



Bayesiaanse statistiek in de audit — *Terugkomdag Belastingdienst 2025*

Ferry Geertman & Koen Derks

1

Introductie sprekers

Ferry Geertman

- (IT) Auditor
- Docent Nyenrode data science
- Docent Tilburg University
- Docent Erasmus
- Secretaris Stuurgroep Statistical Auditing

Dr. Koen Derks

- Universitair docent @ Nyenrode
- Center of Accounting, Auditing & Control
- Statisticus
- Lid Stuurgroep Statistical Auditing
- www.statisticalauditing.com

Deel 1:

Introductie Bayesiaanse statistiek

Of: Hoe maak je statistiek intuïtief voor auditors?

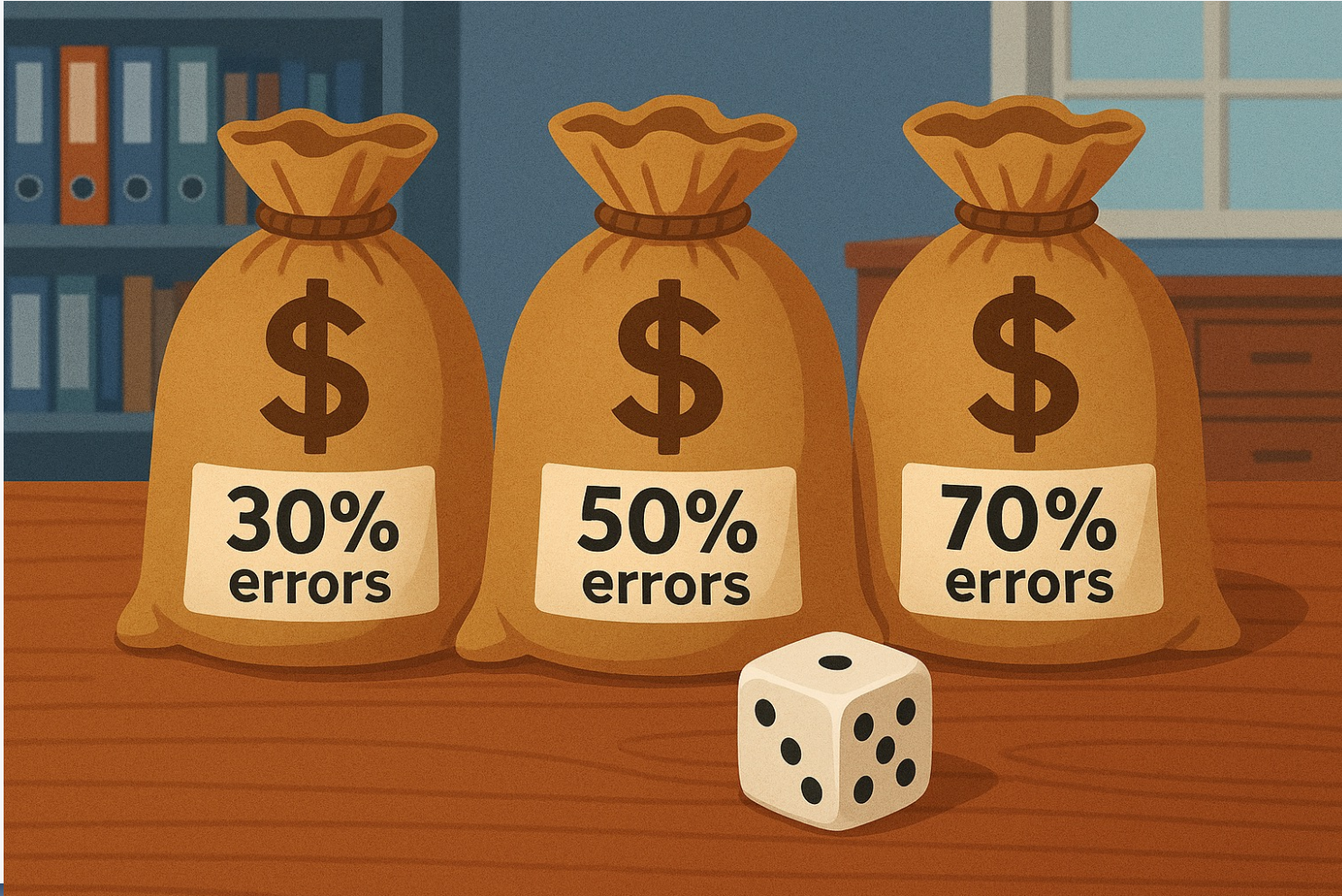
Bayesiaanse analyse - principe

- De (on)zekerheid of “mate van overtuiging” wordt gekwantificeerd met waarschijnlijkheden (kansen), de prior
- De waargenomen data wordt gebruikt om de mate van overtuiging, de ‘prior’ aan te passen
- Deze aanpassing leidt tot een nieuwe “mate van overtuiging” – de ‘posterior’



Weer nieuwe informatie





- Extra info:
- We selecteren 1 euro uit de zak in vorige slide door worp dobbelsteen bepaald is.
- Die euro blijkt fout te zijn
- Uit welke zak geld komt deze euro?

Bayesiaanse aanpak

- **Focus:** De parameter θ is *onbekend* en wordt benaderd als een kansverdeling — we hebben dus een **geloof/overtuiging** over wat θ zou kunnen zijn.
- **Data + voorkennis combineren:**
 - Je begint met een **prior** (je uitgangspositie over θ).
 - Je verzamelt data en berekent de **likelihood** (hoe waarschijnlijk is de data gegeven bepaalde waarden van θ ?).
 - Je combineert die twee om een **posterior** te krijgen (nieuwe inschatting van θ na het zien van de data).
- Interpretatie: Kans is een maat voor *geloof* of *zekerheid* over een parameter, gegeven zowel data als eerdere aannames.

Frequentistische aanpak

- **Focus:** De parameter (bijv. het foutpercentage van een populatie) is *vast*, maar onbekend.
- **Data is variabel:** Als je het experiment 100 keer herhaalt, krijg je steeds andere samples en andere schattingen. De data varieert dus, maar de parameter is hetzelfde.
- **Interpretatie:** De kans komt voort uit *frequentie van gebeurtenissen* als je het experiment herhaalt. Bijvoorbeeld: "Als ik dit 100 keer zou doen, zou het 95% van de tijd binnen dit interval vallen."

In één overzicht

Aspect	Frequentistisch	Bayesiaans
Parameter	Vast en onbekend	Willekeurig met een kansverdeling
Kans	Gebaseerd op herhaling	Geloof/subjectieve onzekerheid
Wat varieert?	Data	Geloof over de parameter
Wat krijg je?	Puntschatting + betrouwbaarheidsinterval	Volledige kansverdeling (posterior)

Hoe ondersteunt de wiskunde dit alles

- Je 'overtuigingen' moeten voldoen aan de wetten van kans
 - Groter of gelijk aan 0
 - Optellen to 1
- Posterior verdelingen die tot dezelfde familie behoren als de prior-verdeling
 - Een beta-prior met een binomial likelihood leidt tot een beta posterior
 - Een gamma-prior met een Poisson likelihood leidt tot een gamma posterior
 - Dit rekt makkelijk
- Als die niet kan/mogelijk is dan extra rekenwerk nodig

Kenmerken van de Beta-verdeling

Eigenschap	Betekenis	
Gedefinieerd op	Interval van 0 tot 1 (fracties)	
Vormparameters	a en b (α en β)	
Flexibiliteit	Kan zeer veel vormen aannemen (scheef, U-vorming, klokvormig)	
Interpretatie	a – 1 # fouten, b -1 #goede	
Gemiddelde	$a/(a+b)$	
Modus	$(a-1) / (a+b-2)$	Als a > 1, b >1

Intermezzo:

Introductie Shiny app

Zelf aan de slag

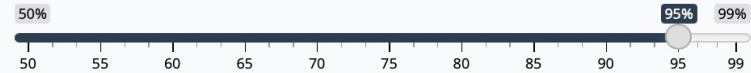
Basis

Regressie

Classificatie

Controleparameters

Betrouwbaarheid



Materialiteit



► Instellingen voor de prior-verdeling

► Gegevens uit de steekproef

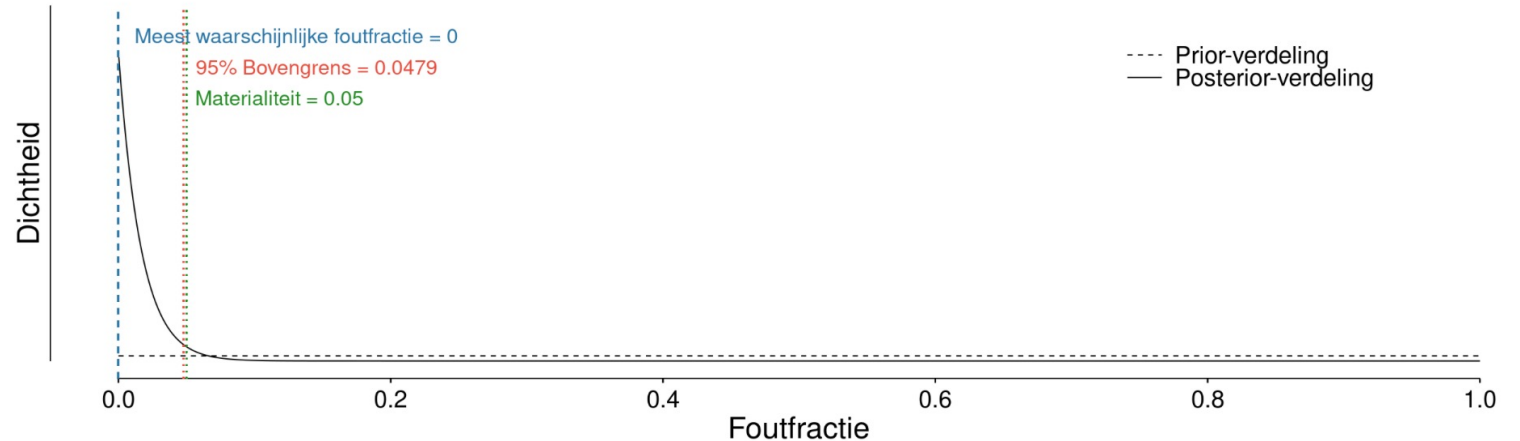
► Geavanceerde opties

Uitleg Bayesiaanse statistiek

Statistiek	Waarde
Meest waarschijnlijke foutfractie	0
95% Bovengrens voor de foutfractie	0.0479
Kansmassa onder materialiteit	0.9562
Kansmassa boven materialiteit	0.0438

Interpretatie:

De prior-verdeling is **Beta(1, 1)**. Na het zien van een steekproef van **60** met **0** fouten is de meest waarschijnlijke foutfractie in de populatie **0**. De 95% bovengrens voor deze foutfractie ligt op **0.0479**. Dit betekent dat de kans dat de foutfractie in de populatie onder de 4.79% ligt gelijk is aan 95%. Daarnaast is de kansmassa onder de materialiteit **0.9562**, wat betekent dat de kans dat de werkelijke, maar onbekende, foutfractie onder de materialiteit ligt 95.62% is.



Link naar applicatie: <https://koenderks.shinyapps.io/BayesDemo/>

13

Basis

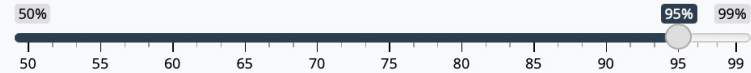
Regressie

Classificatie

Tabbladen

Controleparameters

Betrouwbaarheid



Materialiteit



► Instellingen voor de prior-verdeling

► Gegevens uit de steekproef

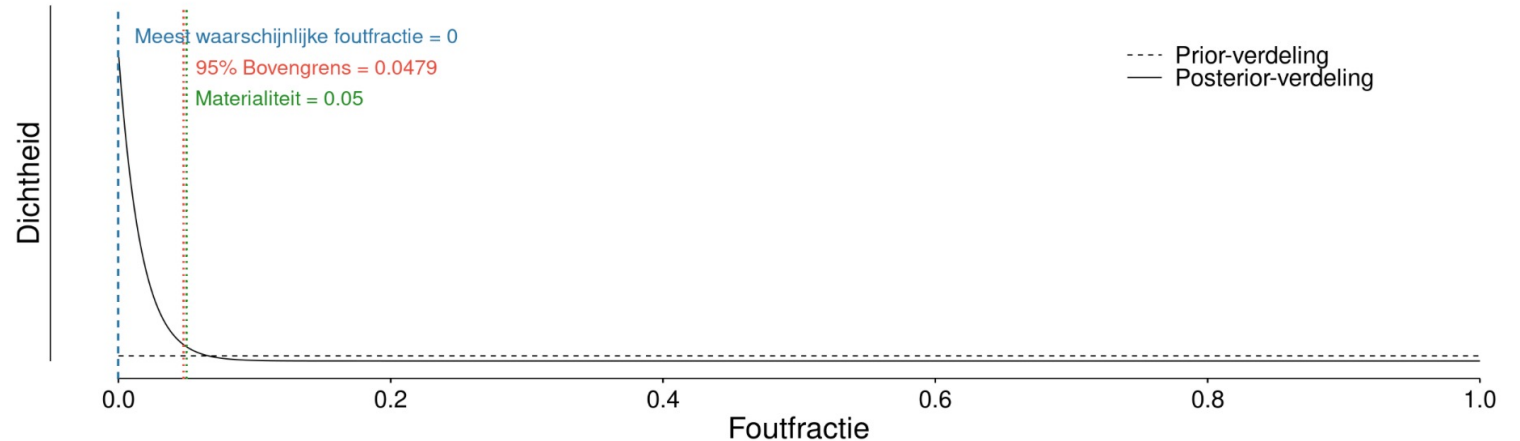
► Geavanceerde opties

► Uitleg Bayesiaanse statistiek

Statistiek	Waarde
Meest waarschijnlijke foutfractie	0
95% Bovengrens voor de foutfractie	0.0479
Kansmassa onder materialiteit	0.9562
Kansmassa boven materialiteit	0.0438

Interpretatie:

De prior-verdeling is $\text{Beta}(1, 1)$. Na het zien van een steekproef van 60 met 0 fouten is de meest waarschijnlijke foutfractie in de populatie 0. De 95% bovengrens voor deze foutfractie ligt op **0.0479**. Dit betekent dat de kans dat de foutfractie in de populatie onder de 4.79% ligt gelijk is aan 95%. Daarnaast is de kansmassa onder de materialiteit **0.9562**, wat betekent dat de kans dat de werkelijke, maar onbekende, foutfractie onder de materialiteit ligt 95.62% is.



