

Inferentiële statistiek

Peter Kop en Marianne van Dijke

Utrecht 2 juni



Universiteit
Leiden

STATISTIEK
— IN PERSPECTIEF —



Universiteit
Utrecht

Deze workshop

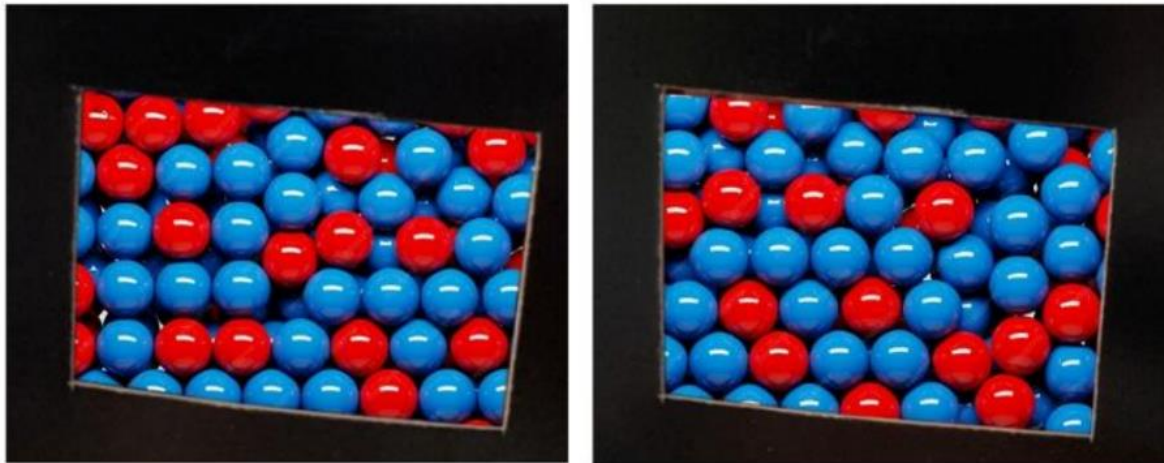
- Schets mogelijke leerlijn inferentiële statistiek
- Didactische route: van informele activiteiten en simulaties naar formele procedures
- Aandacht voor empirische cyclus, samenhang profielvakken en onderzoeksopdrachten
- Onderzoek naar verschil tussen groepen (populaties) op basis van steekproeven: m.b.v. data exploreren, bootstrapping (simulaties), formules
- Extra: Bayesiaanse aanpak als alternatief



Concrete start

Zijn beide Black Boxjes gevuld met dezelfde proportie balletjes (populatie)?

Onderbouw je antwoord met (steekproef)data

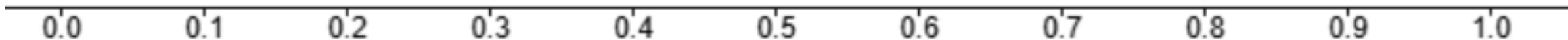


- A: Ja, ze zijn hetzelfde gevuld
- B: Nee, ze zijn verschillend gevuld
- C: Twijfel, ik weet het niet

Steekproeven, variatie, en variatieband

Stel dat 30% van een populatie met omvang 1000 een bepaalde eigenschap (kleur) heeft

Wat is het interval van veelvoorkomende steekproefresultaten (middelste 95%) bij een steekproef van 40?



 A: [0,17; 0,45]

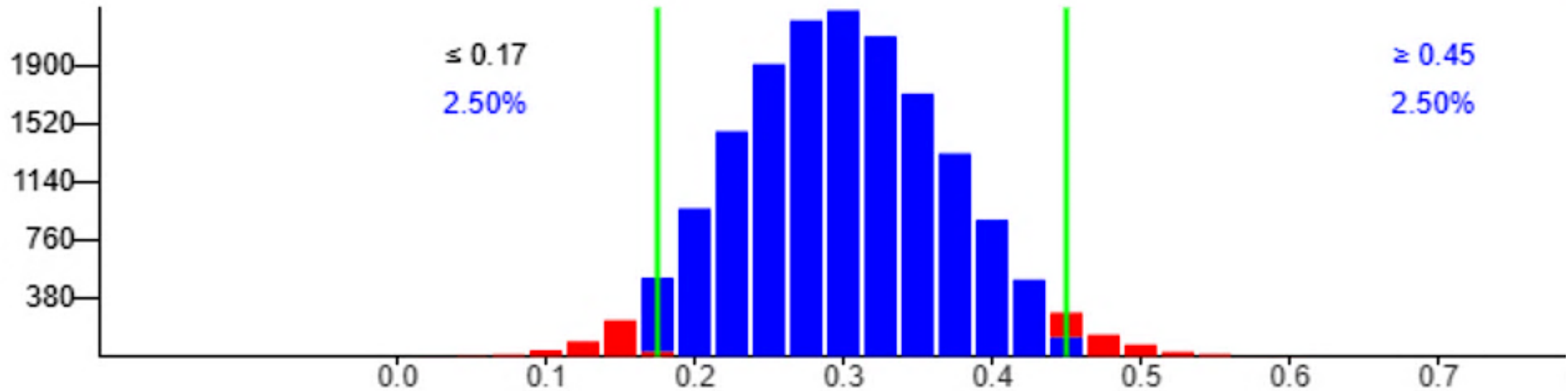
 B: [0,20; 0,40]

 C: [0,23; 0,37]

Variatieband

Simulatie VU-stat

<https://www.vustat.eu/apps/yesno/index.html?language=3>



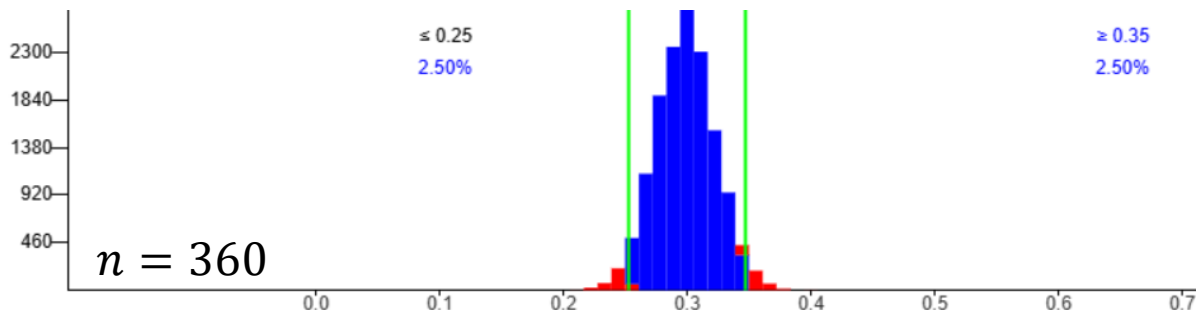
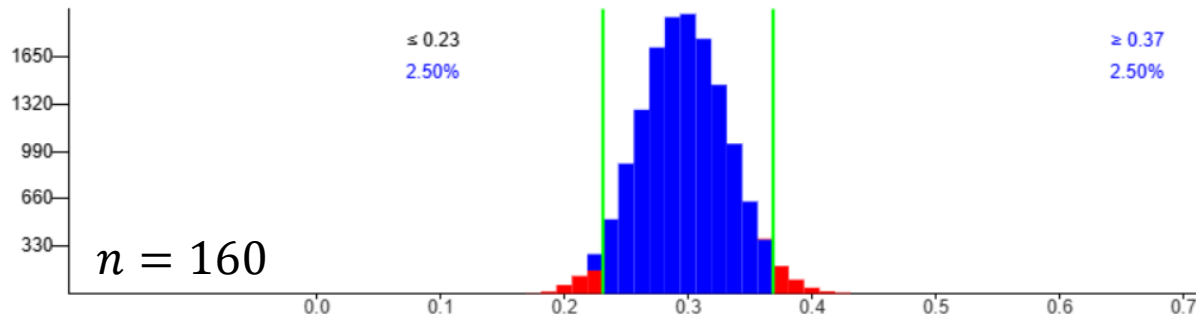
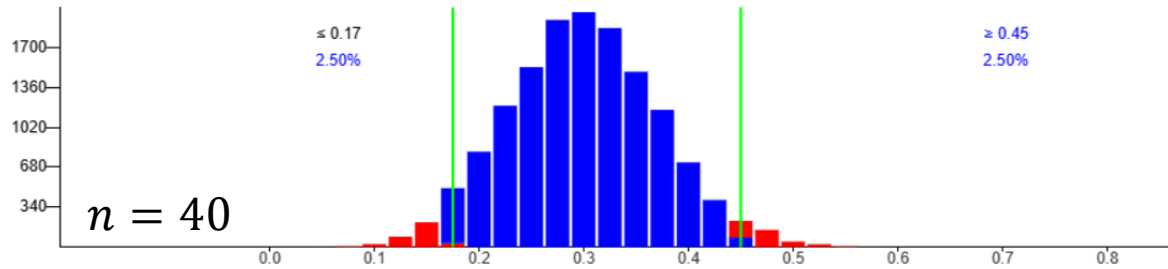
Variatieband: [0,17; 0,45]

