

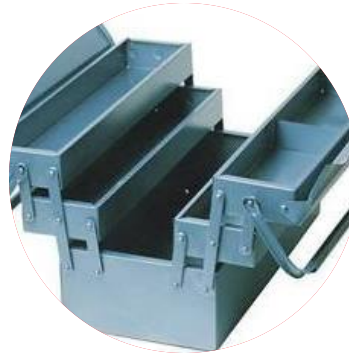
Statistiek B – contactmoment 1

Analysetechnieken

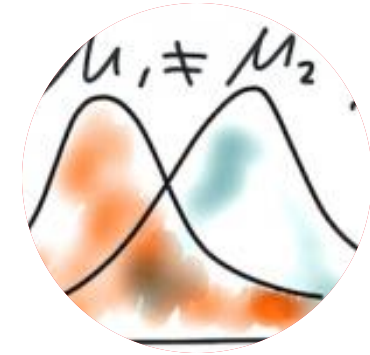
t-test



Praktisch



Analysetechnieken



t-test

Praktisch

BIVARIATE EN MULTIVARIATE STATISTIEK

Een Open Leerpakket in R

Sven De Maeyer, Liesje Coertjens en Jan Aerts



ACADEMIA
PRESS

Cursusmateriaal

- Cursus website: <https://statb-2024-2025.netlify.app/>
- Open leerpakket (via ACCO)
 - 'Bivariate en multivariate statistiek met R. Een openleerpakket.' (De Maeyer, Coertjens en Ardies, 2012)
 - Hoofdstuk 10 is herwerkt:
Studiemateriaal > Openleerpakket > Hoofdstuk 10 HERWERKT
 - Errata: *Studiemateriaal > Openleerpakket > Wijzigingen aan het openleerpakket*
- Videolessen
- ZSO's

Opbouw statistiek B

- **Voorbereid student is er 2** (of meer) **waard**
 - Hoofdstukken uit OLP doornemen
 - Videolessen bekijken
 - Online (R) oefeningen maken
 - ZSO's van afgeronde hoofdstukken maken
 - Vragen -> discussieforum
- **Contactmoment:**
 - Korte herhaling
 - Korte instructie leerinhoud voor dat contactmoment
 - Oefenen!

Opbouw statistiek B

≠ volgorde boek

<u>C1</u>	19 /02/2025	18u30 - 22u	Plenair: S.R.008 Computer labs: S.N. 202 en S.N. 302	Van onderzoeksvraag naar methode t-test
<u>C2</u>	05 /03/2025	17u - 19u	Plenair: S.R.008 Computer labs: S.N. 202 en S.N. 302	ANOVA Kruistabel
<u>C3</u>	19 /03/2025	16u - 18u	Plenair: S.R.008 Computer labs: S.N. 102 en S.N. 202	Covariantie en correlatie Bivariate regressie
<u>C4</u>	26 /03/2025	19u30 - 22u	Plenair: S.R.008 Computer labs: S.N. 202 en S.N. 302	Assumpties bij regressie Meervoudige regressie
<u>C5</u>	23 /04/2025	17u - 19u	Plenair: S.R.008 Computer labs: S.N. 202 en S.N. 302	Regressie met dummyvariabelen en interacties
<u>C6</u>	07 /05/2025	19u30 - 21u30	Plenair: S.R.008 Computer labs: S.N. 202 en S.N. 302	Regressie met dummyvariabelen en interacties
<u>C7</u>	21 /05/2025	18u - 21u30	Plenair: S.R.008 Computer labs: S.N. 202 en S.N. 302	Oefenexamen

Evaluatie

- **Schriftelijk PC-examen in twee delen:**

- GESLOTEN DEEL (70%): focus op kennis, inzicht en output interpreteren
- OPEN DEEL (30%): focus op analyseren in R (~output genereren)

Je moet minimaal de helft halen op het gesloten deel om te kunnen slagen voor Statistiek B. Anders bedraagt de score voor Statistiek B maximaal 8/20.

- **Vorbereiden?**

- Laatste sessie ~ proefexamen
- ScRipt

Statistiek A?

- **We breiden uit...**

- Van *“Hoe hoog scoort een gemiddelde leerling in Vlaanderen op deze kennistest techniek?”*
- Naar *“Scoren jongens en meisjes anders op techniek?”*

- **Wat als je niet geslaagd bent voor statistiek A?**

- No worries! 😊
- Meetniveaus variabelen en z-scores!

Analysetechnieken



Meetniveau variabelen



Meetniveau variabelen



■ Kwalitatieve variabelen

- 2 categorieën

Ja
Nee

- 3 of meer categorieën

Onderwijsvorm

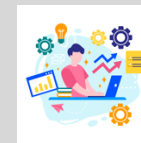
■ Kwantitatieve variabelen

Motivatie

Welke variabelen?