Link naar applicatie: <https://koenderks.shinyapps.io/BayesDemo/>

Basis

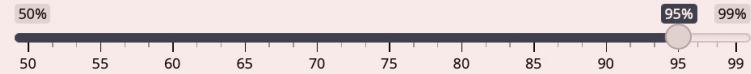
Regressie

Classificatie

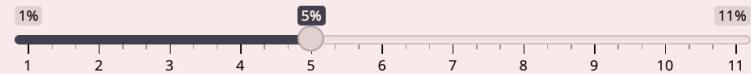
Tabbladen

Controleparameters

Betrouwbaarheid



Materialiteit



► Instellingen voor de prior-verdeling

► Gegevens uit de steekproef

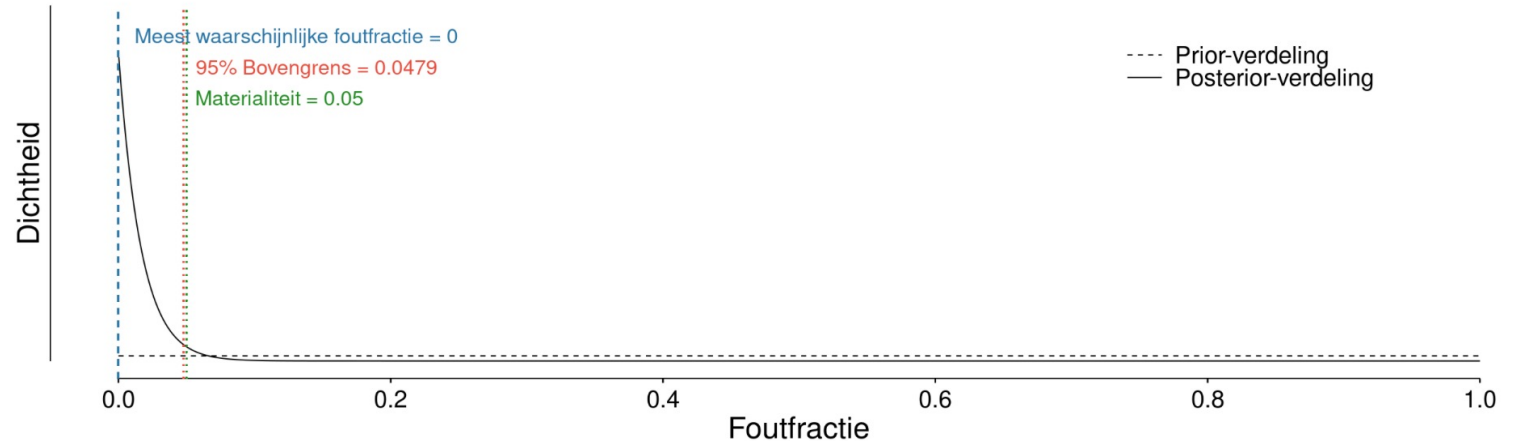
► Geavanceerde opties

Uitleg Bayesiaanse statistiek

Statistiek	Waarde
Meest waarschijnlijke foutfractie	0
95% Bovengrens voor de foutfractie	0.0479
Kansmassa onder materialiteit	0.9562
Kansmassa boven materialiteit	0.0438

Interpretatie:

De prior-verdeling is $\text{Beta}(1, 1)$. Na het zien van een steekproef van 60 met 0 fouten is de meest waarschijnlijke foutfractie in de populatie 0. De 95% bovengrens voor deze foutfractie ligt op **0.0479**. Dit betekent dat de kans dat de foutfractie in de populatie onder de 4.79% ligt gelijk is aan 95%. Daarnaast is de kansmassa onder de materialiteit **0.9562**, wat betekent dat de kans dat de werkelijke, maar onbekende, foutfractie onder de materialiteit ligt 95.62% is.



Link naar applicatie: <https://koenderks.shinyapps.io/BayesDemo/>

Basis

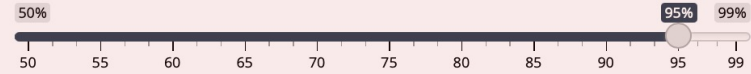
Regressie

Classificatie

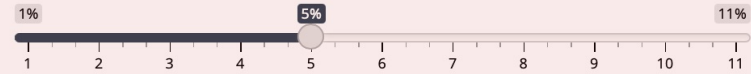
Tabbladen

Controleparameters

Betrouwbaarheid



Materialiteit



► Instellingen voor de prior-verdeling

► Gegevens uit de steekproef

► Geavanceerde opties

► Uitleg Bayesiaanse statistiek

Statistiek

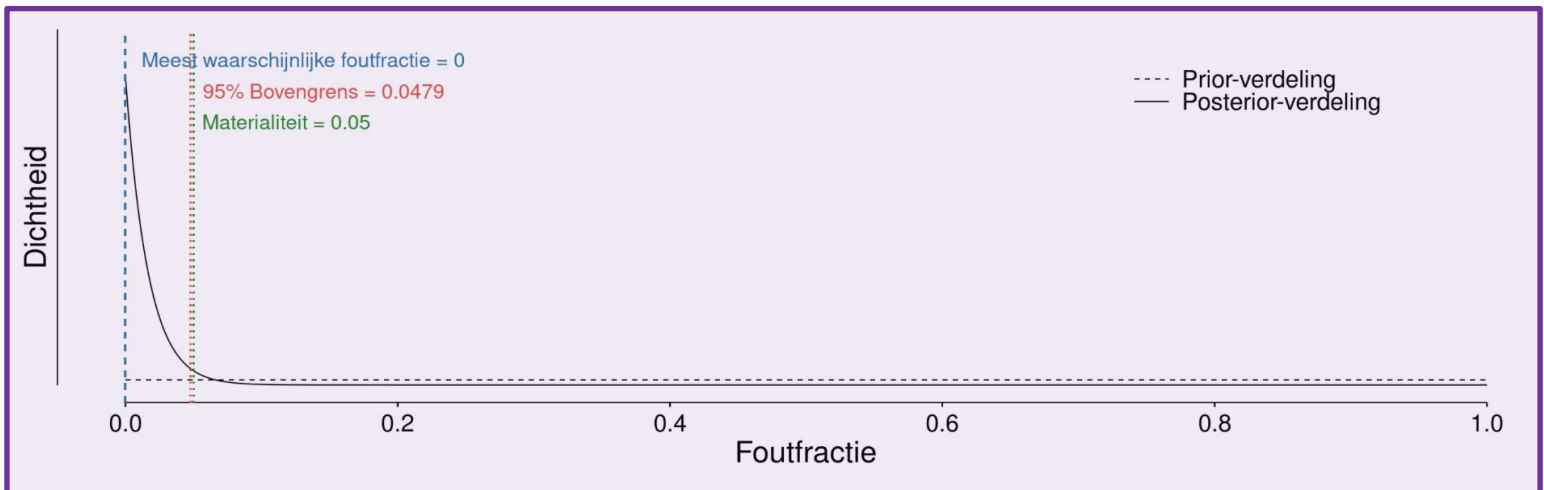
Uitkomsten

Waarde

Meest waarschijnlijke foutfractie	0
95% Bovengrens voor de foutfractie	0.0479
Kansmassa onder materialiteit	0.9562
Kansmassa boven materialiteit	0.0438

Interpretatie:

De prior-verdeling is **Beta(1, 1)**. Na het zien van een steekproef van 60 met 0 fouten is de meest waarschijnlijke foutfractie in de populatie 0. De 95% bovengrens voor deze foutfractie ligt op **0.0479**. Dit betekent dat de kans dat de foutfractie in de populatie onder de 4.79% ligt gelijk is aan 95%. Daarnaast is de kansmassa onder de materialiteit **0.9562**, wat betekent dat de kans dat de werkelijke, maar onbekende, foutfractie onder de materialiteit ligt 95.62% is.



Link naar applicatie: <https://koenderks.shinyapps.io/BayesDemo/>

Basis

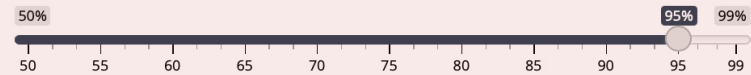
Regressie

Classificatie

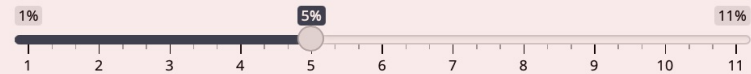
Tabbladen

Controleparameters

Betrouwbaarheid



Materialiteit



► Instellingen voor de prior-verdeling

► Gegevens uit de steekproef

► Geavanceerde opties

Input opties voor tabblad

► Uitleg Bayesiaanse statistiek

Uitleg

Statistiek

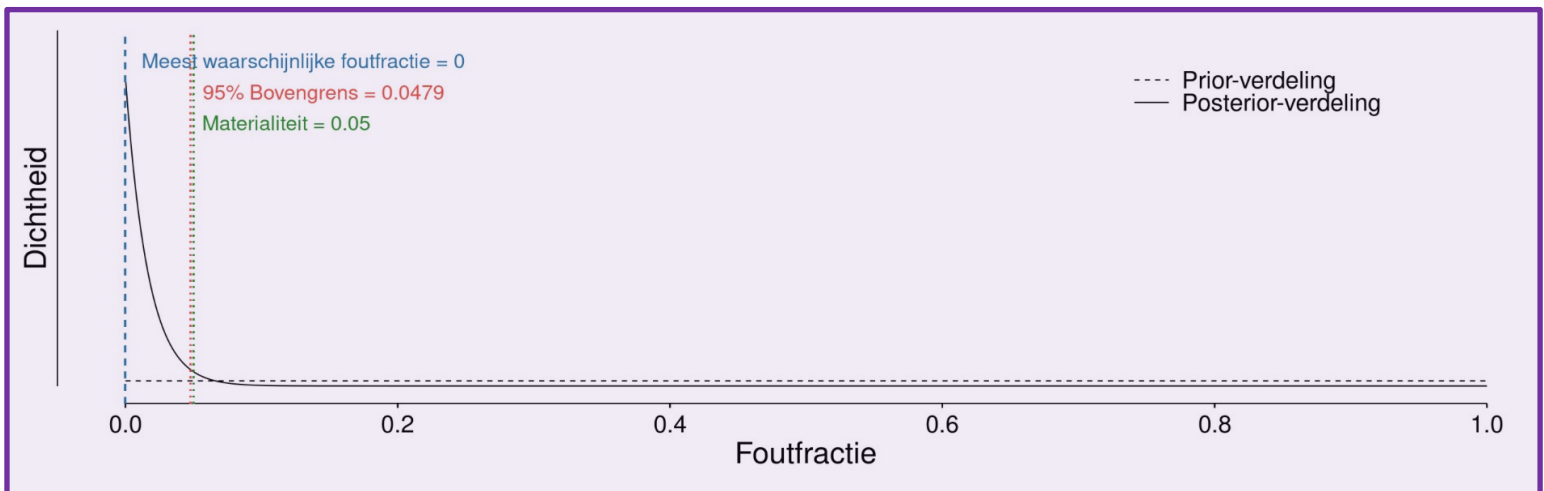
Uitkomsten

Waarde

Meest waarschijnlijke foutfractie	0
95% Bovengrens voor de foutfractie	0.0479
Kansmassa onder materialiteit	0.9562
Kansmassa boven materialiteit	0.0438

Interpretatie:

De prior-verdeling is $\text{Beta}(1, 1)$. Na het zien van een steekproef van 60 met 0 fouten is de meest waarschijnlijke foutfractie in de populatie 0. De 95% bovengrens voor deze foutfractie ligt op **0.0479**. Dit betekent dat de kans dat de foutfractie in de populatie onder de 4.79% ligt gelijk is aan 95%. Daarnaast is de kansmassa onder de materialiteit **0.9562**, wat betekent dat de kans dat de werkelijke, maar onbekende, foutfractie onder de materialiteit ligt 95.62% is.



Link naar applicatie: <https://koenderks.shinyapps.io/BayesDemo/>

17

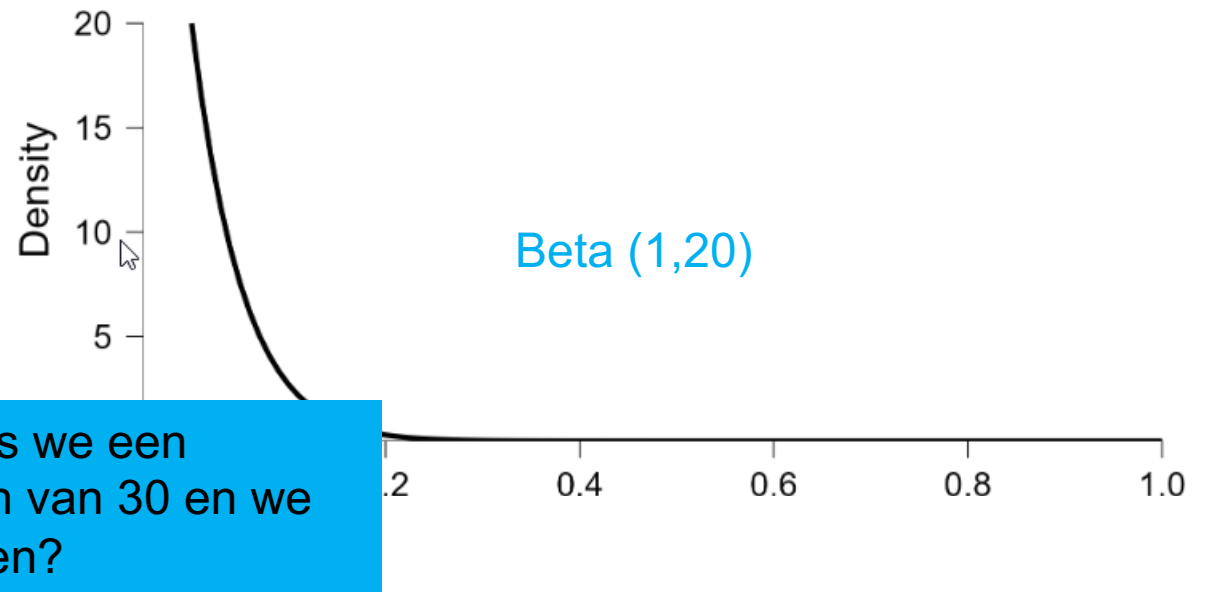
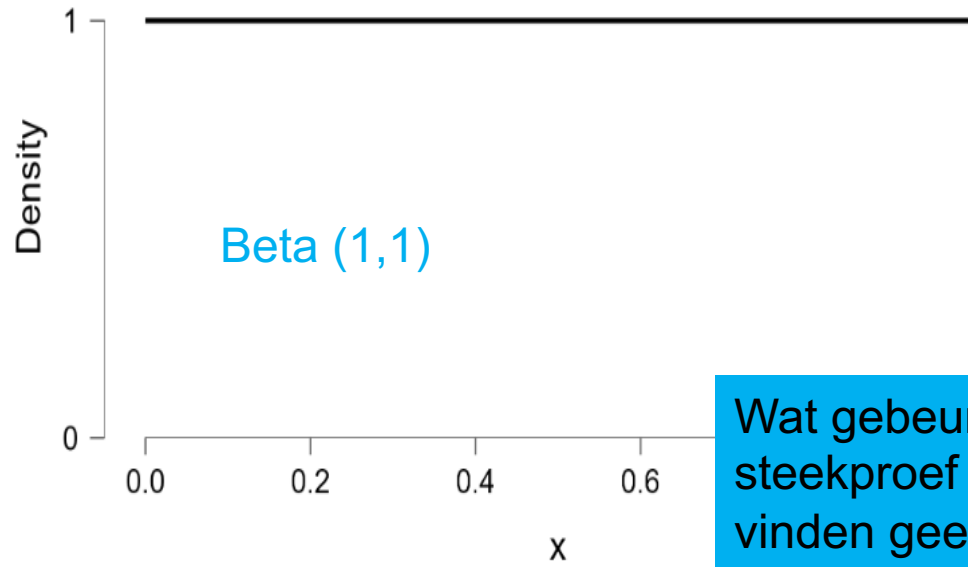
Zelf aan de slag met de BayesDemo applicatie