Variatieband – de rol van steekproefomvang

Steekproef
omvang n

x4

x9



Kentallen

Aantal
Gemiddelde 0.300
Stand.afw. 0.072

Standaard
afwijking σ

$$\times \frac{1}{\sqrt{4}}$$

Kentallen

Aantal
Gemiddelde 0.300
Stand.afw. 0.036

$$\times \frac{1}{\sqrt{9}}$$

Kentallen

Aantal
Gemiddelde 0.300
Stand.afw. 0.024



Universiteit
Leiden

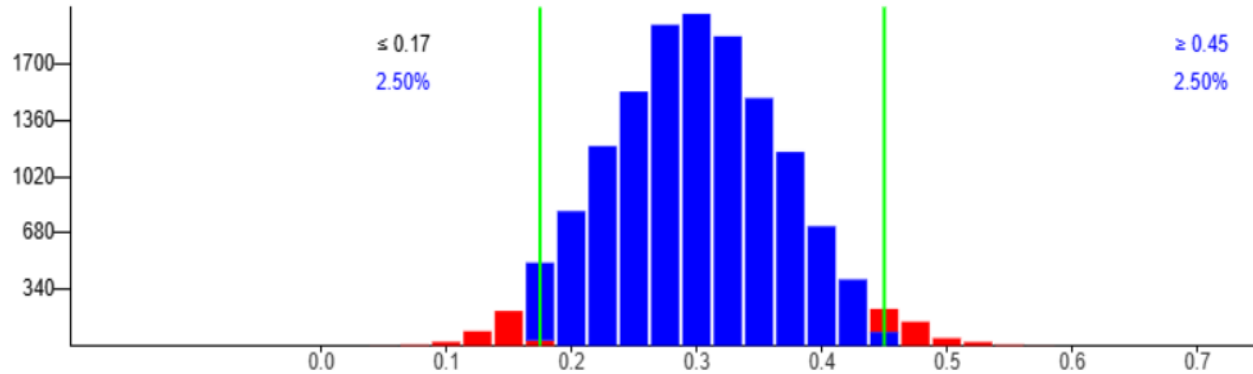


Universiteit
Utrecht

Variatieband - $\sqrt{n}$ wet

Steekproef
omvang

n



Kentallen

Aantal

Gemiddelde 0.300

Stand.afw. 0.072

Standaard
afwijking

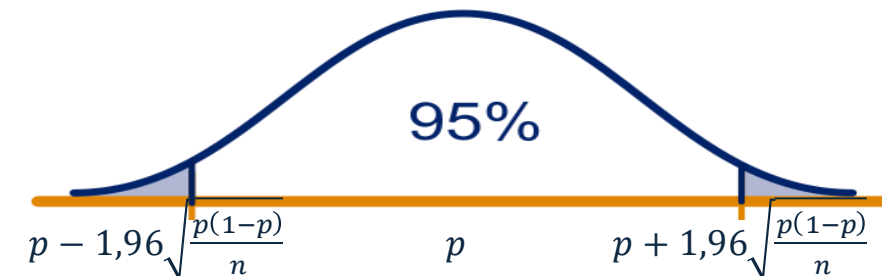
$$\frac{\sigma}{\sqrt{n}}$$

Voor de standaardafwijking van de steekproevenverdeling geldt dan $\sigma_s = \sqrt{\frac{p(1-p)}{n}}$

Voor de **variatieband** geldt

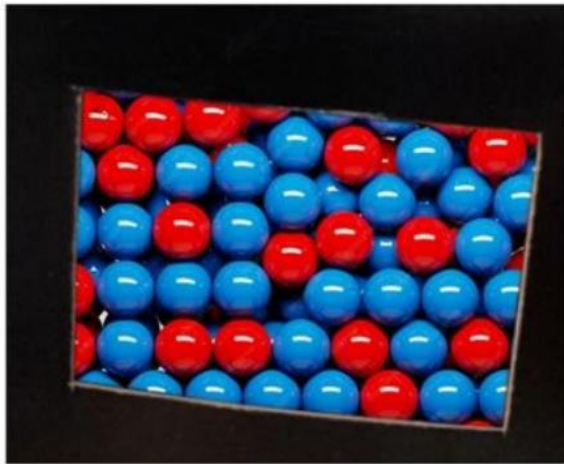
- bij proporties $[p - 1,96\sqrt{\frac{p(1-p)}{n}}, p + 1,96\sqrt{\frac{p(1-p)}{n}}]$, waarbij p de populatieproportie is;
- bij gemiddelden $[\mu - 1,96 \cdot \frac{\sigma}{\sqrt{n}}, \mu + 1,96 \cdot \frac{\sigma}{\sqrt{n}}]$, waarbij μ het populatiegemiddelde en σ de populatiestandaardafwijking is.

$$\left[0,3 - 1,96 \sqrt{\frac{0,3 \cdot 0,7}{40}}, p + 1,96 \sqrt{\frac{0,3 \cdot 0,7}{40}} \right] = [0,16; 0,44]$$