- *OV1) Kan attestering (a-attest, c-attest) in het 6de jaar ASO verklaard worden door sociale economische achtergrond (laag, midden, hoog)?*
- *OV2) Hangt de werkmotivatie van werknemers (schaalscore) samen met de bedrijfsgrootte (uitgedrukt in aantal werknemers)?*

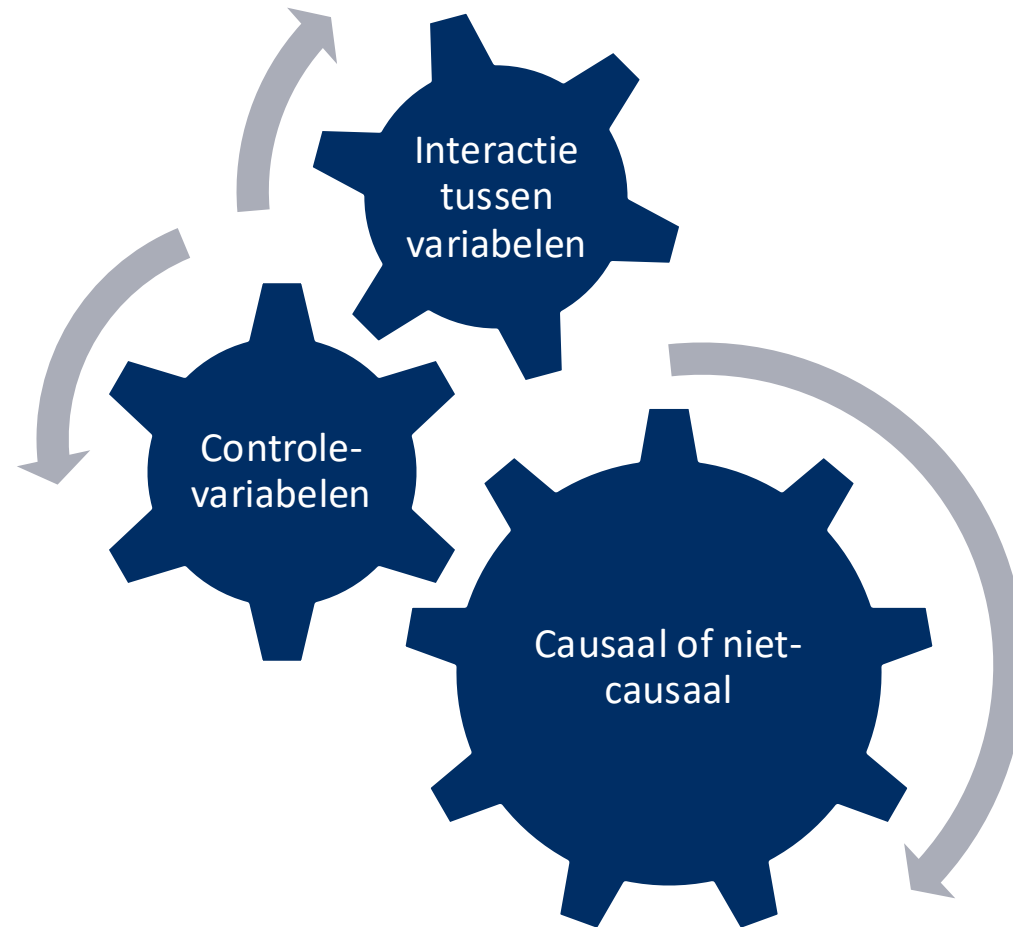
- 1. Wat zijn de variabelen?**
- 2. Teken deze variabelen**