- Opdracht 1 (30 minuten)
- Opdracht 2 (30 minuten)
- Opdracht 3 (30 minuten)
- Opdracht 4 (30 minuten)
- Opdracht 5 (10 minuten)
- Opdracht 6 (30 minuten)

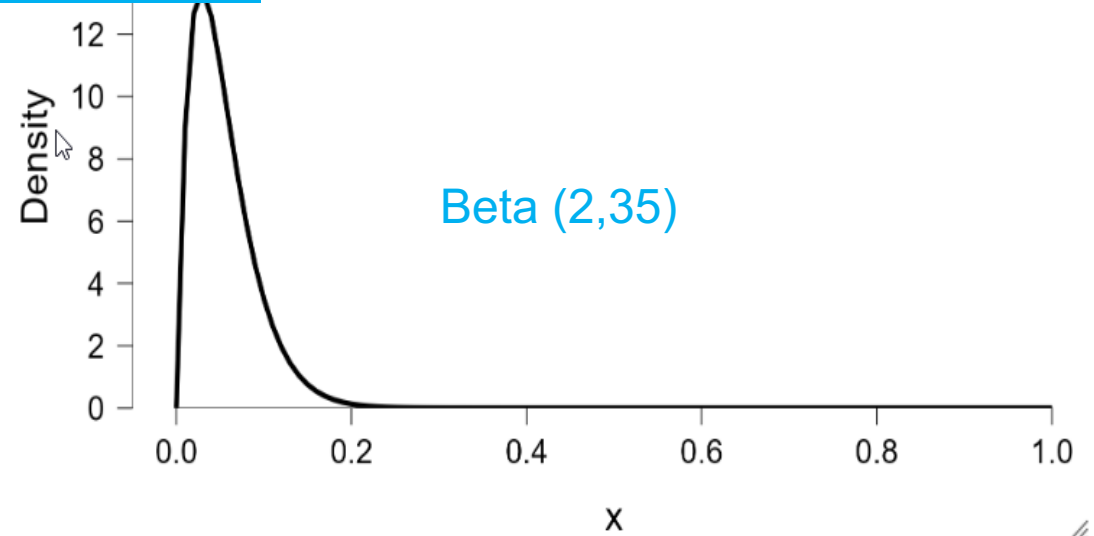
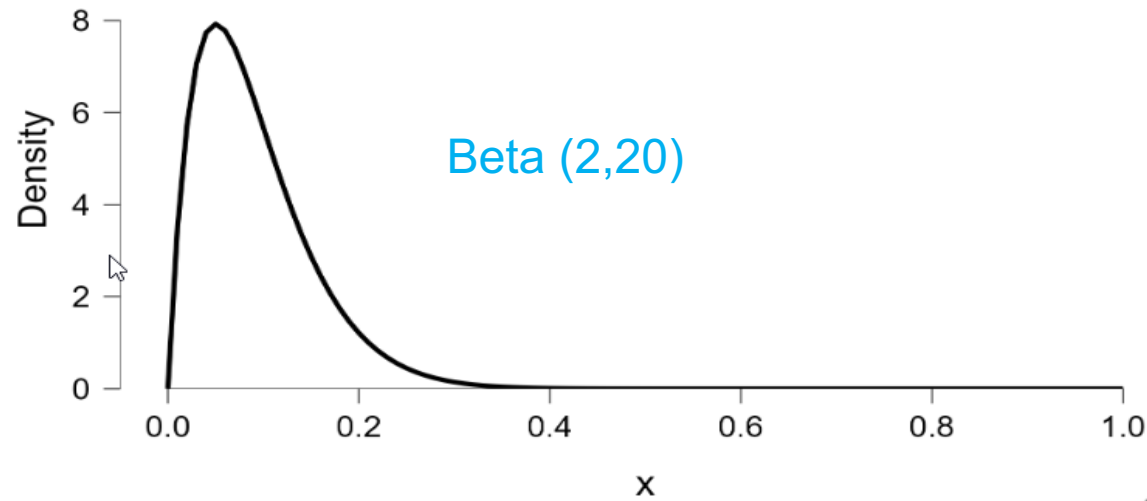
Deel 2:

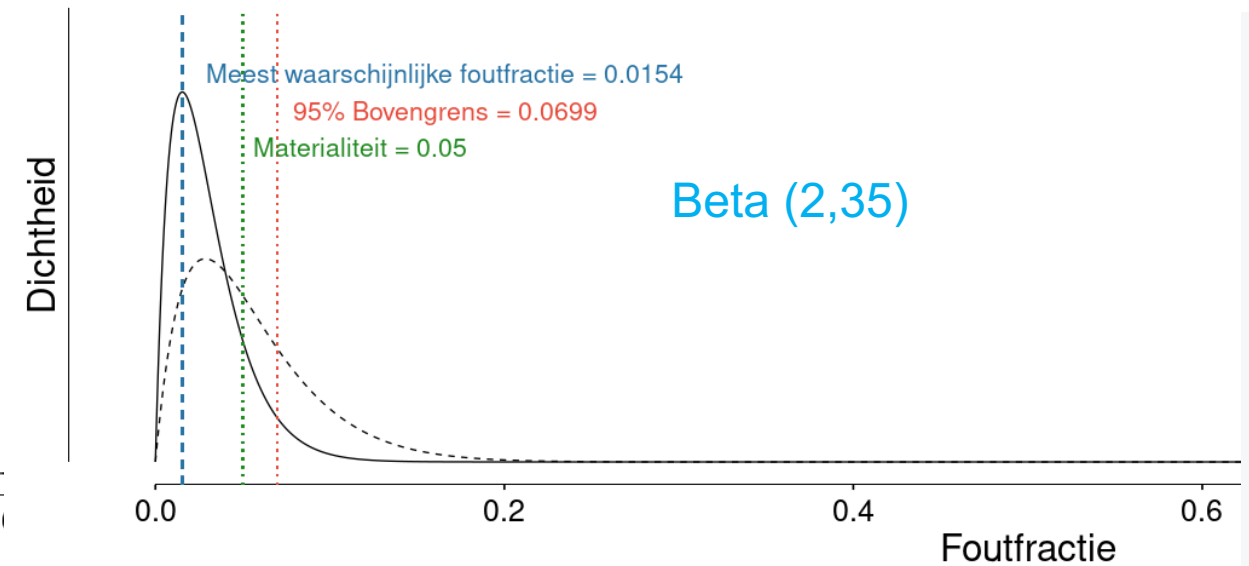
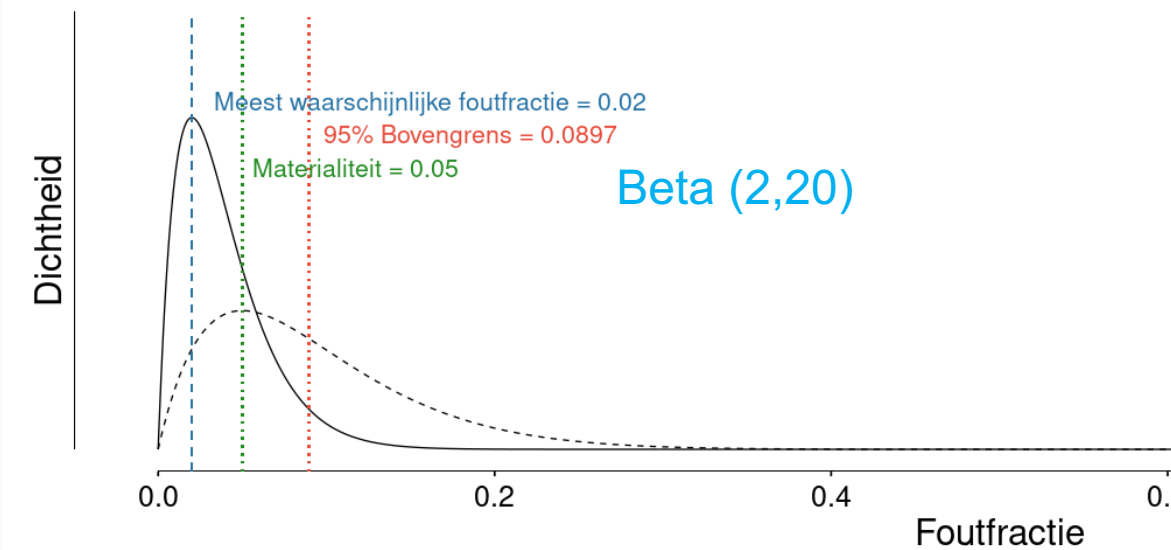
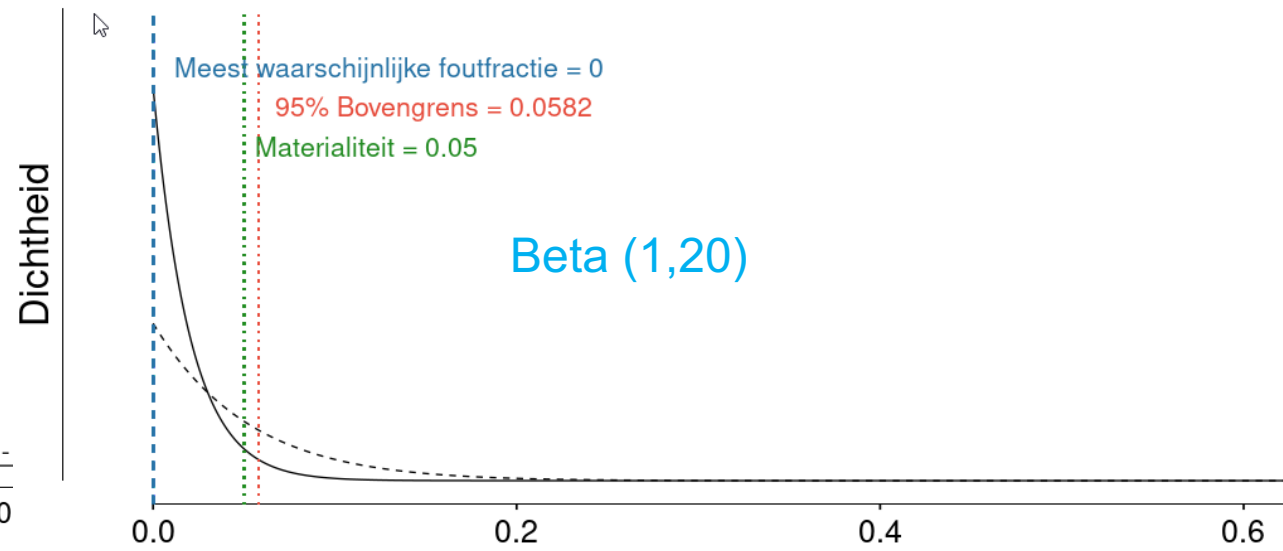
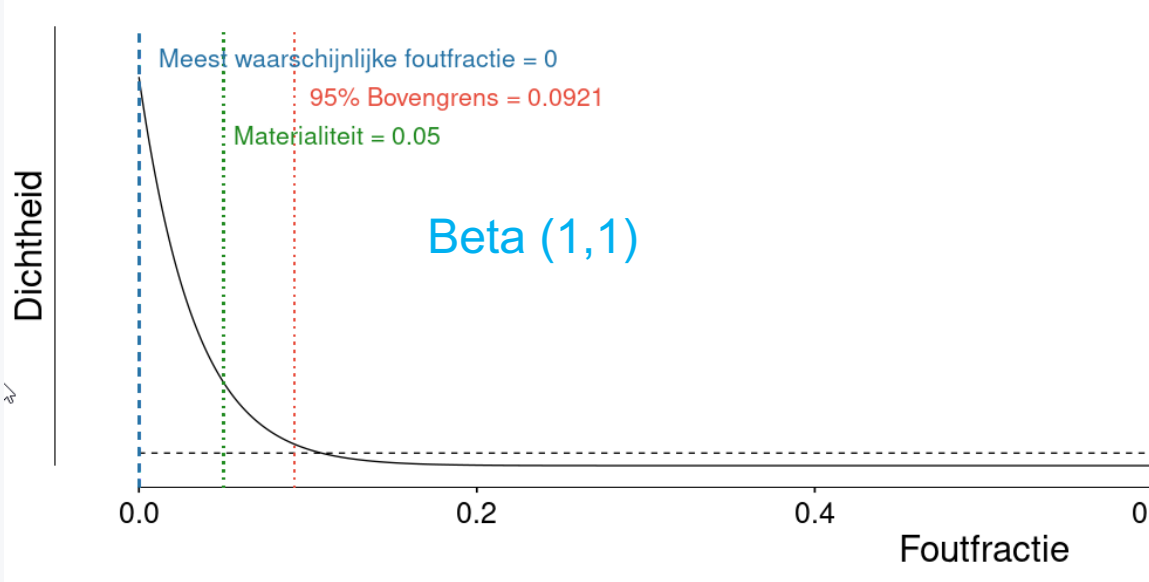
Meenemen van positieve voorinformatie

Of: Wat is het effect van de prior-verdeling?



Wat gebeurt er als we een steekproef nemen van 30 en we vinden geen fouten?





Zelf aan de slag met de BayesDemo applicatie

- Opdracht 1 (30 minuten)
- **Opdracht 2 (30 minuten)**
- Opdracht 3 (30 minuten)
- Opdracht 4 (30 minuten)
- Opdracht 5 (10 minuten)
- Opdracht 6 (30 minuten)

Intermezzo: Terugkoppeling

Wat vond je van het vorige deel?