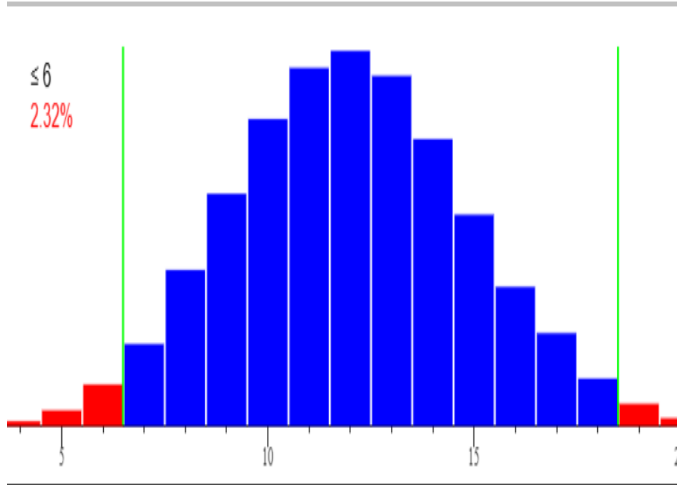


Didactische route

Informeel



Simulaties

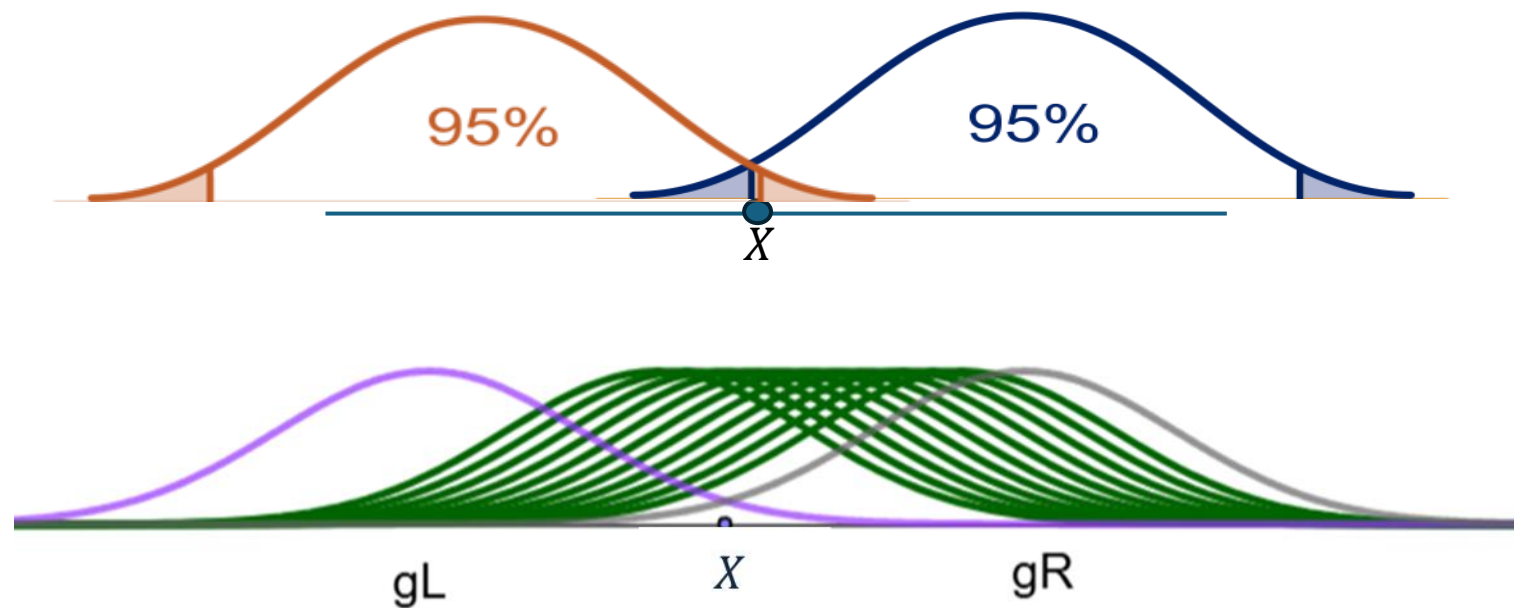


Formules

$$\left[p - 1,96 \sqrt{\frac{p(1-p)}{n}}, p + 1,96 \sqrt{\frac{p(1-p)}{n}} \right],$$

Van steekproef naar populatie?

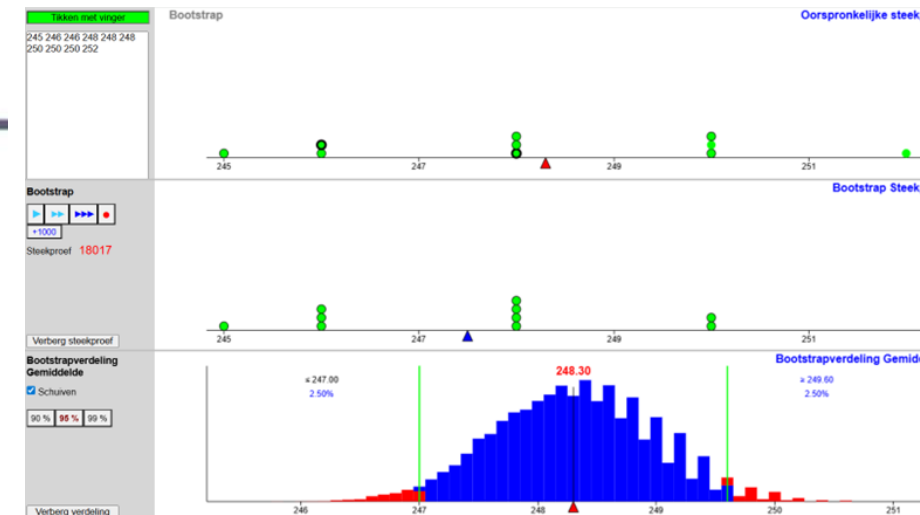
Steekproef met gemiddelde X en steekproefstandaardafwijking S



Via simulatie
Bootstrapmethode

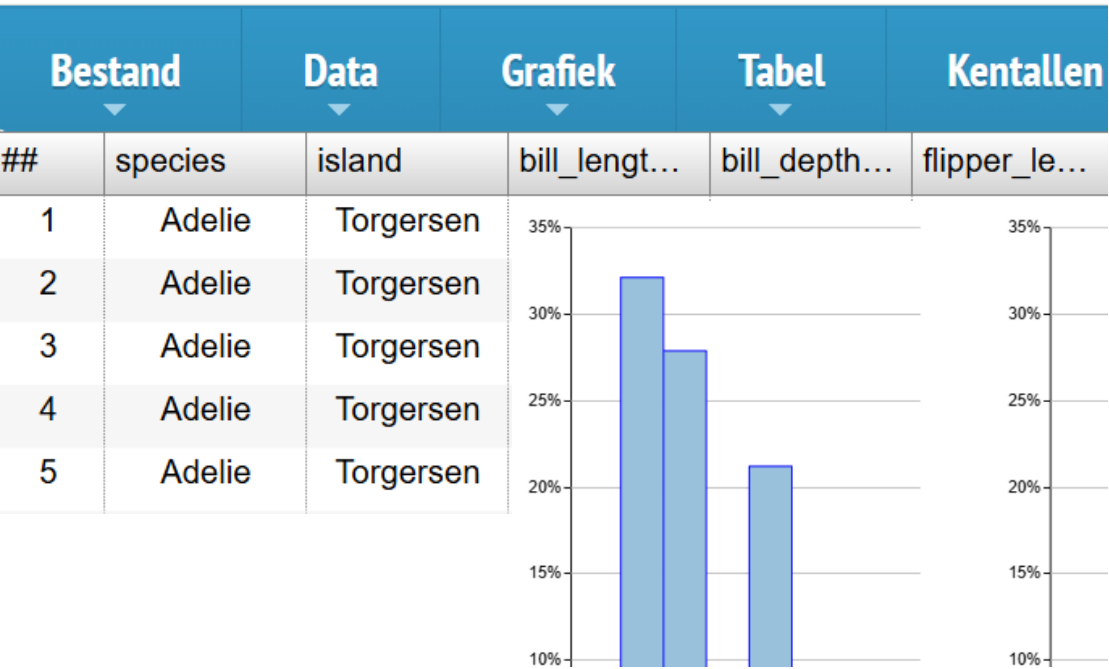
Betrouwbaarheidsinterval voor populatiegemiddelde

$$\left[X - 1,96 \cdot \frac{S}{\sqrt{n}} ; X + 1,96 \cdot \frac{S}{\sqrt{n}} \right]$$



