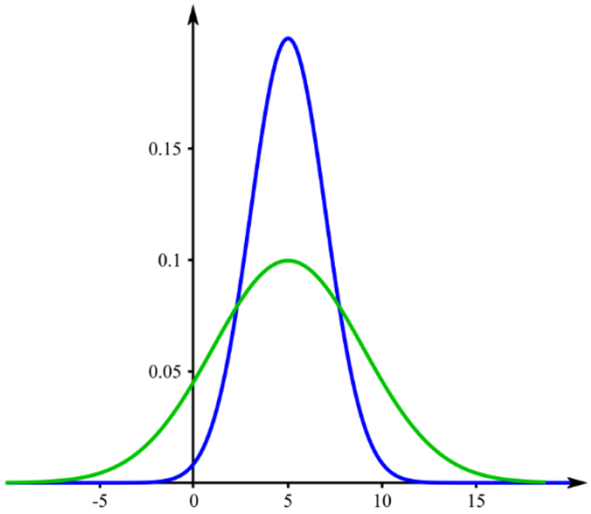
Article Talk

From Wikipedia, the free encyclopedia

The **variability hypothesis**, also known as the **greater male variability hypothesis**, is the hypothesis that males generally display greater variability in traits than females do.

It has often been discussed in relation to human cognitive ability, where some studies appear to show that males are more likely than females to have either very high or very low IQ test scores. In this context, there is controversy over whether such sex-based differences in the variability of intelligence exist, and if so, whether they are caused by genetic differences, environmental conditioning, or a mixture of both.

Sex-differences in variability have been observed in many abilities and traits — including physical, psychological and genetic ones — across a wide range of sexually dimorphic species.



https://en.wikipedia.org/wiki/Variability_hypothesis

In de Media

De Morgen – 17 maart 2016

Schok in wiskunde: priemgetallen volgen patroon

STEFAN GROMMEN

Priemgetallen, de bouwstenen van de wiskunde, volgen wel degelijk een patroon. Die ontdekking doet de wiskundige wereld op haar grondvesten daveren.

Priemgetallen zijn getallen die groter zijn dan 1 en enkel deelbaar door zichzelf en 1. Ze worden beschouwd als de bouwstenen van de wiskunde omdat alle getallen ofwel een priemgetal zijn, ofwel het product van meerdere priemgetallen (bijvoorbeeld 23244 = 2 x 2 x 3 x 13 x 149).

Nog altijd proberen wetenschappers alle eigenschappen van priemgetallen in kaart te brengen. Zo ligt het wel degelijk vast of een getal al dan niet een priemgetal is – ze zijn dus eigenlijk sowieso niet willekeurig – maar wiskundigen weten nog altijd niet hoe ze moeten voorspellen wanneer er een zal voorkomen. De consensus is net daarom om priemgetallen te behandelen 'alsof ze willekeurig voorkomen'.

Precies die consensus wordt nu onderuit gehaald door een nieuwe ontdekking van onderzoekers Kannan Soundararajan en Robert Lemke Oliver van de

'Dit is krankzinnig. We bestuderen priemgetallen al lang, en niemand heeft dit ooit opgemerkt'

ANDREW GRANVILLE
UNIVERSITY COLLEGE LONDON

universiteit van Stanford. Want in het laatste cijfer van priemgetallen zit wel degelijk een soort van patroon.

Alle priemgetallen, 2 en 5 uitgezonderd, moeten eindigen op 1, 3, 7 of 9 want anders zijn ze ook deelbaar door 2 of 5. In theorie, als ze willekeurig voorkomen, zou elk van die vier cijfers 25 procent kans moeten hebben om het priemgetal af te sluiten.

Maar nadat een computerprogramma door de eerste 400 miljard priemgetallen grasduinde, bleek dat priemgetallen er niet van houden om hetzelfde laatste cijfer te hebben als hun voorganger. Veel vaker komt het voor dat een priemgetal eindigt op een ander cijfer dan het priemgetal ervoor. Alsof ze 'voorkeuren' hebben dus, of, zoals het zo beeldend klinkt in de Angelsaksische pers, alsof "we een grote priemgetallensamenzwering hebben ontmaskerd".

Zo is de kans dat een priemgetal eindigend op 1 gevolgd wordt door nog een priemgetal dat eindigt op 1 maar 18,5 procent (in plaats van de verwachte 25 procent). Veel meer kans (30 procent) is er dat een 3 of 7 volgen op een 1. Een 9 volgend op een 1 heeft 22 procent kans. Ook bij andere combinaties noteerde de computer vergelijkbare patronen.

"We bestuderen priemgetallen al lang en niemand heeft dit ooit al opgemerkt", stelt wiskundig theoreticus Andrew Granville van de universiteit van Montreal en de University College London aan *Quanta Magazine*. "Dit is krankzinnig." Sommige wetenschappers geloofden het nieuws zelfs niet tot ze de resultaten van de analyse zagen. "Je zou je nu kunnen afvragen wat we nog allemaal gemist hebben over priemgetallen."

Overigens: praktische implicaties – bijvoorbeeld op de manier waarop we priemgetallen in cryptografie gebruiken – heeft deze ontdekking niet.

Welke toets kan je hier op toepassen?
Wat is je besluit?

Chi-kwadraat toets

- H_0 : na een priemgetal eindigend op 1 volgt willekeurig een priemgetal eindigend op 1,3,7 of 9
  H_1: na een priemgetal eindigend op 1 volgt niet willekeurig een priemgetal eindigend op 1,3,7 of 9
- Gegevens:

Periode	Verwacht (f_e)	Observatie (f_o)
eindigend op 1	4.10^{11} /4	4.10^{11} . 0,185
eindigend op 3	4.10^{11} /4	4.10^{11} . 0,2975
eindigend op 7	4.10^{11} /4	4.10^{11} . 0,2975
eindigend op 9	4.10^{11} /4	4.10^{11} . 0,22

- Met Python:

```
>>> measured_values = [74e9,119e9, 119e9, 88e9]
>>> expected_values = [100e9,100e9,100e9,100e9]
>>> chisquare(measured_values , expected_values)
```

- Besluit (α = 5%)?

Vragenlijst

6

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

1. Herneem de oefening van vorige keer:

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)

b. 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 bijtreden? En dit vanaf welk significantieniveau (α)?

Vragenlijst

2.a Bepaal het 99% betrouwbaarheidsinterval voor de afstand tot KdG (verwijder eerst de uitschieters!).

2.b Iemand beweert dat de afstand tot KdG gemiddeld 19 km bedraagt. Vanaf welk significantieniveau (α) kan je de bewering niet weerleggen?

Vragenlijst

3. In de onderstaande tabel staat de verdeling van de Belgische bevolking over de verschillende bloedgroepen (Bron: *nl.wikipedia.org/wiki/Bloedgroep*). Kan je dit op basis van de steekproef weerleggen? Neem $\alpha = 0{,}05$.

O+	A+	B+	AB+	O-	A-	B-	AB-
38%	34%	8,5%	4,1%	7%	6%	1,5%	0,8%

O	A	B	AB
45%	40%	10%	5%

Oefeningen

- vraag 1
- vraag 2
- vraag 7

KdG Karel de Grote Hogeschool

Vragenlijst

Oplossingen