Deel 3:

Prior-verdeling op basis van data-analyse

Regressie-analyse

Als de data-analyse te relateren is aan de foutfractie, dan kan die als voorinformatie worden gebruikt

Resultaten van data analyse (**CAB § 6.5.2.1**)

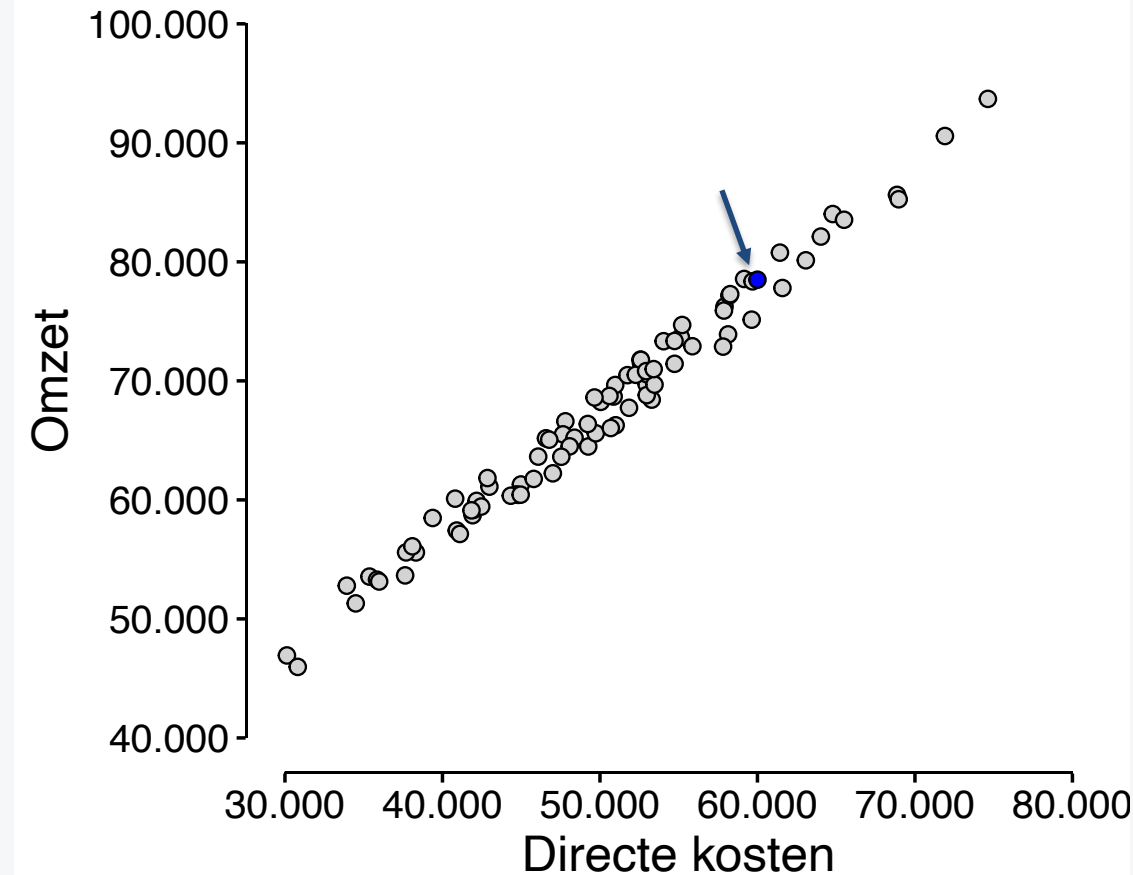
Data-analyse geeft op zich geen redelijke zekerheid over de mate waarin de populatie vrij is van materiële fouten.

populatie vooraf op te schonen. Data-analyse kan ook positieve voorinformatie leveren die aanleiding geeft om de gegevensgerichte werkzaamheden in de statistische steekproef te reduceren.



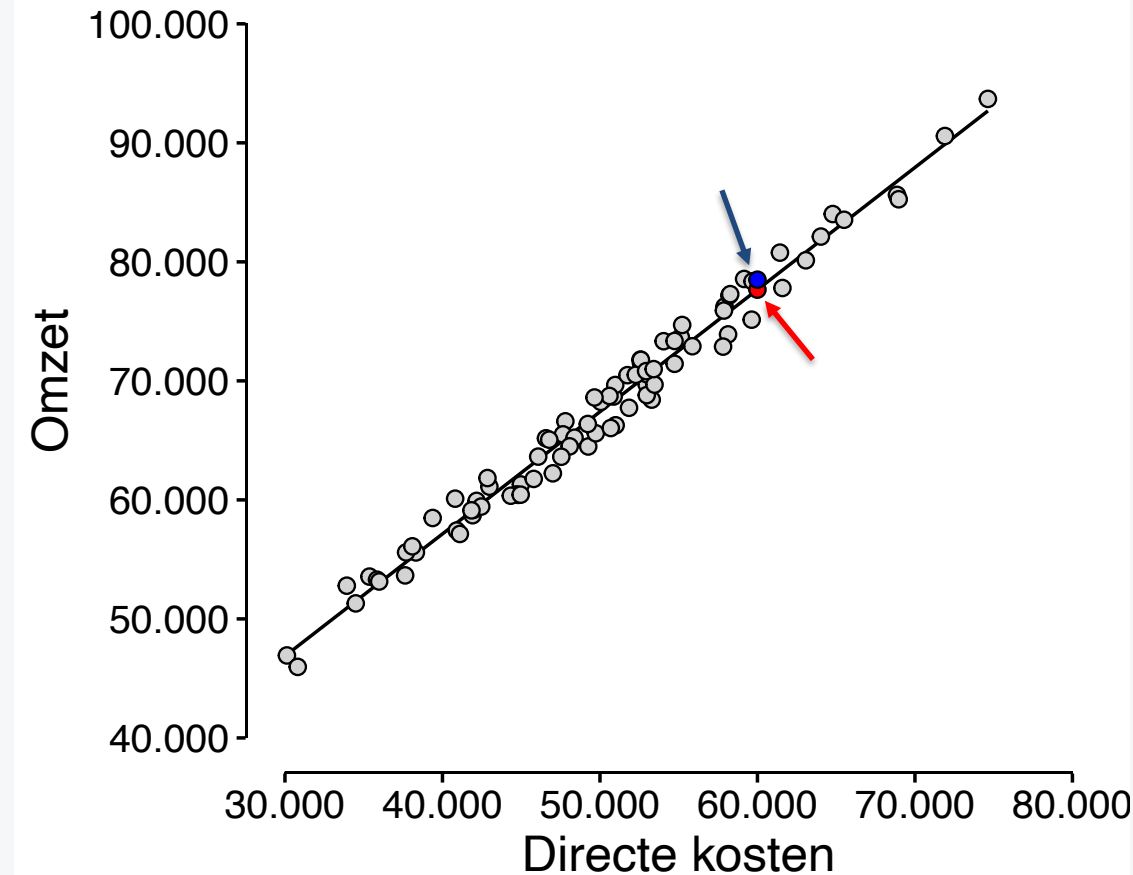
Regressie: Voorspellen van omzet ovb kosten

- Auditor kent omzet ● en directe kosten van auditee (€ 78.500 en € 60.000)
- Kent ook omzet en directe kosten van 100 soortgelijke bedrijven ○



Regressie: Voorspellen van omzet ovb kosten

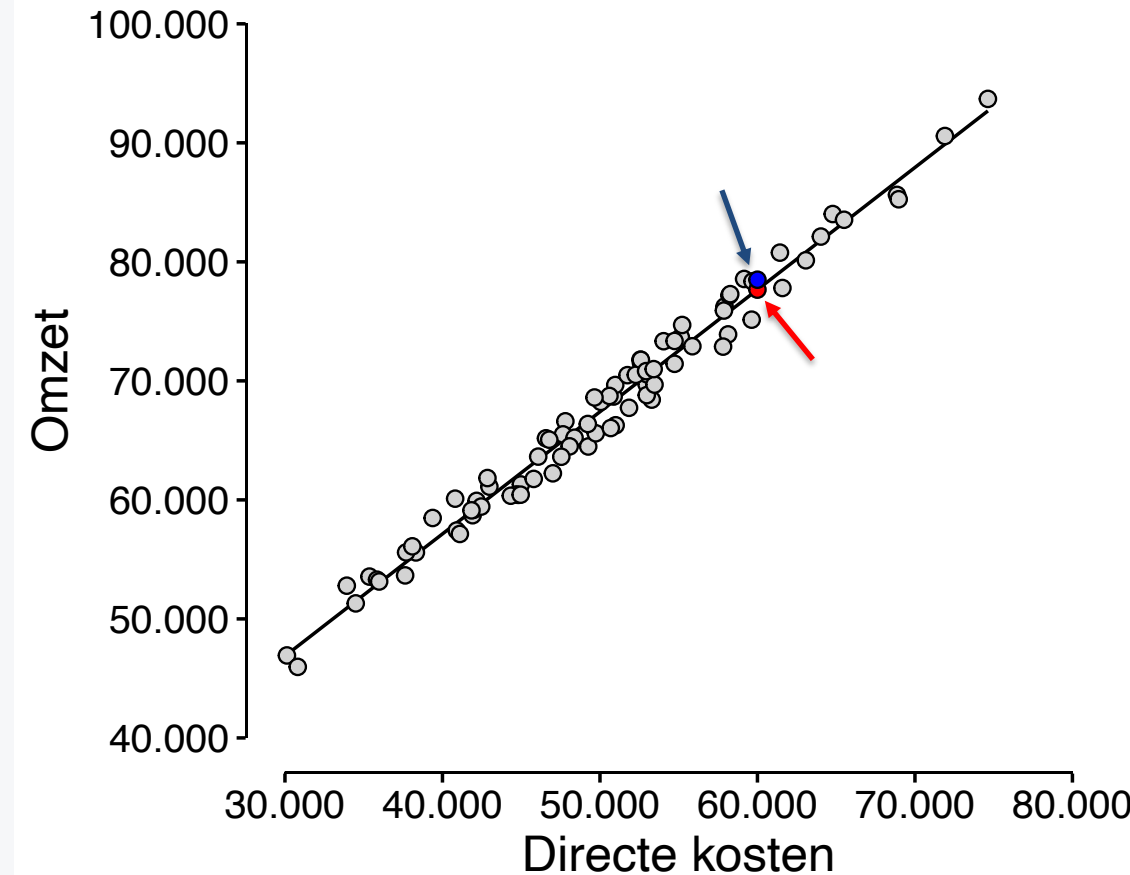
- Auditor kent omzet  en directe kosten van auditee (€ 78.500 en € 60.000)
- Kent ook omzet en directe kosten van 100 soortgelijke bedrijven 
- Past lineaire regressie toe om verband tussen directe kosten en omzet te modelleren en omzet te voorspellen 



Regressie: Voorspellen van omzet ovb kosten

- Auditor kent omzet ● en directe kosten van auditee (€ 78.500 en € 60.000)
- Kent ook omzet en directe kosten van 100 soortgelijke bedrijven ○
- Past lineaire regressie toe om verband tussen directe kosten en omzet te modelleren en omzet te voorspellen ●

$$\text{Omzet} = \beta_0 + \beta_1 * \text{Kosten} + \epsilon$$

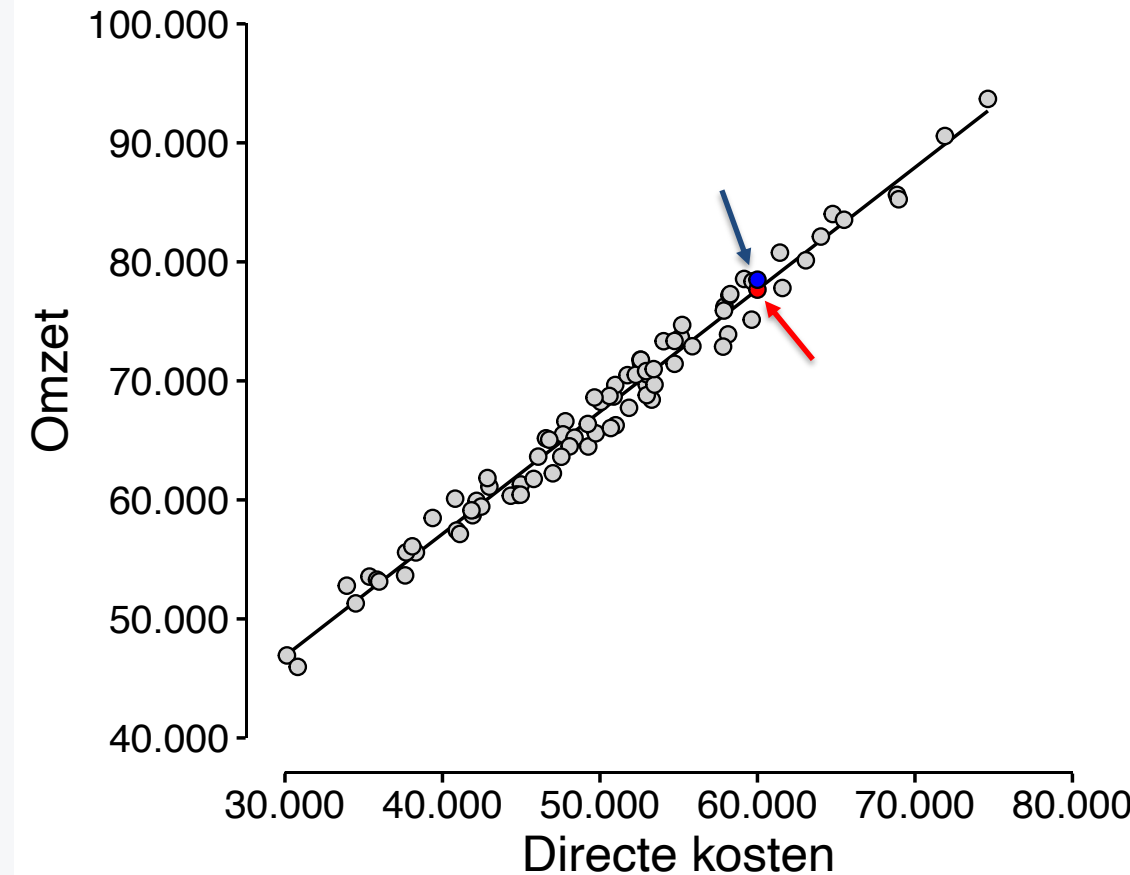


Regressie: Voorspellen van omzet ovb kosten

- Auditor kent omzet ● en directe kosten van auditee (€ 78.500 en € 60.000)
- Kent ook omzet en directe kosten van 100 soortgelijke bedrijven ●
- Past lineaire regressie toe om verband tussen directe kosten en omzet te modelleren en omzet te voorspellen ●

$$\text{Omzet} = \beta_0 + \beta_1 * \text{Kosten} + \epsilon$$

$$\text{Omzet} = 16.089,27 + 1,026 * \text{Kosten}$$



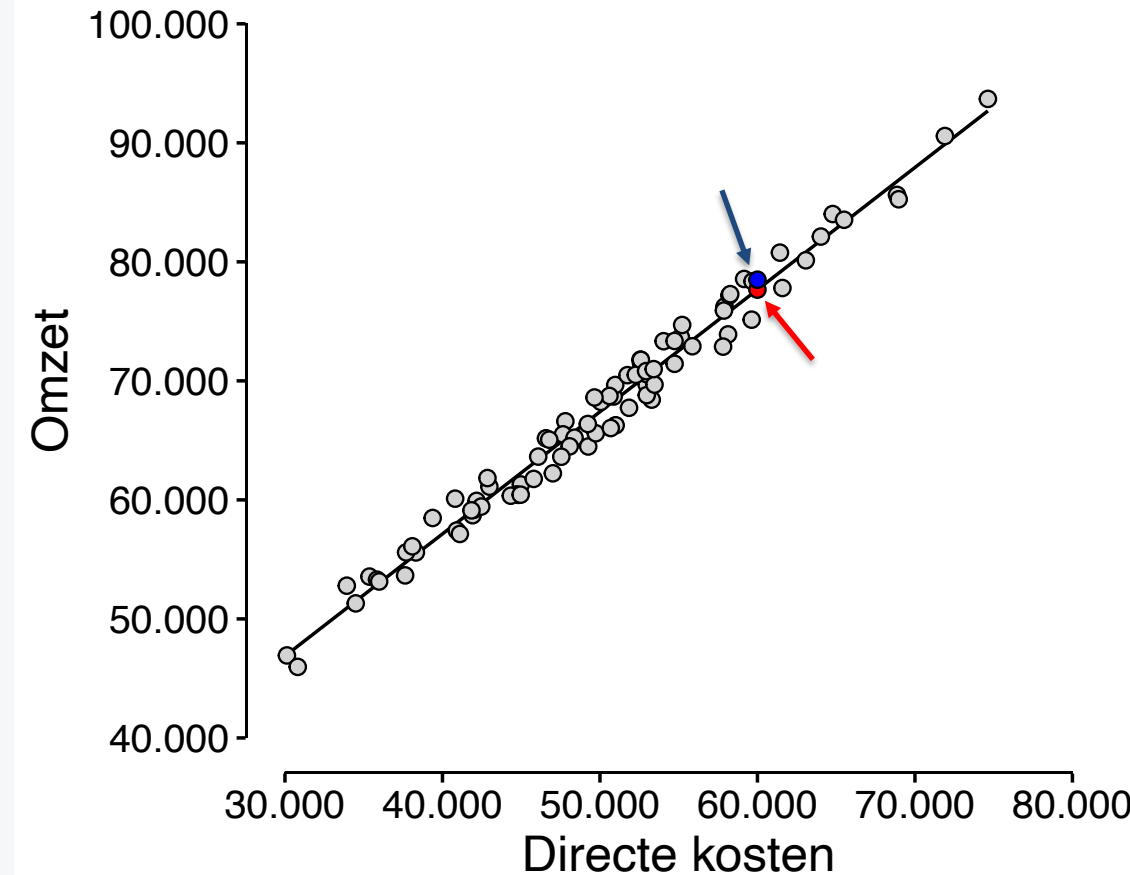
Regressie: Voorspellen van omzet ovb kosten