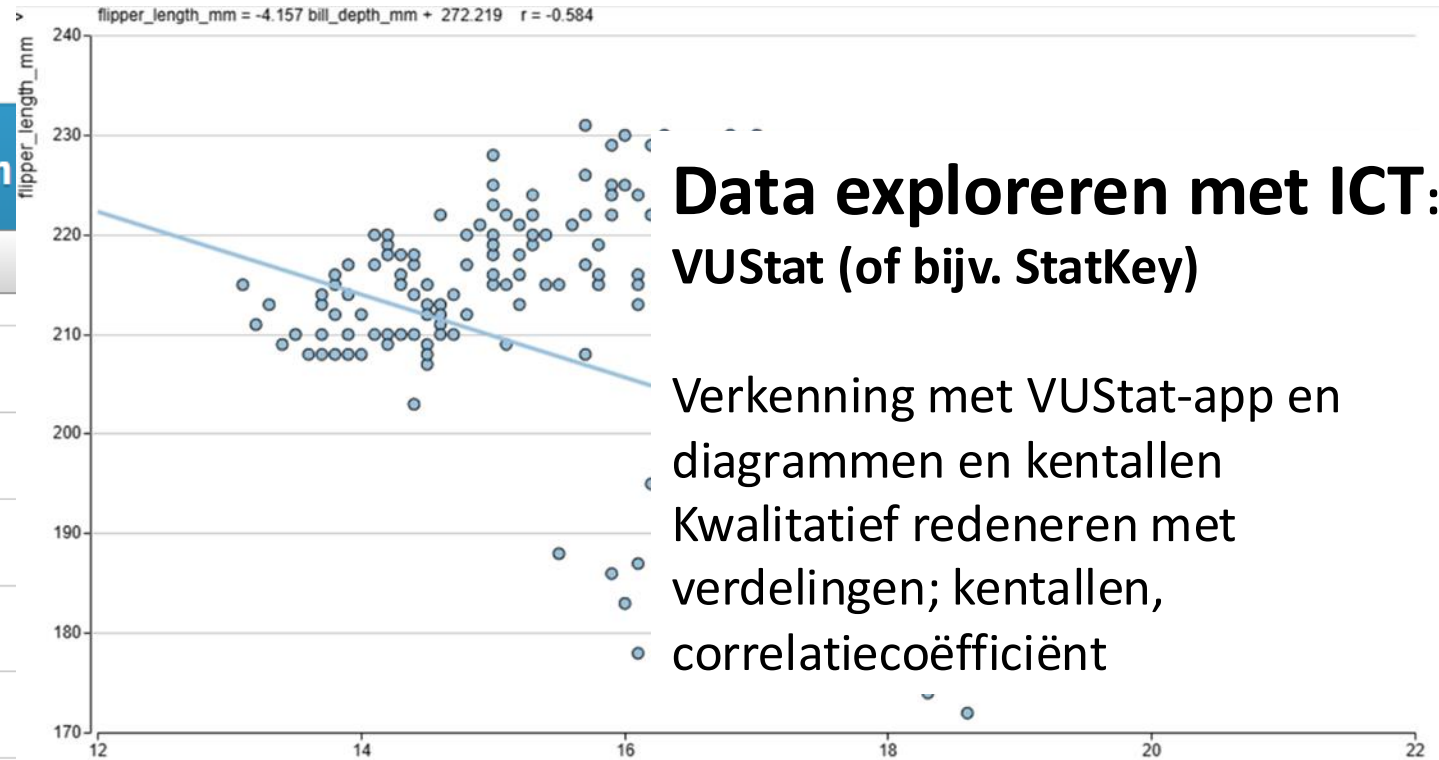
Empirische cyclus

Pinguïnbestand in VUstat



Korte onderzoeksvragen, denk aan:

- Wat is de body-mass van de pinguïnpopulatie?
- (Is er verschil tussen de mannetjes en vrouwtjes pinguïns m.b.t. body-mass?)



Via <https://www.vustat.eu/apps>
Bestand -> Datasets -> NWD datasets -> Pinguin.json

Bootstrapping

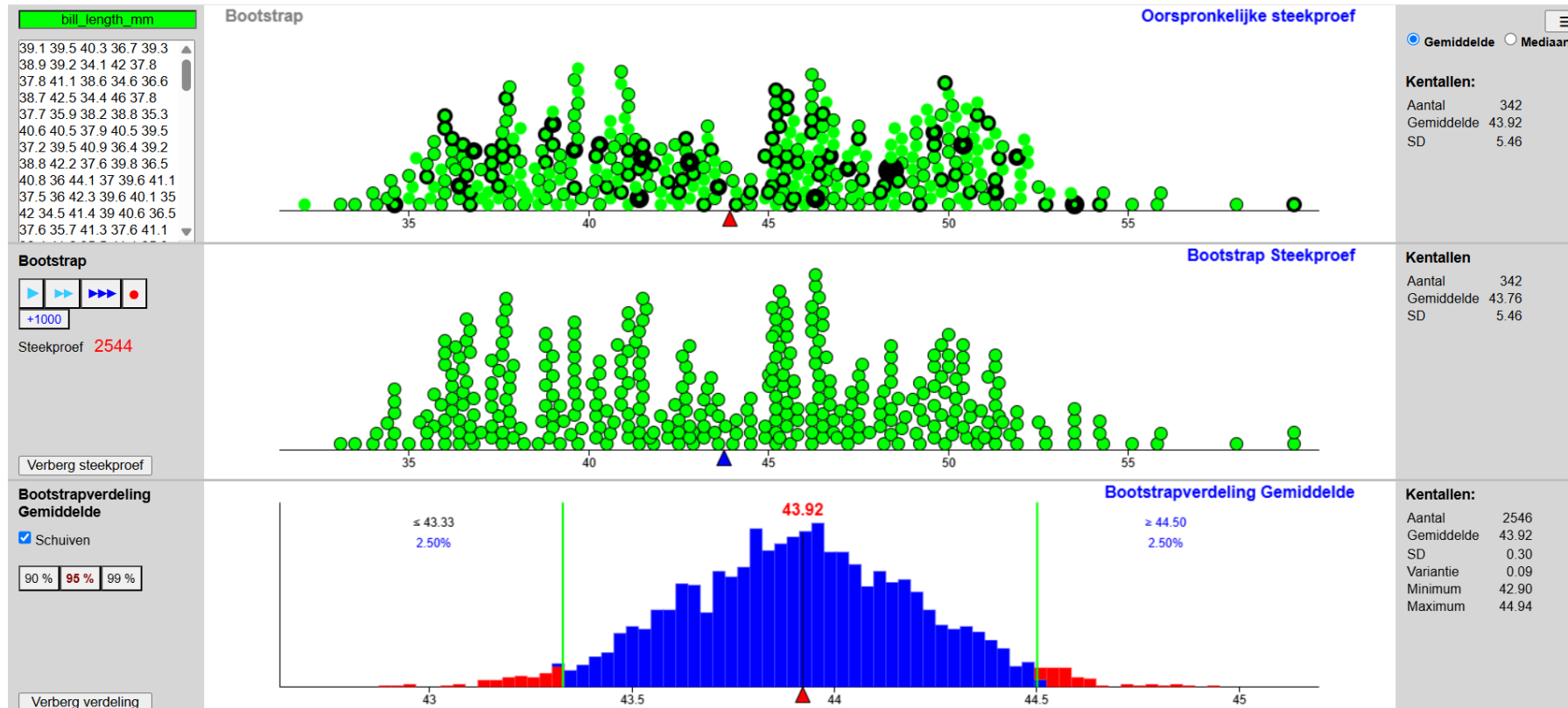
Wat is de body-mass van **de totale** pinguïnpopulatie ($\pm 5,9$ miljoen)?

BI via bootstrapping

Via <https://www.vustat.eu/apps>

Bestand -> Datasets -> NWD datasets -> Pinguin.json

Resampling → 1 variabele

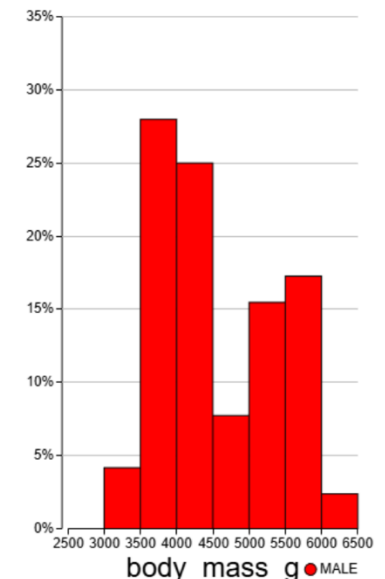
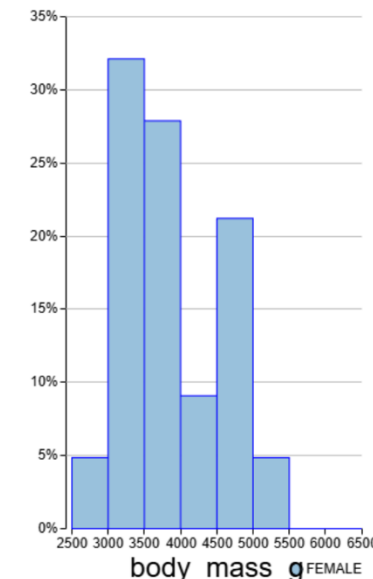
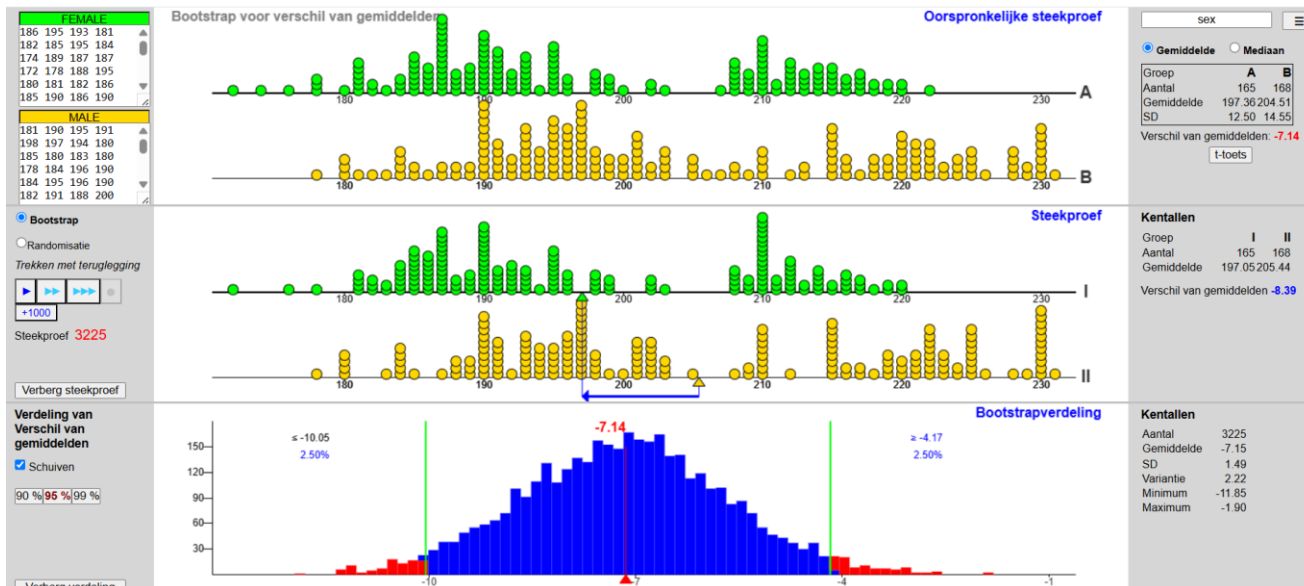


Groepen vergelijken

Is er verschil tussen de mannetjes en vrouwtjes pinguïns m.b.t. body-mass?



BI van verschil via bootstrapping of formules



demonstratiefilmpje zie <https://www.youtube.com/watch?v=oAA480O3U1E>

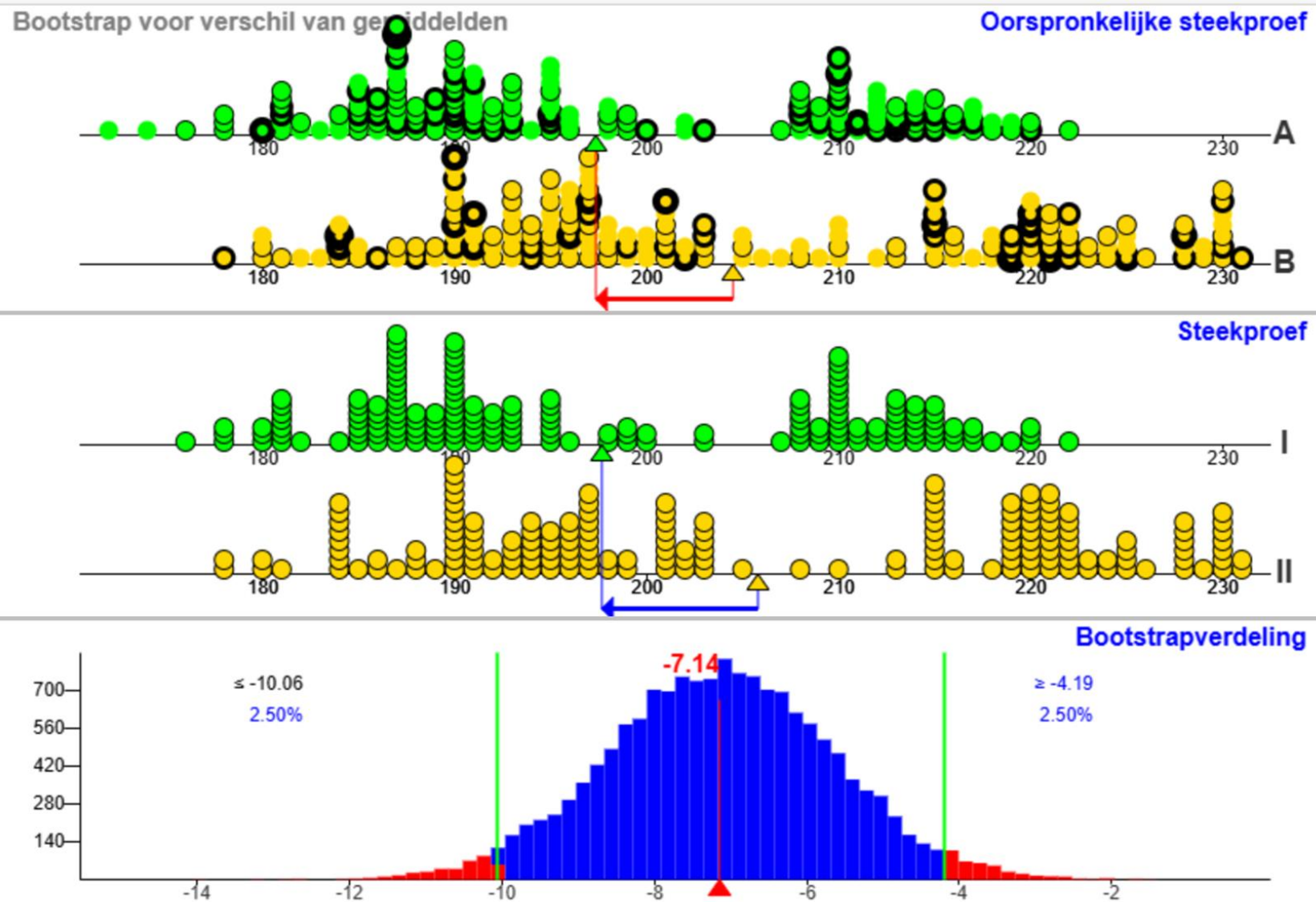
Korte onderzoeksvragen, denk aan:

- Wat is de body-mass van de pinguïnpopulatie?
- Is er verschil tussen de mannetjes en vrouwtjes pinguïns m.b.t. body-mass?

Groepen vergelijk

Is er verschil tussen de mar
body-mass?

	flipper_length_mm	
sex	FEMALE	MALE
Waarnemingen	165	168
Gemiddelde	197.4	204.5
SD	12.5	14.5
Minimum	172.0	178.0
Eerste kwartiel	187.0	193.0
Mediaan	193.0	200.5
Derde kwartiel	210.0	219.0
Maximum	222.0	231.0



$$BI \text{ Verschil: } X_1 - X_2 \pm \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

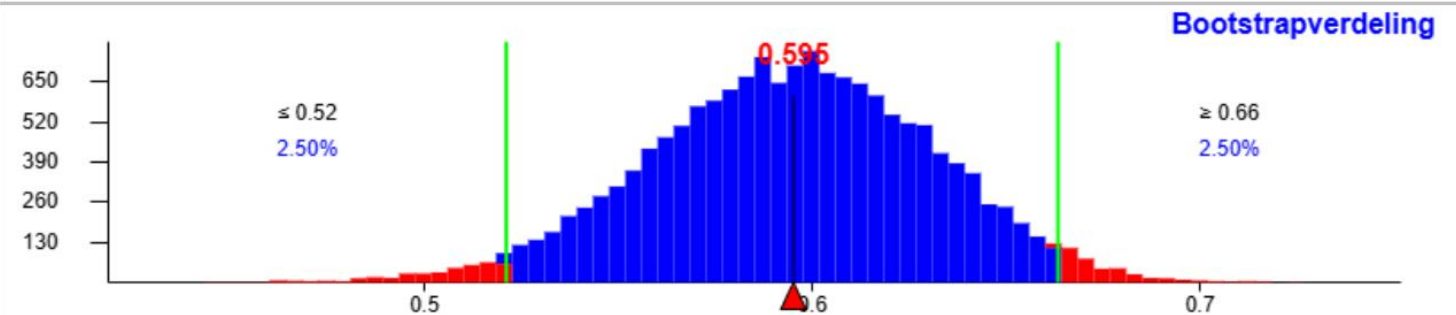
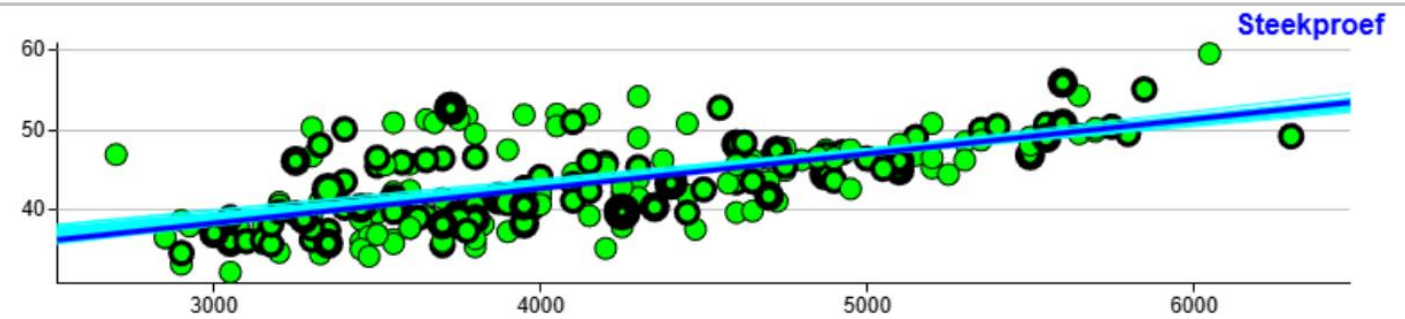
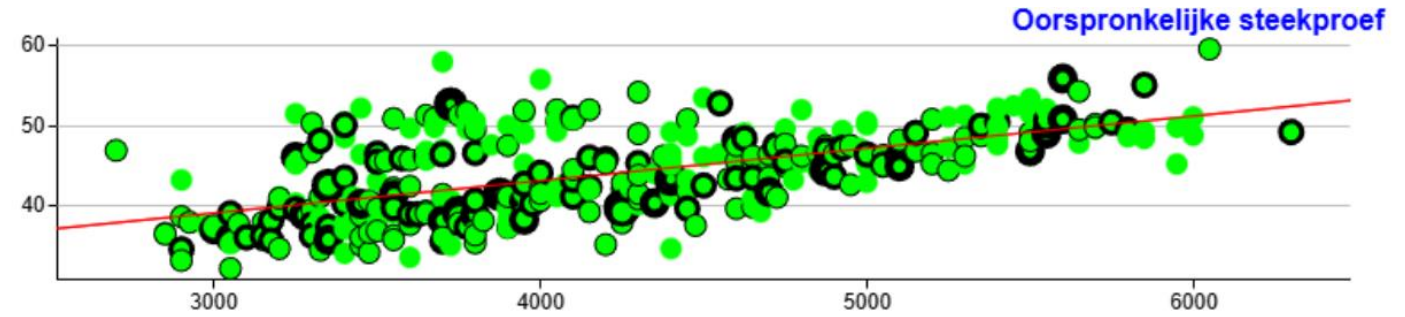
Hypothesetoetsen extra enkel bij verschil groepen H_0 vs H_1 ; berekening aannemen H_0 waar; bereken kans op waargenomen afwijking (of groter)(p-waarde)