- Auditor kent omzet ● en directe kosten van auditee (€ 78.500 en € 60.000)
- Kent ook omzet en directe kosten van 100 soortgelijke bedrijven ●
- Past lineaire regressie toe om verband tussen directe kosten en omzet te modelleren en omzet te voorspellen ●

$$\text{Omzet} = \beta_0 + \beta_1 * \text{Kosten} + \epsilon$$

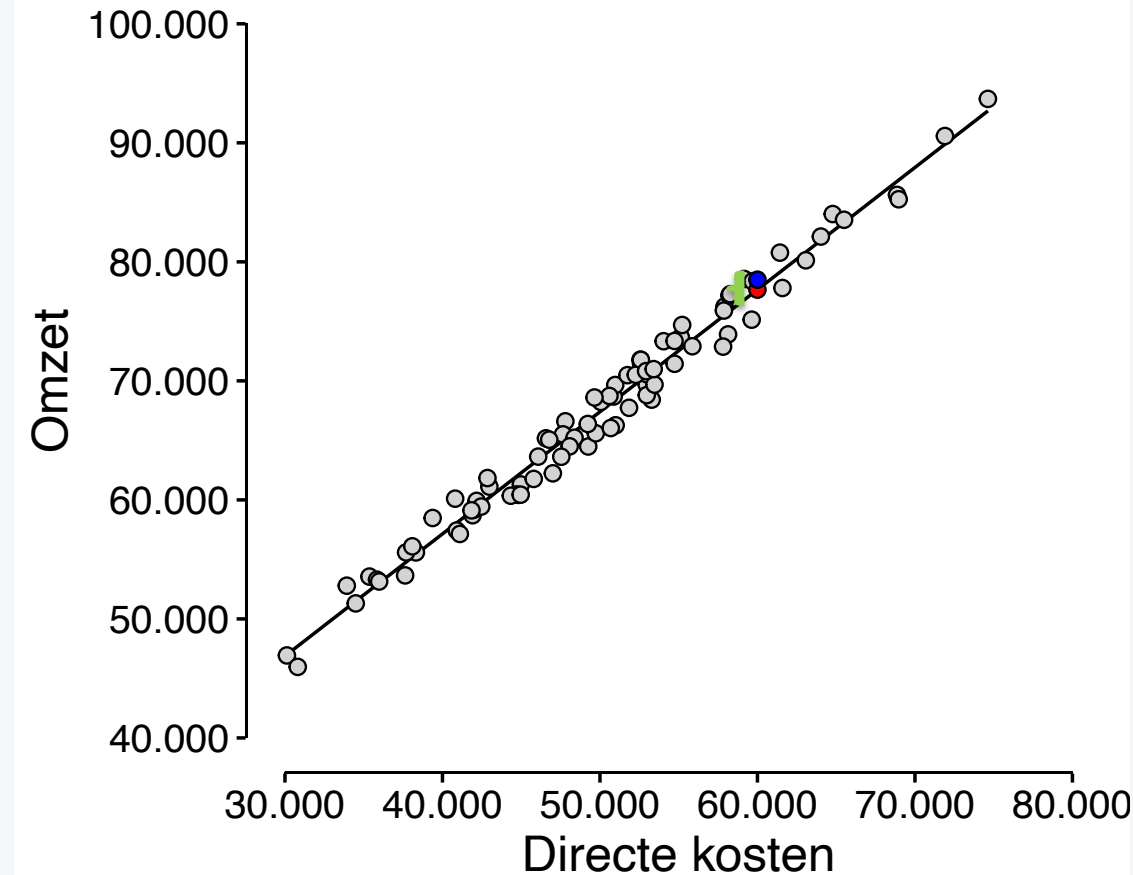
$$\text{Omzet} = 16.089,27 + 1,026 * \text{Kosten}$$

$$\text{● } 77.649,27 = 16.089,27 + 1,026 * 60.000$$

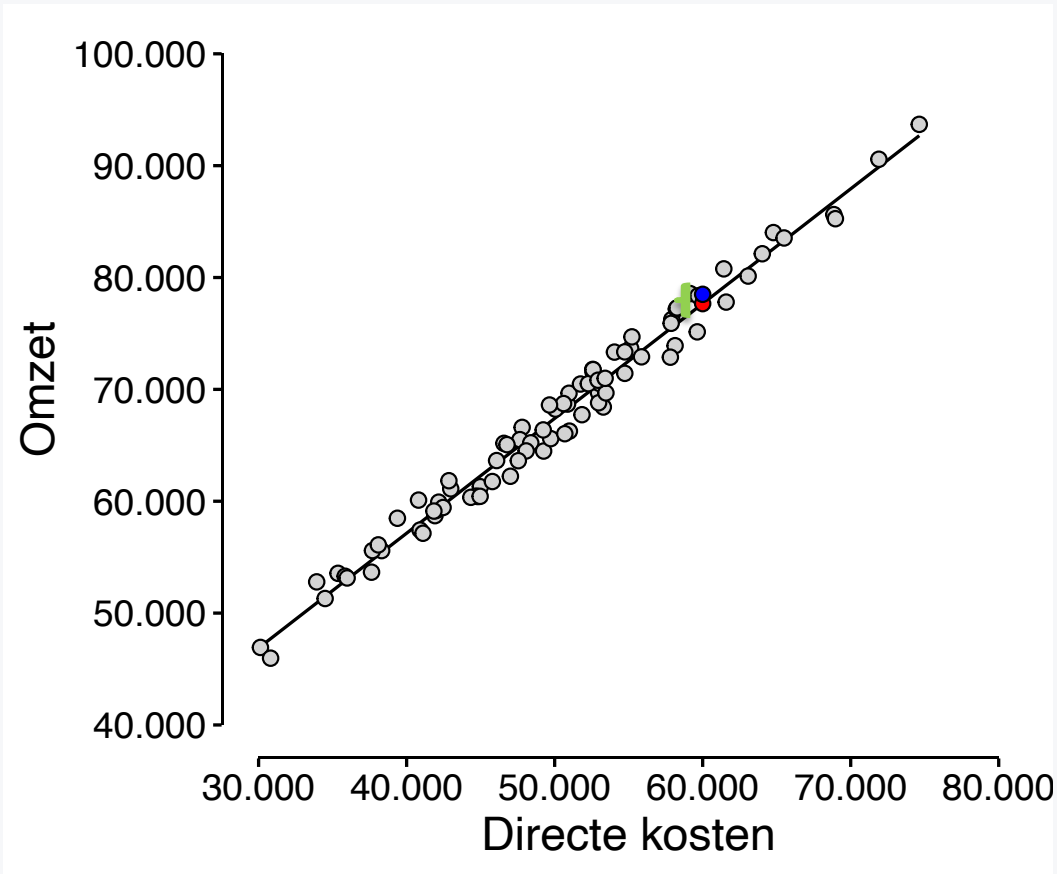


Regressie: Voorspellen van omzet ovb kosten

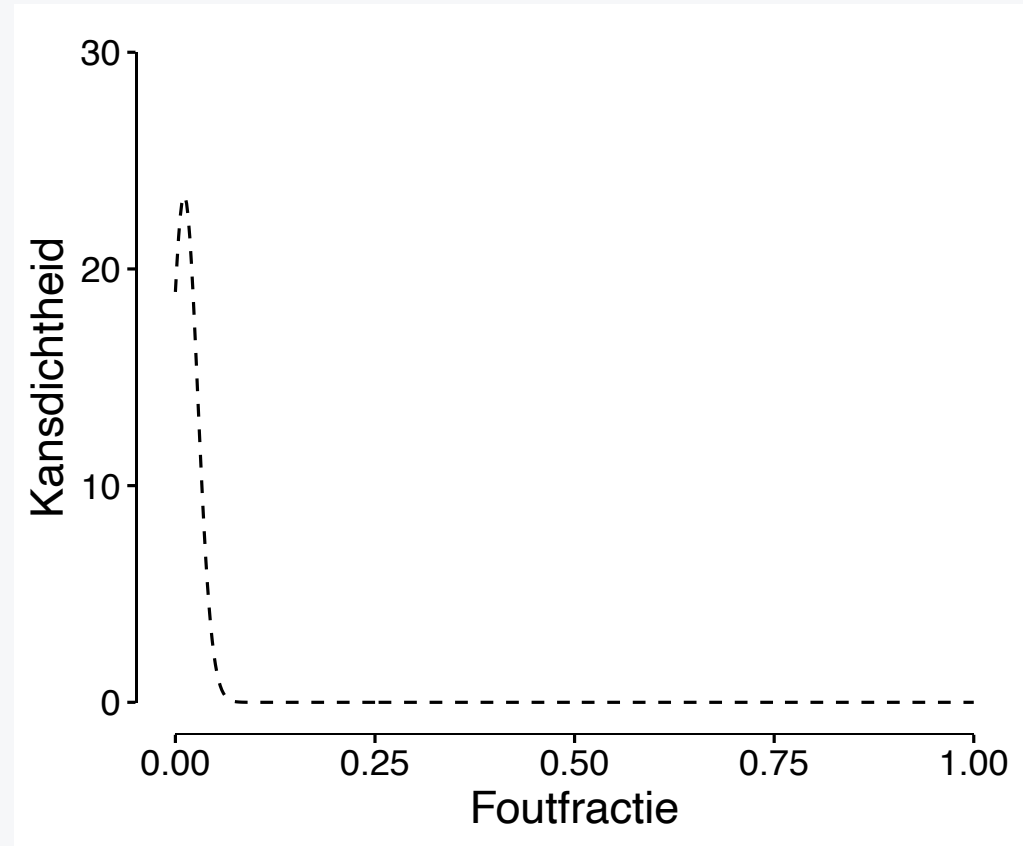
- Auditor kent omzet ● en directe kosten van auditee (€ 78.500 en € 60.000)
- Kent ook omzet en directe kosten van 100 soortgelijke bedrijven ○
- Past lineaire regressie toe om verband tussen directe kosten en omzet te modelleren en omzet te voorspellen ●
- Afwijking { tussen voorspelde ● en opgevoerde ● omzet geeft een indicatie van mogelijke overstatement in omzet



Een relatief kleine afwijking van de voorspelde omzet kan gezien worden als positieve voorinformatie



Vertaling



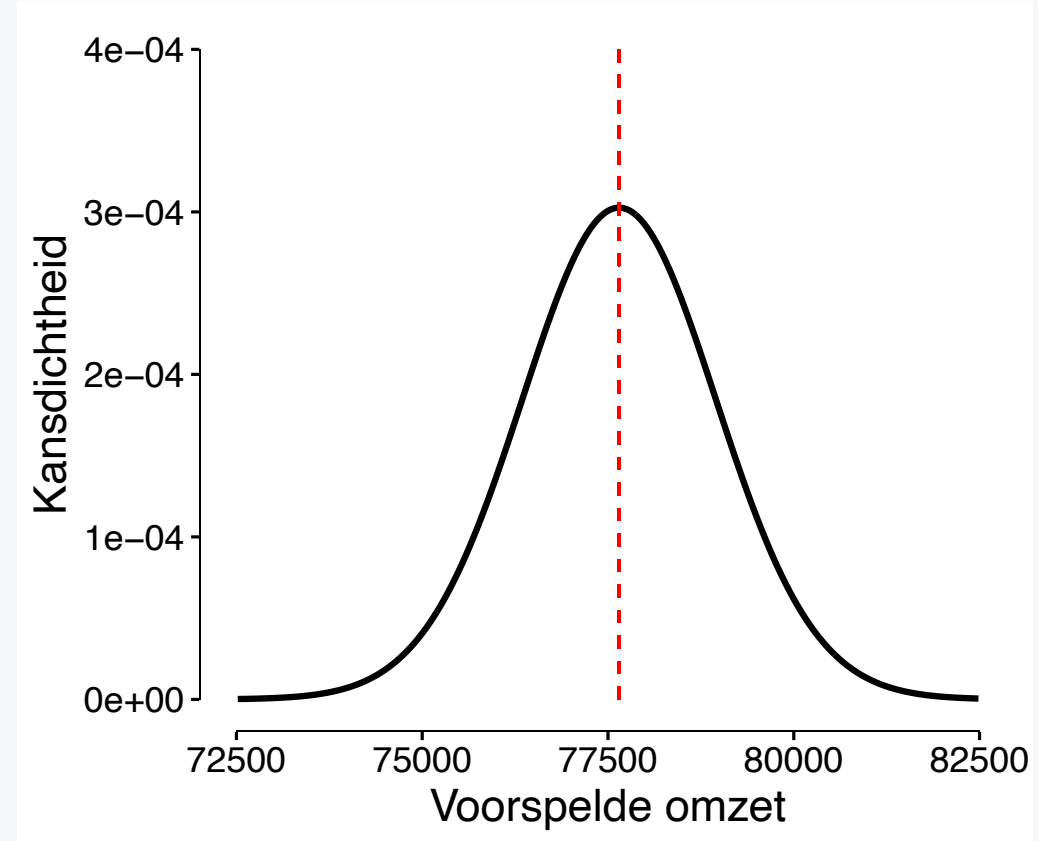
Rekenvoorbeeld

Voorspelde omzet volgt t-verdeling:

Gemiddelde (μ) = 77.649,27 ●

Standaardafwijking (σ) = 1.314,37

Vrijheidsgraden = 83



Student-t(77649,27; 1314,37; 83) 33

Rekenvoorbeeld

Voorspelde omzet volgt t-verdeling:

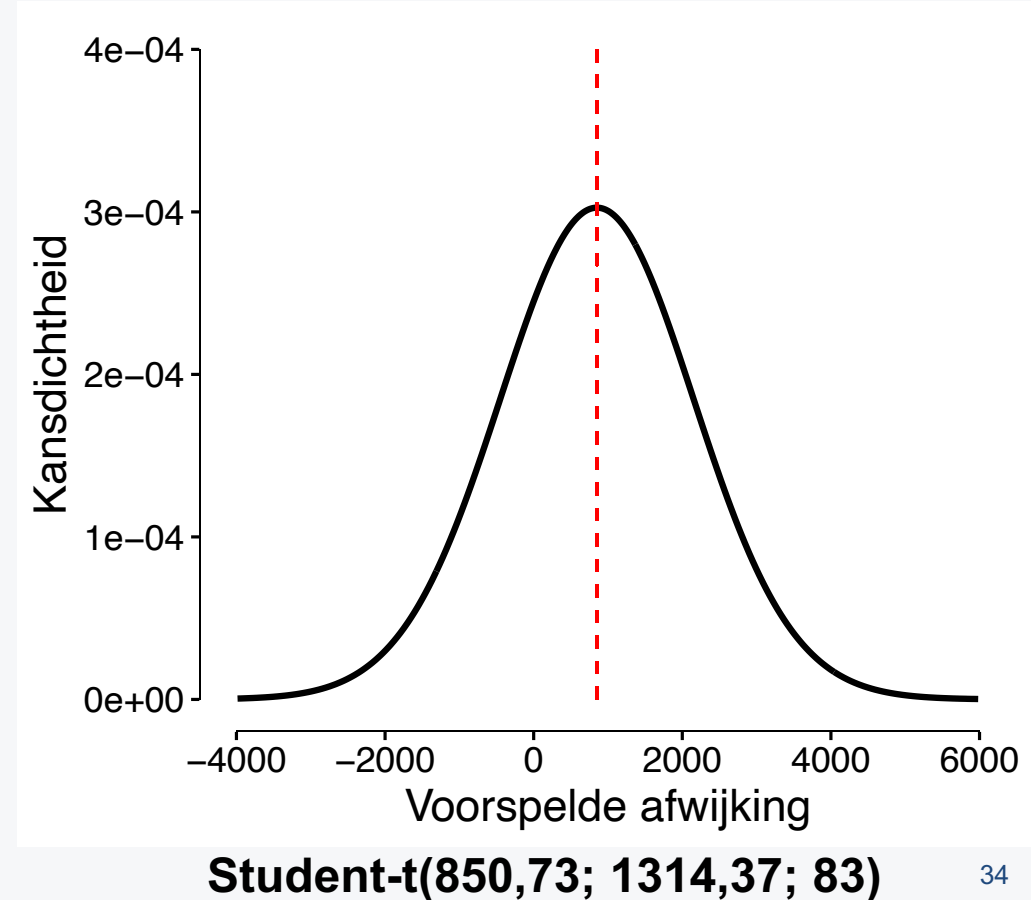
Gemiddelde (μ) = 77.649,27 ●

Standaardafwijking (σ) = 1.314,37

Vrijheidsgraden = 83

Voorspelde afwijking (overstatement):

$$\mu = 78.500 \text{ ●} - 77.649,27 \text{ ●} = 850,73 \text{ }$$



Rekenvoorbeeld

Voorspelde omzet volgt t-verdeling:

Gemiddelde (μ) = 77.649,27 ●

Standaardafwijking (σ) = 1.314,37

Vrijheidsgraden = 83

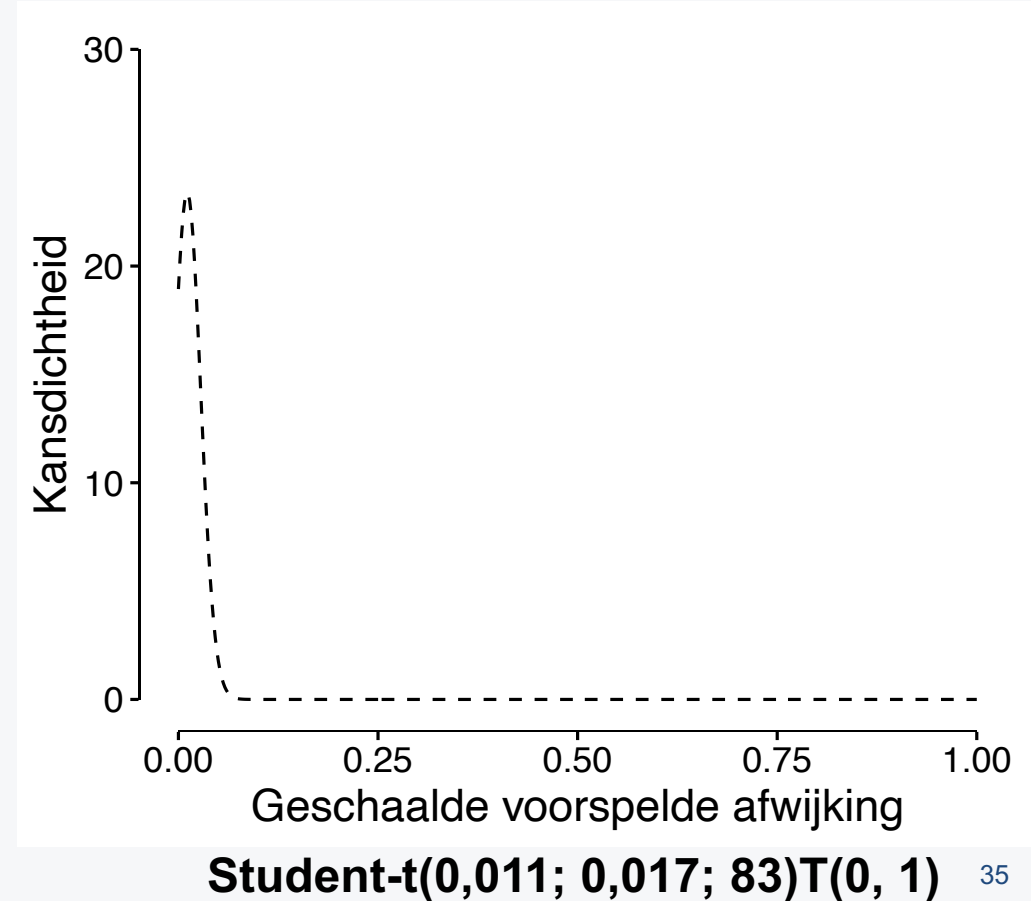
Voorspelde afwijking (overstatement): :

$$\mu = 78.500 \text{ ●} - 77.649,27 \text{ ●} = 850,73 \text{ }$$

Geschaalde voorspelde afwijking:

$$\mu = 850,73 / 78.500 \text{ ●} = 0,011$$

$$\sigma = 1.314,37 / 78.500 \text{ ●} = 0,017$$



De prior obv regressie reduceert de steekproef

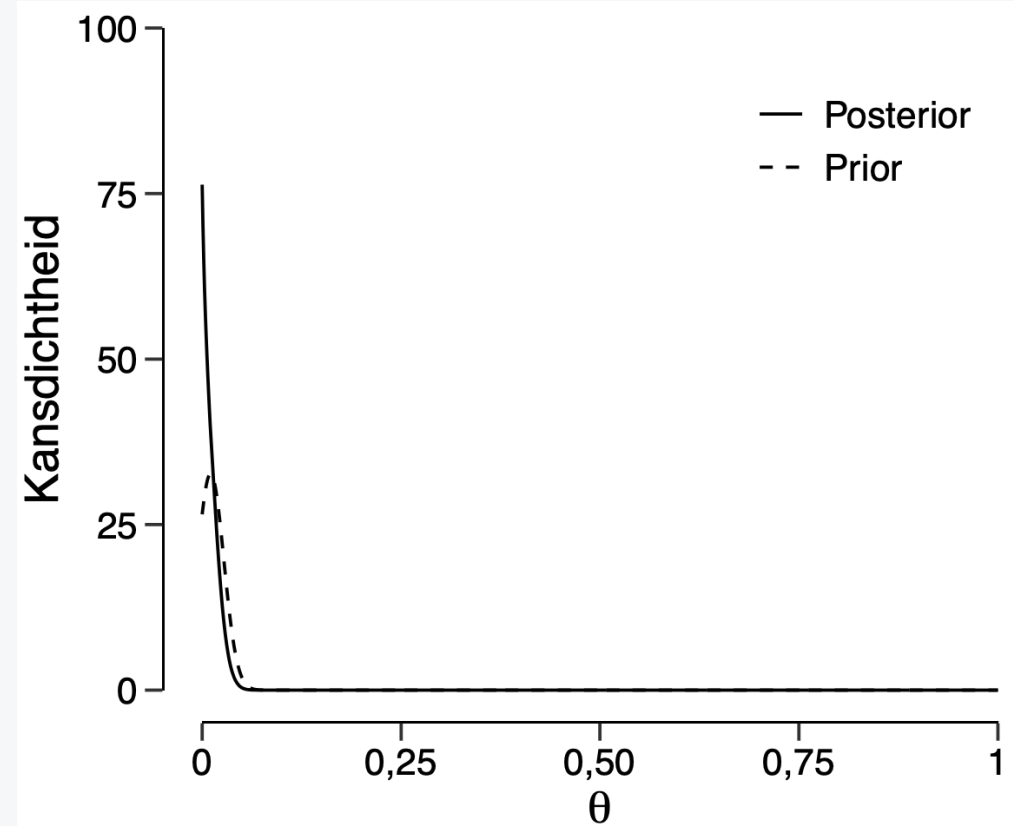
Steekproef:

53 ipv 80 geldeenheden nodig om 95% bovengrens onder 3,75% te krijgen.

Tabel 1. Planningssamenvatting ▼

	Waarde
Uitvoeringsmaterialiteit	0.0375
Auditrisico	0.0500
Toelaatbare fouten	0.0000
Minimale steekproefomvang ^a	53

Hoe veel nodig als er ook nog gesteund kan worden op de interne beheersing?



Zelf aan de slag met de BayesDemo applicatie

- Opdracht 1 (30 minuten)
- Opdracht 2 (30 minuten)
- **Opdracht 3 (30 minuten)**
- Opdracht 4 (30 minuten)
- Opdracht 5 (10 minuten)
- Opdracht 6 (30 minuten)

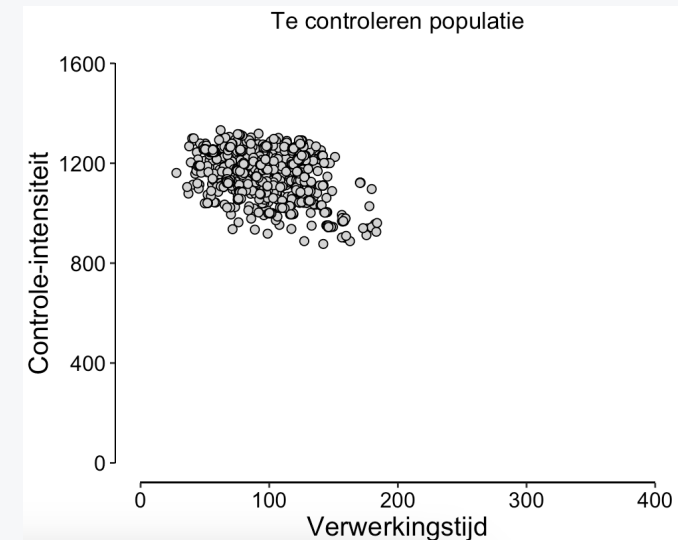
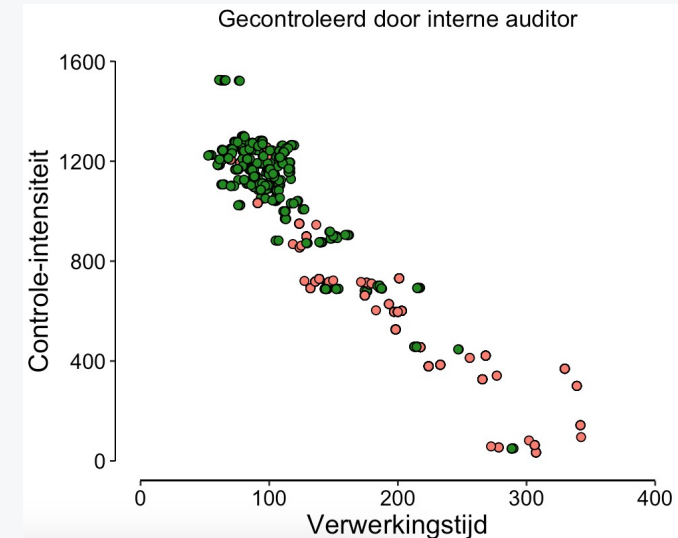
Deel 4:

Prior-verdeling op basis van data-analyse

Classificatie-analyse

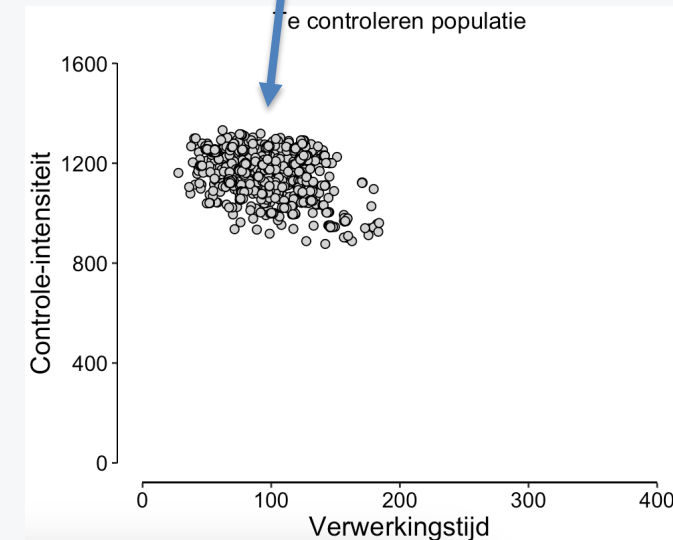
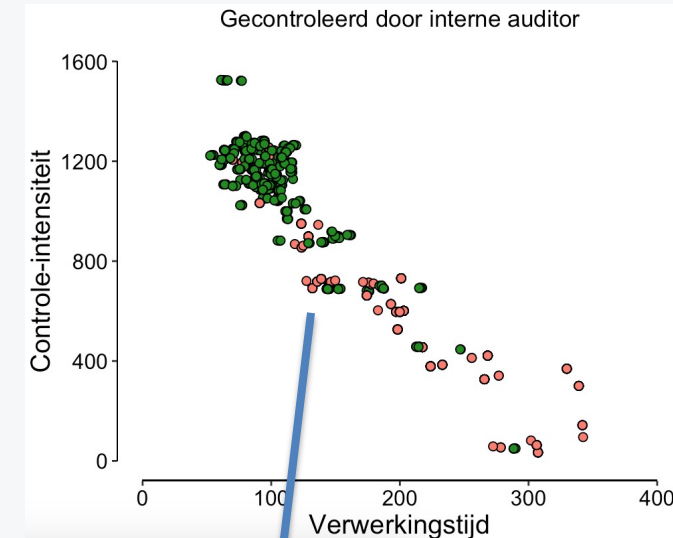
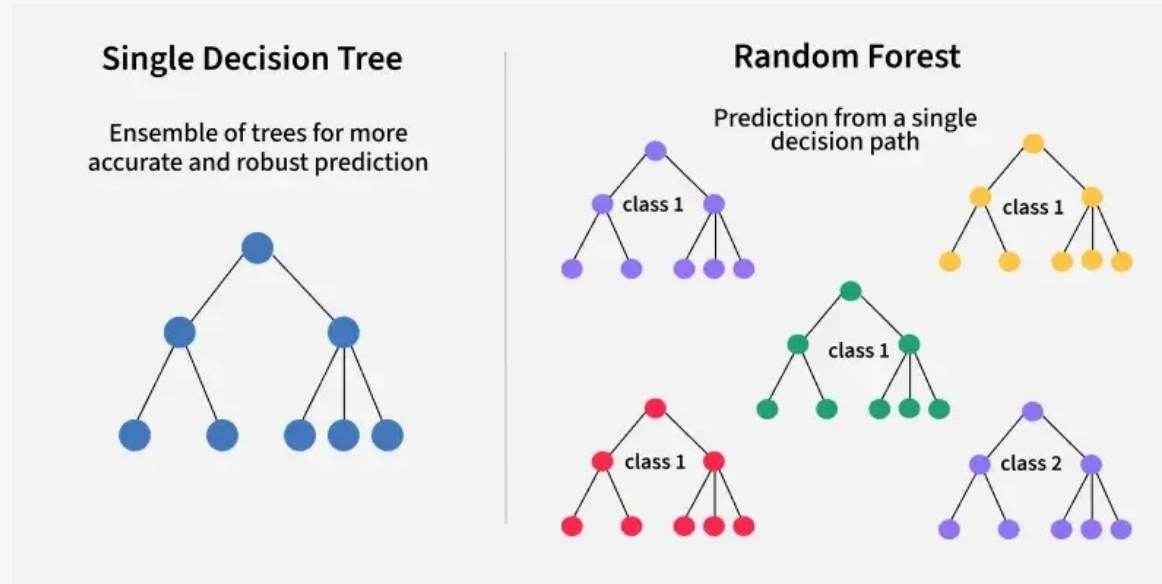
Classificatie: Voorspellen foute facturen