Pilot in Vwo 5-6

Vooraf: data exploreren met AI
2 dagen (van 9-13) werken aan
inferentiële statistiek met
afsluitende onderzoeksopdracht

Leerlingen gebruiken veel
bootstrapping voor BI Verschil
(nauwelijks formules)

Maar ook bij correlatiecoëfficiënt
(was niet geoefend)



Aan de slag

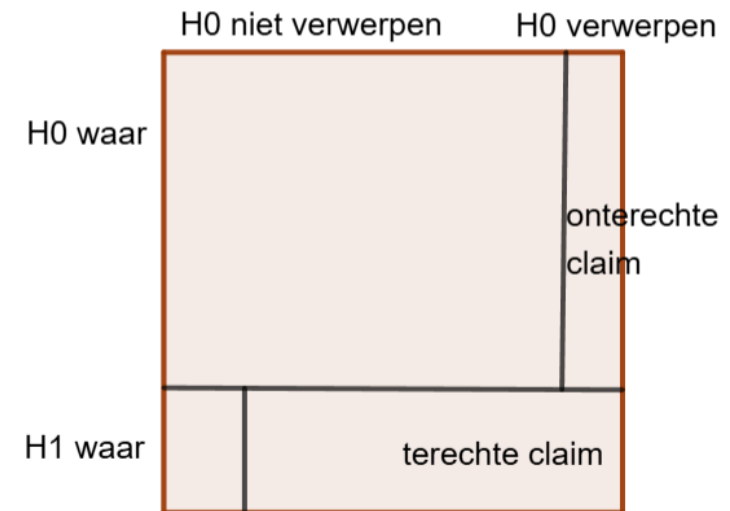


Opdracht: in twee of drietallen (20 min)

- Ga naar www.vustat.eu/apps/ en kies voor data-analyse
- Open dataset: Penguin.json
- Bedenk zelf onderzoeksvragen m.b.t. verschillen tussen groepen in populatie
- Formuleer eerst hypothese, maak diagrammen en kentallen en gebruik kwalitatief redeneren, trek voorlopige conclusies
- Onderbouw dit met betrouwbaarheidsintervallen via bootstrapping (en/of formules)

Kanttekening bij frequentistische benadering

- Interpretatie van betrouwbaarheidsintervallen en p-waarden (overschrijdingskansen) bij toetsen is lastig en subtiel. Fout is: “de kans is 95% dat echte populatiegemiddelde in BI ligt.”
- Zouden we niet een uitspraak willen als “met een kans van 95% ligt het echte populatiegemiddelde tussen ... en ...”? Of “gegeven de data, welke hypothese (nulhypothese of alternatieve hypothese) is meest waarschijnlijk?”
- Wat doen we met onze eventuele voorkennis?
- Wat als we eerder of later moeten stoppen met verzamelen van data
- Bij replicatiestudies blijkt steeds dat veel significante resultaten (H_0 verwerpen) niet opnieuw gevonden worden: veel claims onjuist



Geen test is 100% betrouwbaar

Gegeven:

- als je ziek bent test je positief met kans 0,95 (test positief = test zegt je bent ziek)
- als je niet ziek bent dan test je negatief met kans 0,90

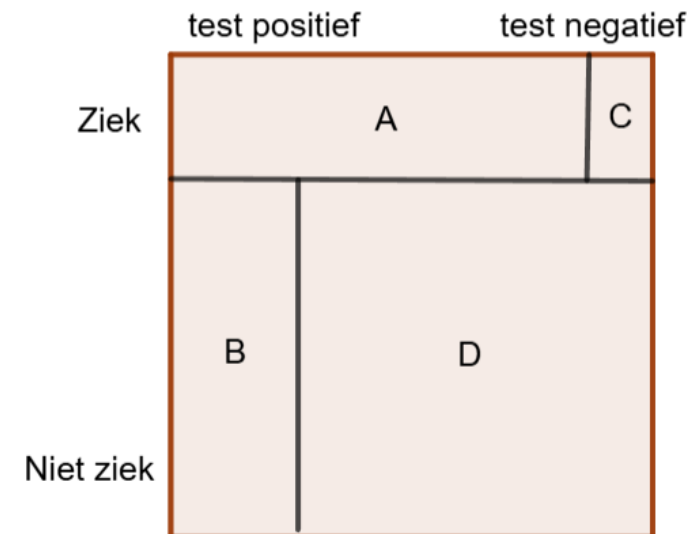
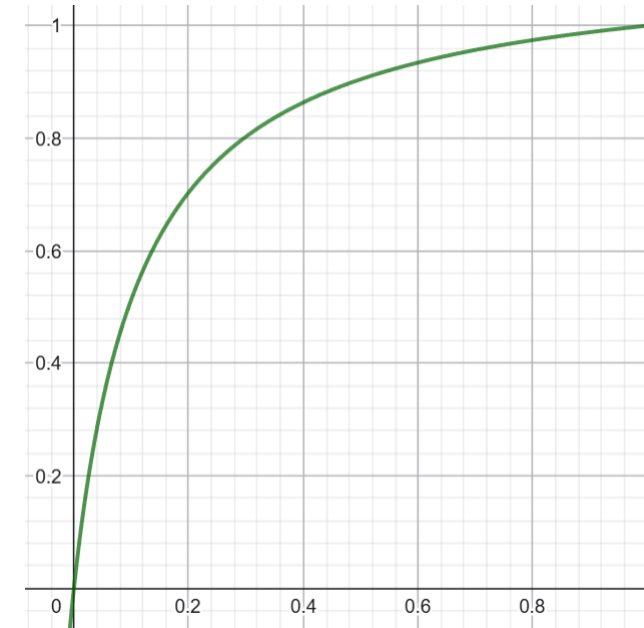
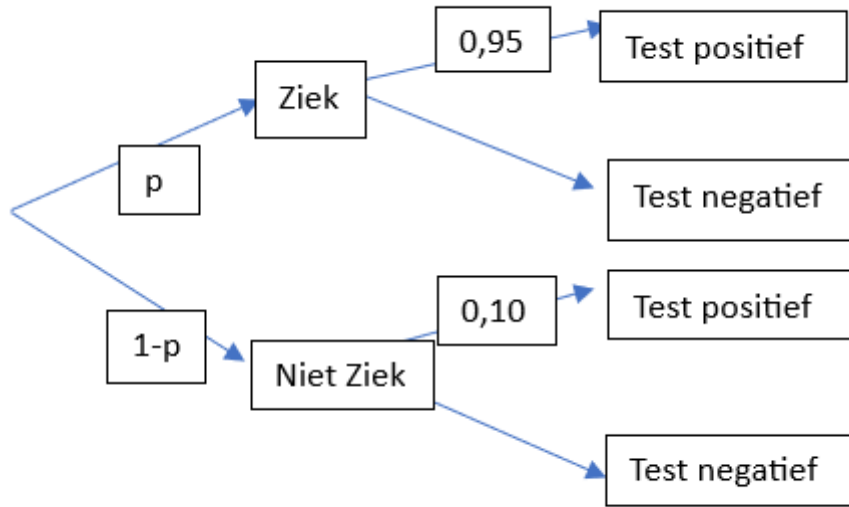
Iemand test positief.