- Een interne auditor heeft wekelijks 10 inkoopfacturen gecontroleerd op fouten, zocht expliciet naar fouten
- Gegevens over controle-intensiteit en verwerkingstijd zijn beschikbaar voor alle inkoopfacturen
- Auditor wil voortbouwen op het werk van interne auditor
- Beschikbare data: 520 al gecontroleerde inkoopfacturen
- Nog niet gecontroleerd: 1960 facturen



Classificatie: Voorspellen foute facturen

- Doel classificatiemodel: Voorspellen 'fout' of 'niet fout' voor elke ongeziene post o.b.v. bekende gegevens
- In dit geval een '*random forest*' model gebruikt

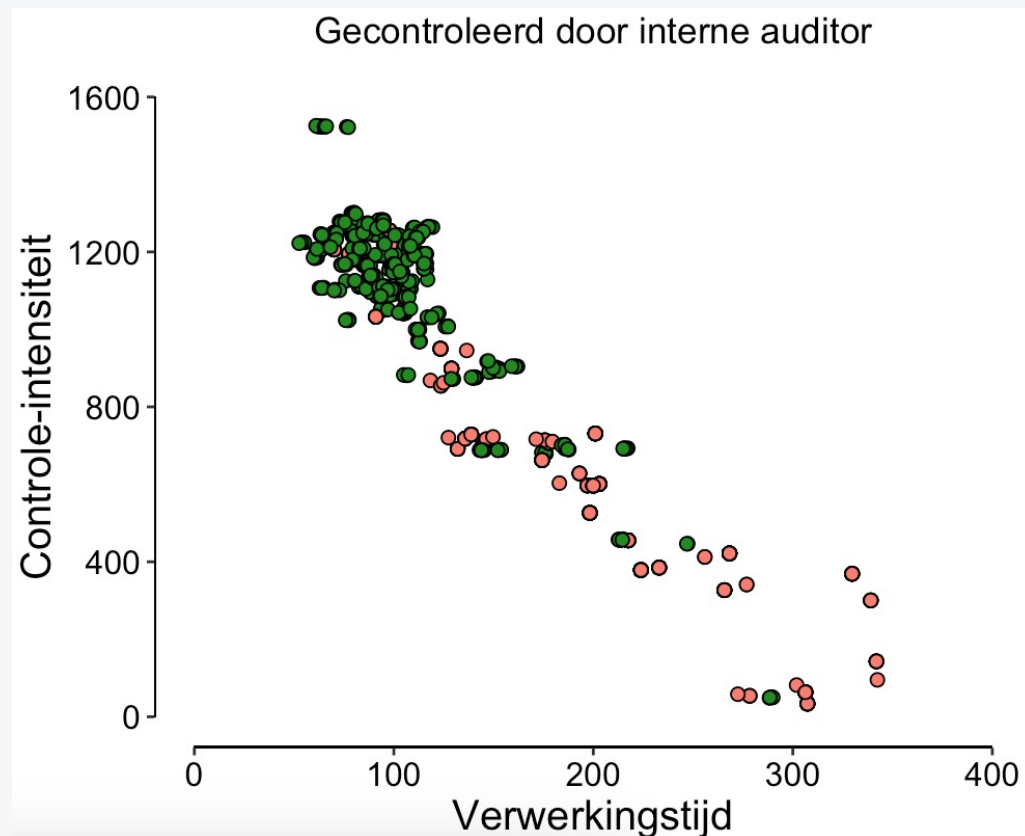


Data worden verdeeld in train- en testset

Train: 420

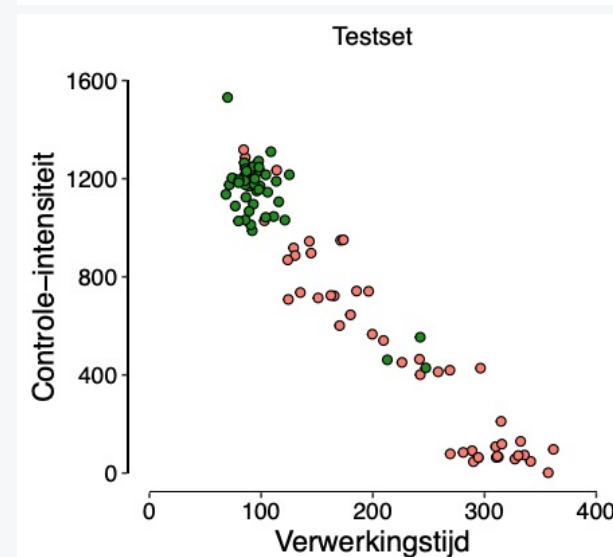
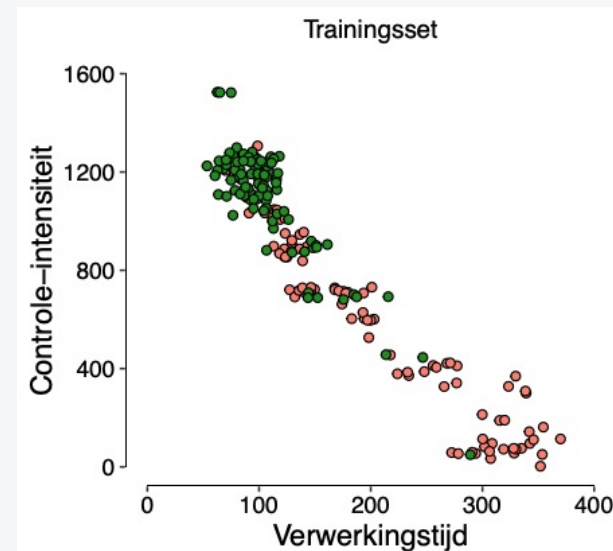
Test: 100

Totaal: 520



80 % = 420

20 % = 100

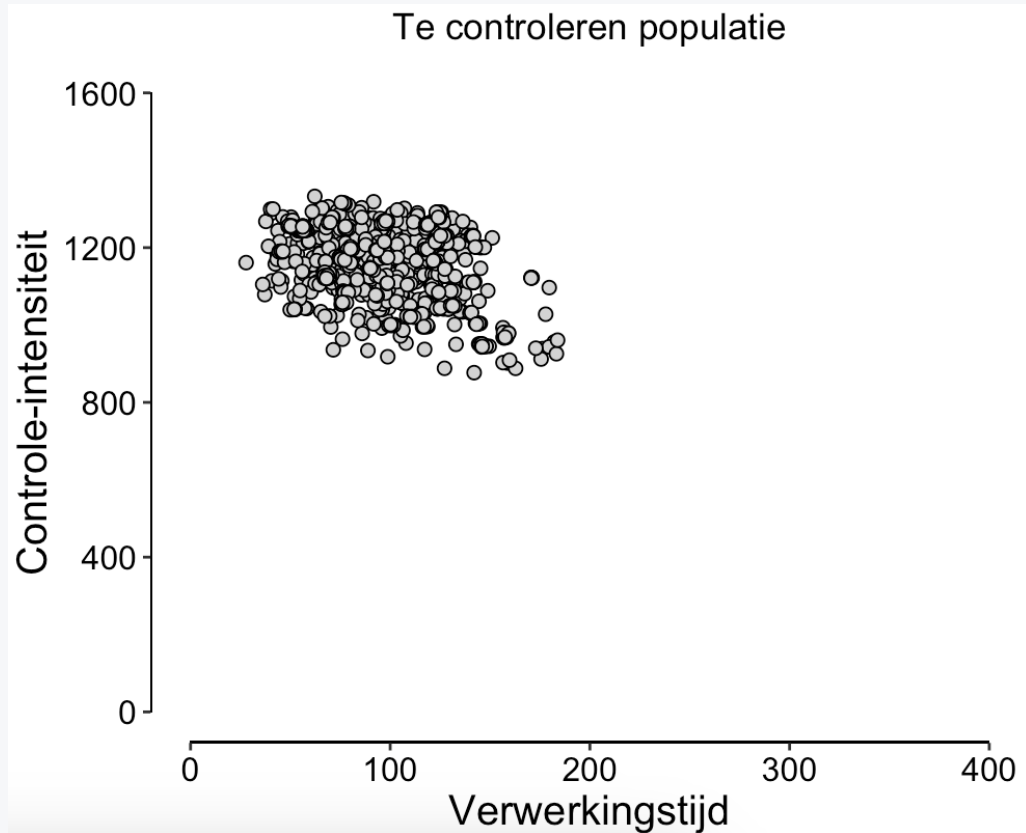


Model wordt geëvalueerd met de verwarringsmatrix

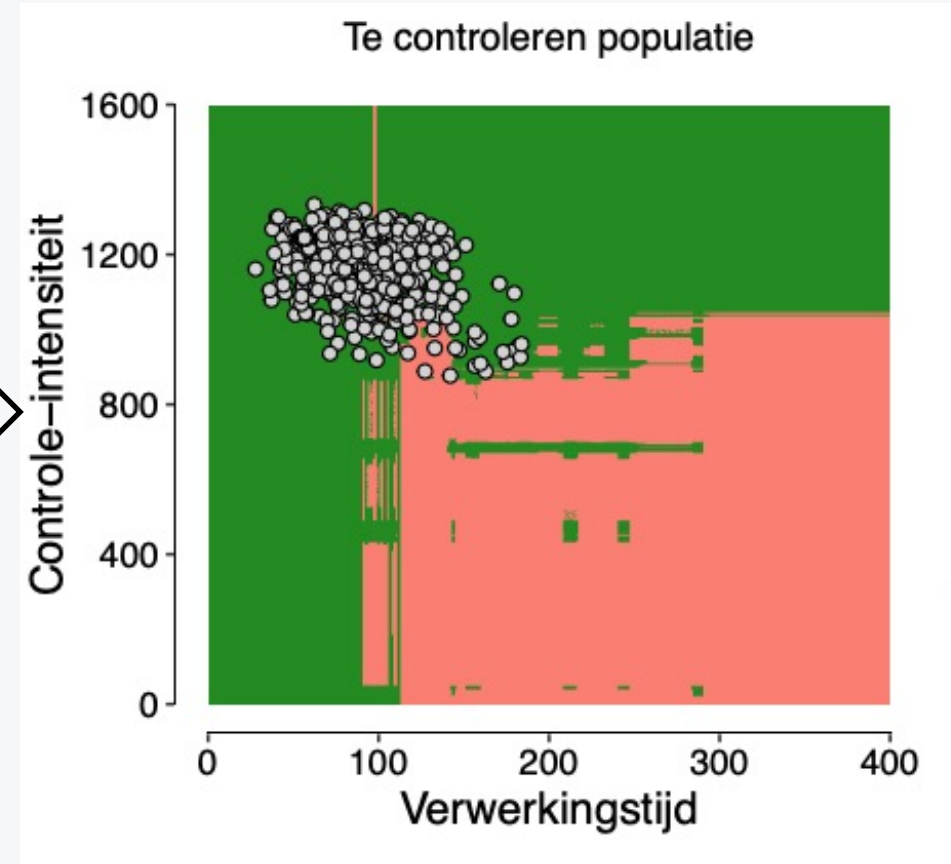
		Werkelijk	
		Fout	Niet fout
Voorspeld	Fout	11 (TP)	1 (FP)
	Niet fout	4 (FN)	84 (TN)

Bijv. nauwkeurigheid (*accuracy*) = $(TP+TN) / (TP+FN+TN+FP)$
 $(11 + 84) / (11 + 4 + 84 + 1) = 0,95$

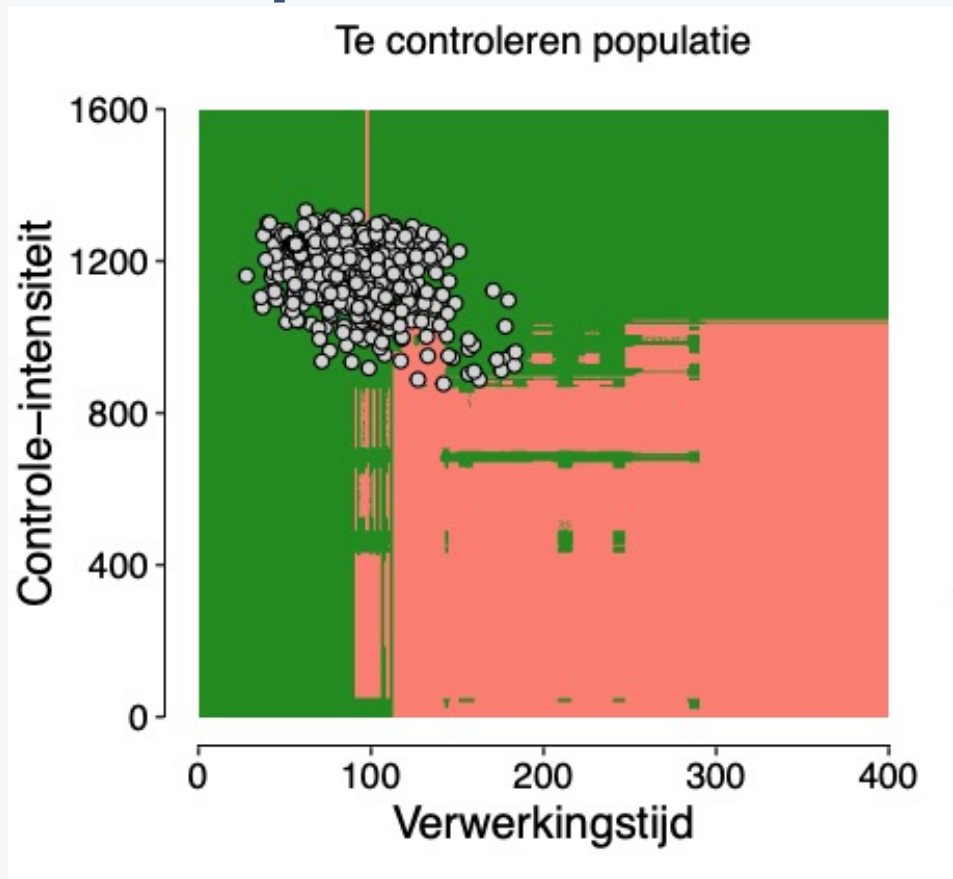
Getraind model maakt voorspelling voor populatie