Wat is de kans dat die inderdaad ziek is?

	Test positief	Test negatief	Totaal
Ziek	0,95	0,05	1
Niet ziek	0,10	0,90	1
Totaal			

Kans dat iemand die positief test ziek is?

$$\frac{0,95 \cdot p}{0,95 \cdot p + 0,10 \cdot (1-p)}$$



Bayes-tabel

Kies prevalentie; data: test positief

$$P(\text{ziek}|\text{test positief}) = \frac{P(\text{test positief}|\text{ziek}) \cdot P(\text{ziek})}{P(\text{test positief})}$$

	Prevalentie	Test positief	Prevalentie*Test positief	Kans op hypothese, gegeven test positief
Ziek	0,20	0,95	$0,20 \cdot 0,95 = 0,19$	$0,19/0,27 = 0,70$
Niet ziek	0,80	0,10	$0,80 \cdot 0,10 = 0,08$	$0,08/0,27 = 0,30$
Totaal			Totaal 0,27	1

In termen van Bayesiaanse statistiek

	Prior	P(data hypothese) = likelihood	Prior * likelihood	Posterior
Hypothese Ziek	0,20	0,95	$0,20 \cdot 0,95 = 0,19$	$0,19/0,27 = 0,70$
Hypothese Niet ziek	0,80	0,10	$0,80 \cdot 0,10 = 0,08$	$0,08/0,27 = 0,30$
Totaal			Totaal 0,27	1

$$P_{\text{posterior}}(\text{hypothese}|\text{data}) = \frac{P(\text{data}|\text{hypothese}) \cdot P_{\text{prior}}(\text{hypothese})}{P(\text{data})}$$

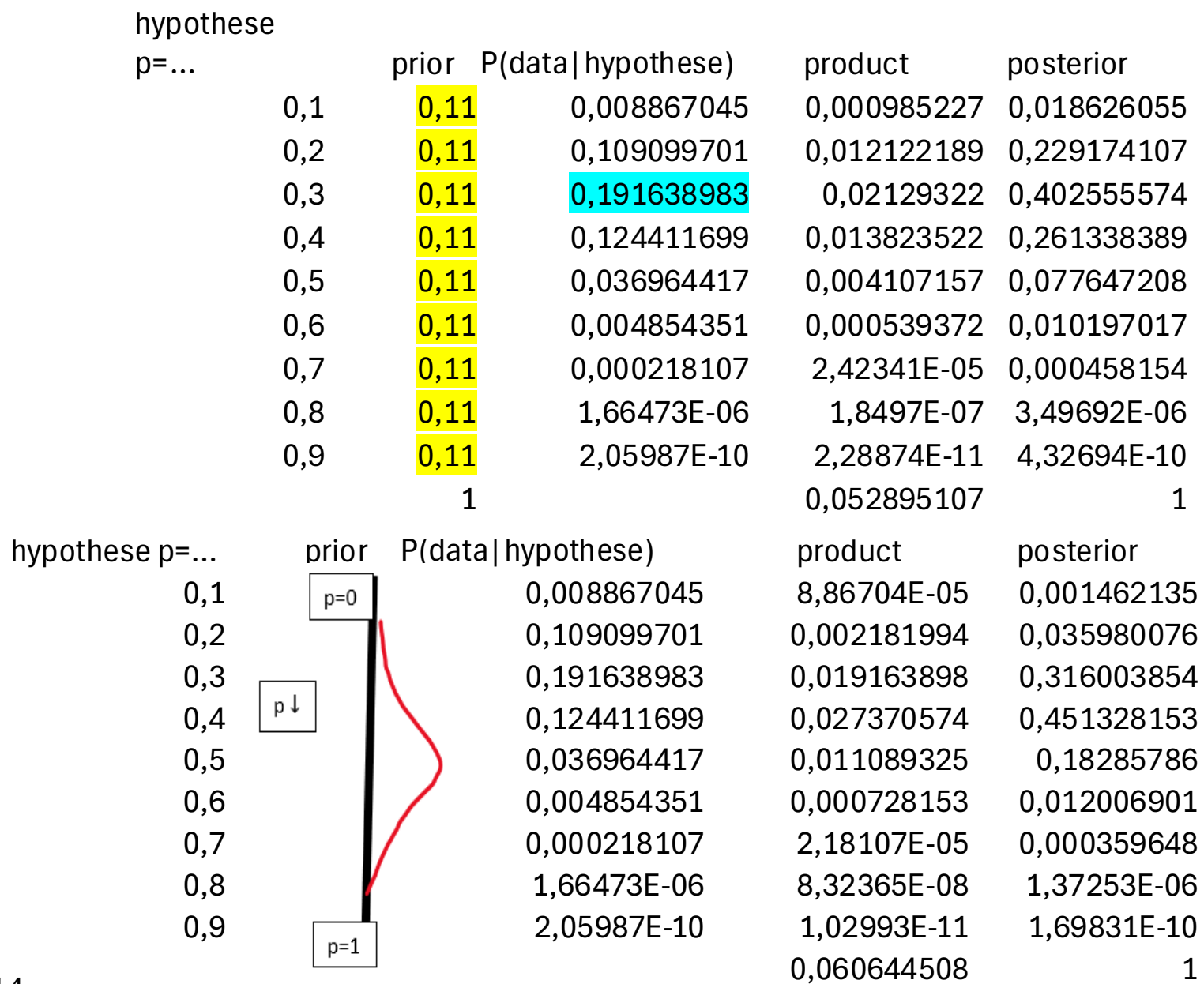
Hypothesen	Prior- kansen	P(data likelihood) =likelihood	Product Prior*likelihood	Posterior-kansen
H1	$P(H1)$	$P(data H1)$	$P(data H1) \cdot P(H1)$	$P(H1 data)$
H2	$P(H2)$	$P(data H2)$	$P(data H2) \cdot P(H2)$	$P(H2 data)$
....				
Totalen	1		$P(data)$	1

Onderzoeksvraag:
 Wat is de proportie
 leerlingen in het
 voortgezet onderwijs
 die afgelopen december
 regelmatig, dat wil
 zeggen ‘minstens drie
 uur’, naar de top 2000
 heeft geluisterd?

Data:
 Steekproef van 20
 leerlingen: 6 hebben
 regelmatig geluisterd.