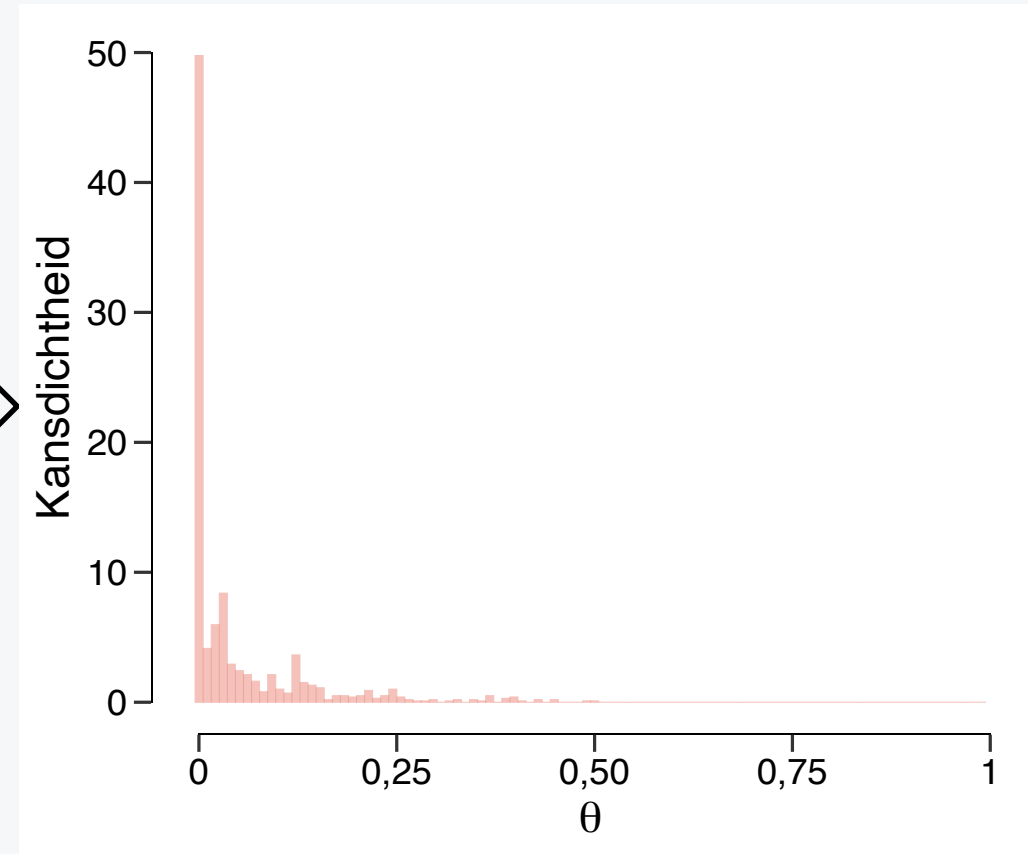
Voorspellen



Relatief kleine voorspelde foutkansen kan gezien worden als positieve voorinformatie



Kansuitspraak



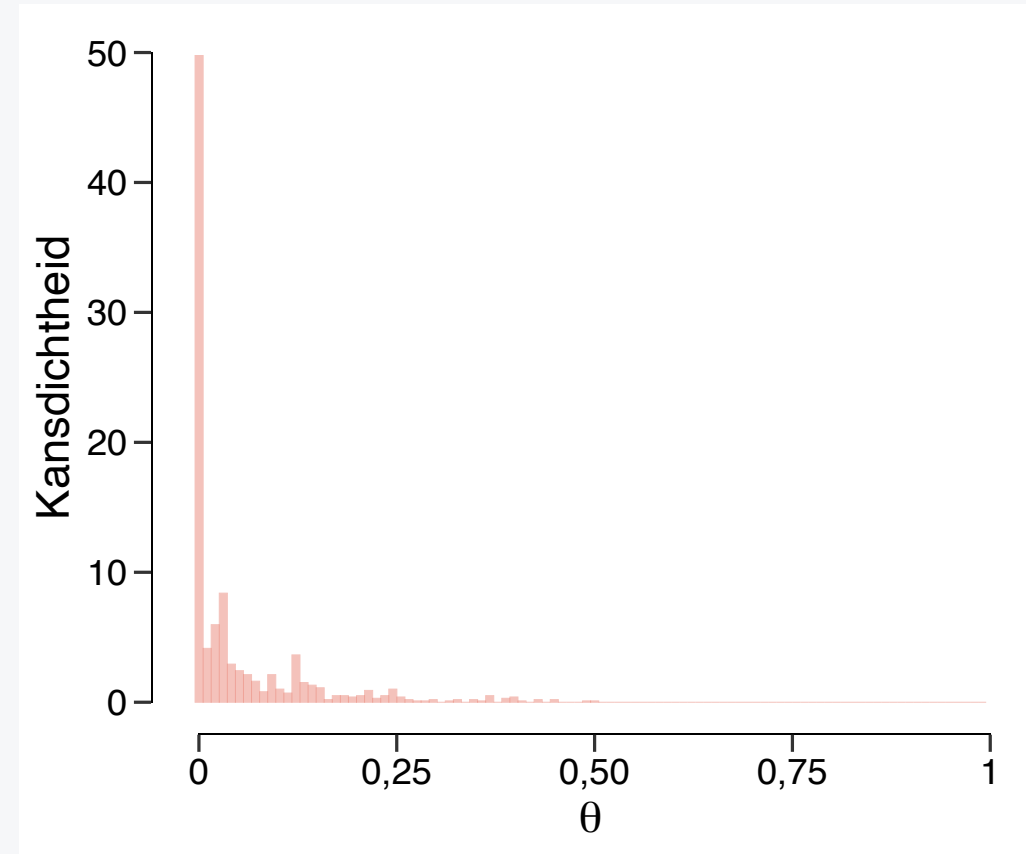
Rekenvoorbeeld

Voorspelde kansen:

Gemiddelde ($\bar{x}$) = 0,047

Variantie (s^2) = 0,007

Vind de beta-verdeling met gemiddelde $\bar{x}$ en variantie s^2 die op de foutkansen past.



Rekenvoorbeeld

Voorspelde kansen:

Gemiddelde ($\bar{x}$) = 0,047

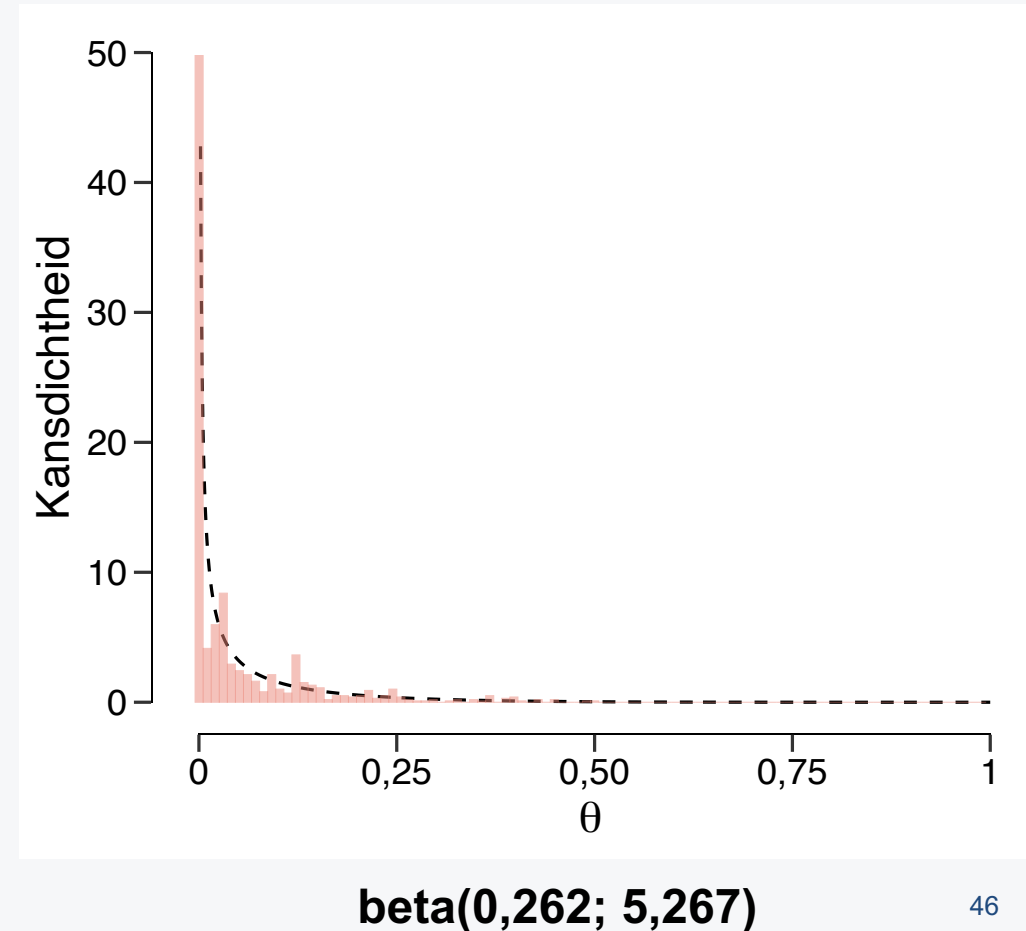
Variantie (s^2) = 0,007

Vind de beta-verdeling met gemiddelde $\bar{x}$ en variantie s^2 die op de foutkansen past.

Parameters voor de prior-verdeling:

$$\alpha = \left(\frac{\bar{x} (1 - \bar{x})}{s^2} - 1 \right) \bar{x} = 0,262$$

$$\beta = \left(\frac{\bar{x} (1 - \bar{x})}{s^2} - 1 \right) (1 - \bar{x}) = 5,267$$



Rekenvoorbeeld

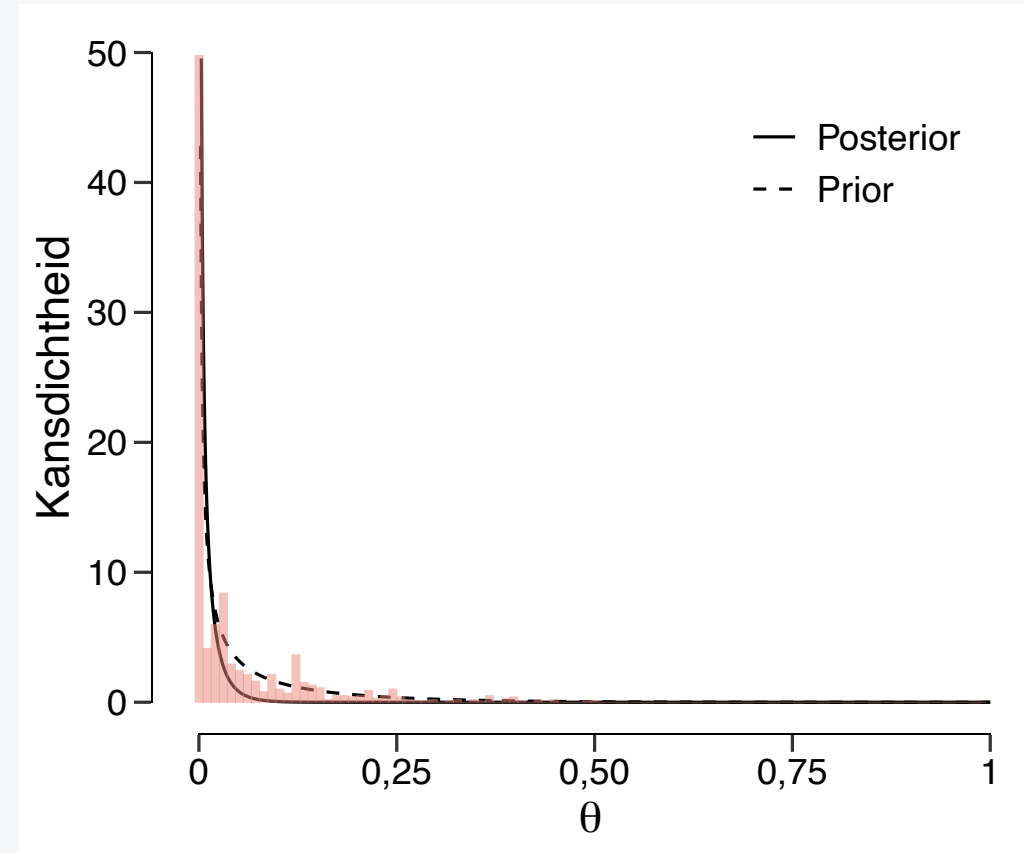
Steekproef:

37 ipv 80 geldeenheden nodig om 95% bovengrens onder 3,75% te krijgen.

Tabel 1. Planningssamenvatting

	Waarde
Uitvoeringsmaterialiteit	0.0375
Auditrisico	0.0500
Toelaatbare fouten	0.0000
Minimale steekproefomvang ^a	37

Hoe veel nodig als er ook nog gesteund kan worden op de interne beheersing?



Zelf aan de slag met de BayesDemo applicatie

- Opdracht 1 (30 minuten)
- Opdracht 2 (30 minuten)
- Opdracht 3 (30 minuten)
- **Opdracht 4 (30 minuten)**
- **Opdracht 5 (10 minuten)**
- Opdracht 6 (30 minuten)

Deel 5:

Vertaling naar Excel

Of: Hoe implementeer ik dit helemaal zelf?

Zelf aan de slag met de BayesDemo applicatie

- Opdracht 1 (30 minuten)
- Opdracht 2 (30 minuten)
- Opdracht 3 (30 minuten)
- Opdracht 4 (30 minuten)
- Opdracht 5 (10 minuten)
- **Opdracht 6 (30 minuten)**

Meer te weten komen? Lees het artikel!

MAB MAANDBLAD VOOR
ACCOUNTANCY EN
BEDRIJFSECONOMIE

Maandblad voor Accountancy en Bedrijfseconomie 98(4) (2024): 131–143
DOI 10.5117/mab.98.124167

Research Article



Toepassing van data-analyse om de steekproef te rationaliseren

Koen Derks, Lotte Mensink, Jacques de Swart, Ruud Wetzels

Received 28 March 2024 | **Accepted** 29 August 2024 | **Published** 18 September 2024

<http://mab-online.nl/article/124167/>