$n=20, p=0,3, X=6$

Binomiale verdeling: $38760 \cdot p^6 \cdot (1 - p)^{14}$



Werken met continue priorverdeling

$$Beta_{a,b} = p^{a-1} \cdot (1-p)^{b-1}$$

$$\text{Max bij } p = \frac{a-1}{a+b-2}$$

Helaas geen kansdichtheidsfunctie

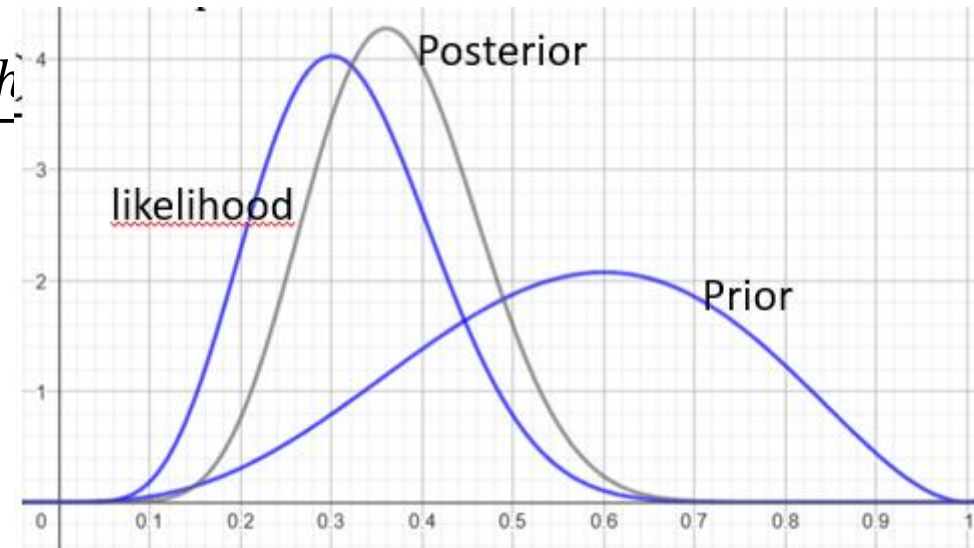
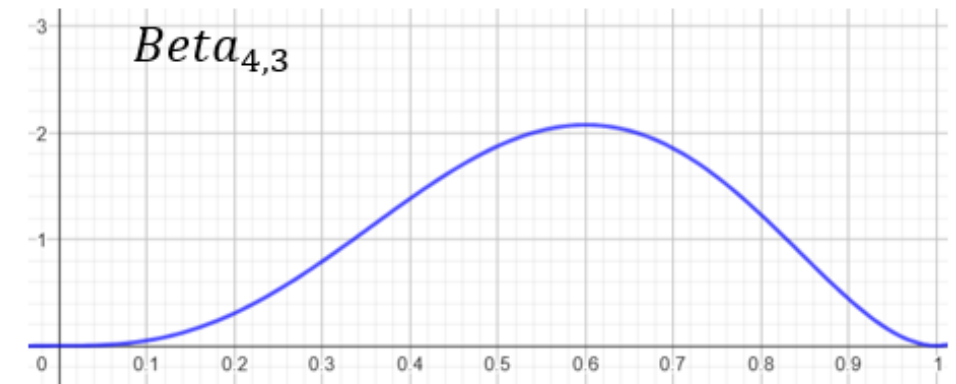
$$\text{Maar } Prior_{a,b} = \frac{p^{a-1}(1-p)^{b-1}}{\int_0^1 x^{a-1}(1-x)^{b-1}dx} \text{ wel!}$$

$$P_{posterior}(hypothesis|data) = \frac{P(data|hypothesis) \cdot P_{prior}(hypothesis)}{P(data)}$$

Werkwijze nu:

- Kies $P_{prior}(hypothesis): Beta_{4,3}$
- $P(data|hypothesis) = \text{likelihood} = 38760 \cdot p^6 \cdot (1-p)^{14}$
- de teller wordt: $p^{4-1} \cdot (1-p)^{3-1} \cdot 38760 \cdot p^6(1-p)^{14}$

$$\text{posteriorverdeling wordt : } Posterior = \frac{p^{4-1} \cdot (1-p)^{3-1} \cdot 38760 \cdot p^6(1-p)^{14}}{\int_0^1 x^{4-1} \cdot (1-x)^{3-1} \cdot 38760 \cdot x^6 \cdot (1-x)^{14} dx} \quad \frac{Beta_{(4+6,2+14)}}{\int_0^1 Beta_{(4+6,2+14)} dx}$$



Afsluiten op kern

Focus op kernconcepten en begrip

- Steeds ergens in empirische onderzoekscyclus bezig
- Kleine onderzoekjes voor exploreren van data en kwalitatief redeneren m.b.v. diagrammen en kentallen , correlatiecoëfficiënt
- Simulaties voor steekproevenverdeling (wortel-n-wet en CLS) en voor betrouwbaarheidsinterval (bootstrapping)
- Verschil tussen groepen onderzoeken m.b.v. betrouwbaarheidsinterval (bootstrapping en formules)
- Eerst frequentistische aanpak; maar Bayesiaanse statistiek misschien als keuzeonderwerp?
- Interpreteren van authentieke statistische informatie in media
Bijv. via 5 min wiskunde

Tabaksindustrie (als voorbeeld voor rechtszaken tegen sociale media)

De tabaksindustrie heeft lang gebruik gemaakt van het feit dat een oorzaak-gevolg verband (causaal verband) tussen roken en longkanker lastig is aan te tonen. (Zie Sanne Blauw in De Correspondent, 07-11-2018) over het zaaien van twijfel door tabaksindustrie.

Welke argumenten werden er gebruikt om twijfel te zaaien?

- a. Een gevonden verband tussen roken en longkanker berust op toeval berusten. Verklaring zou kunnen zijn: andere manier van tellen van longkanker patiënten die werd ingevoerd, of problemen met onderzoeksopzet.
- b. Er kan een ontbrekende factor zijn die zowel “oorzaak als gevolg” beïnvloedt. Zie bijv. ‘Als rokers minder goede resultaten op school halen, kun je dan zeggen dat het door roken komt? Nee, rokers zijn misschien socialer, die minder studeren vanwege drukke sociaal leven?’
- c. Misschien is verband wel andersom. Zie bijv. ‘Als Yale-afgestudeerden meer geld hebben dan de meesten van ons, komt dat dan doordat ze naar Yale zijn geweest? Nee, misschien komen er op Yale vooral studenten uit rijke families?’

Kleine ziekenhuizen in top?

Bij een onderzoek naar de kwaliteit van ziekenhuizen werd gekeken naar het voorkomen van complicaties bij operaties.

Op basis van de bevindingen werd een ranglijst van ziekenhuizen gemaakt.

Het bleek dat er relatief veel kleine ziekenhuizen bij de eerste twintig stonden.

Kun je beter naar klein ziekenhuis gaan?

Mogelijke reacties:

- a. Bij kleine ziekenhuizen nemen artsen meer de tijd.
- b. Als het percentage complicaties in alle ziekenhuizen hetzelfde zou zijn, speelt bij kleine ziekenhuizen de vanwege de kleine steekproefomvang de steekproefvariatie een grotere rol: lage en hoge percentages 'successen' zullen vaker voorkomen
- c. Het is een steekproef; van wege toeval zal er geen patroon zijn

Slechte tweede sprong

Een commentator bij een wedstrijd skischanssprongen voor heren waarbij de eindscore een combinatie is van twee afzonderlijke sprongen analyseert na afloop van de tweede sprong:

“Noorwegen had een geweldige eerste sprong, maar in de tweede verloor hij de wedstrijd”

Wat is het meest waarschijnlijk?

- a. Hij zal waarschijnlijk te gespannen zijn geweest in de hoop zijn voorsprong te beschermen.
- b. Dit is een statistisch verschijnsel en treedt heel vaak op.
- c. Het is puur toeval op toeval; we kunnen er niets over zeggen.

Interpreteren van p-waarde en betrouwbaarheidsinterval

Bekijk onderstaande uitspraken en geef bij iedere uitspraak aan of deze correct is.

1. De p-waarde (overschrijdingskans) is de kans dat de nulhypothese waar is; als een test van de nulhypothese bijvoorbeeld $p = 0,01$ geeft, heeft de nulhypothese slechts een kans van 1% om waar te zijn.
2. Een hoge p-waarde is een bewijs voor de alternatieve hypothese.
3. Het specifieke 95%-betrouwbaarheidsinterval dat door een onderzoek wordt gepresenteerd, is het interval waarin de echte waarde van de parameter met een kans van 95% ligt.
4. Als 0 niet in het 95%-BI van het verschil tussen twee populatiegemiddelden ligt dat is er sprake van een significant verschil.
5. Als twee BI overlappen dan is het verschil tussen de populatiegemiddelden niet significant.

Nog vragen??

Bedankt voor uw belangstelling!

Inspiratie:

Zebraboekje Inferentiele Statistiek (over frequentistische aanpak en Bayesiaanse aanpak; en over E-waarden)

Leerlijn Statistiek op site NVVW (Kop & Van Barneveld, 2022) [Leerlijn statistiek voor vwo](#)

Lesmaterialen Black box – statistiek

<https://www.nvvw.nl/lesmaterialen/projectenfonds/blackbox/>

MATHerials: Toolkits – Tastbare materialen en activiteiten voor hands-on wiskunde