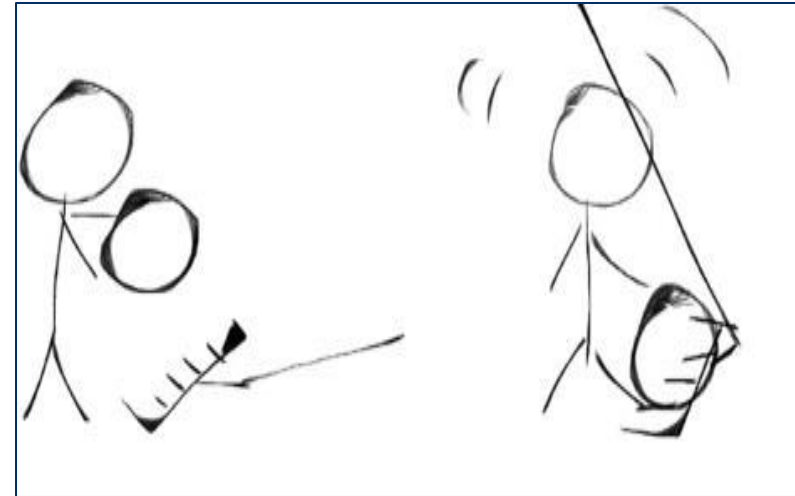
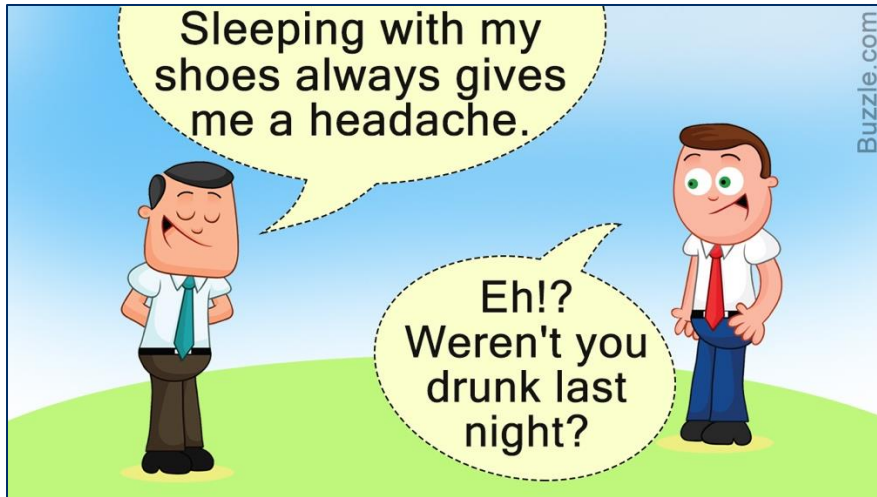
Relaties tussen variabelen?



Relaties tussen variabelen



Causale versus niet-causale verbanden (1)



Causaal?

Causale versus niet-causale verbanden (2)

- **Niet-causale verbanden = associaties zonder causaliteit**

Herkennen in OV? *samenhang, geassocieerd met, ...*

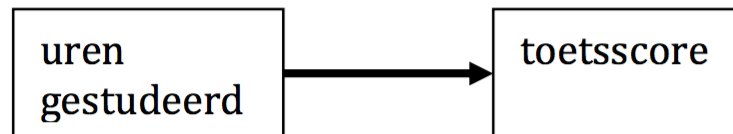
Tekenen?



- **Causale verbanden = associaties in termen van oorzaak (onafhankelijke var.) en gevolg (afhankelijke var.)**

Herkennen in OV? *effect, impact, voorspellen, verklaren, ...*

Tekenen?



Complexere verbanden: controle variabelen

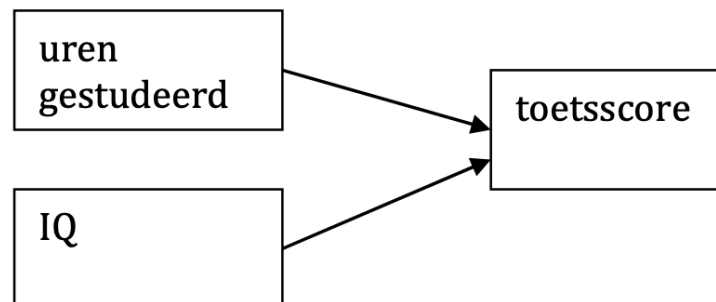
- **Bijvoorbeeld:**

Heeft het aantal uren dat een leerling studeert voor een toets een invloed op de scores voor die toets, ongeacht het IQ van de leerling?

- **Om oneigenlijke verbanden uit te sluiten** (~interne validiteit, zie IMT!)

Herkennen in OV? *Wat is het effect van X op Y, ongeacht Z;
rekening houdend met Z;
controllerend voor Z?*

Tekenen?



Complexere verbanden: interactie tussen variabelen

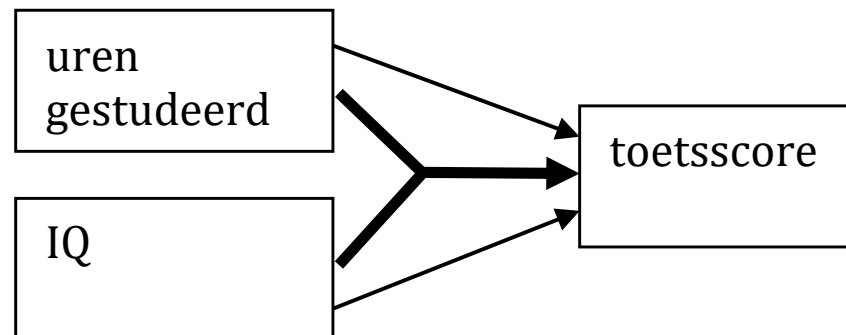
- **Bijvoorbeeld:**

Is het verband tussen het aantal studieuren en de toetsscore afhankelijk van het IQ van een leerling?

- **Om complexere relaties in kaart te brengen**

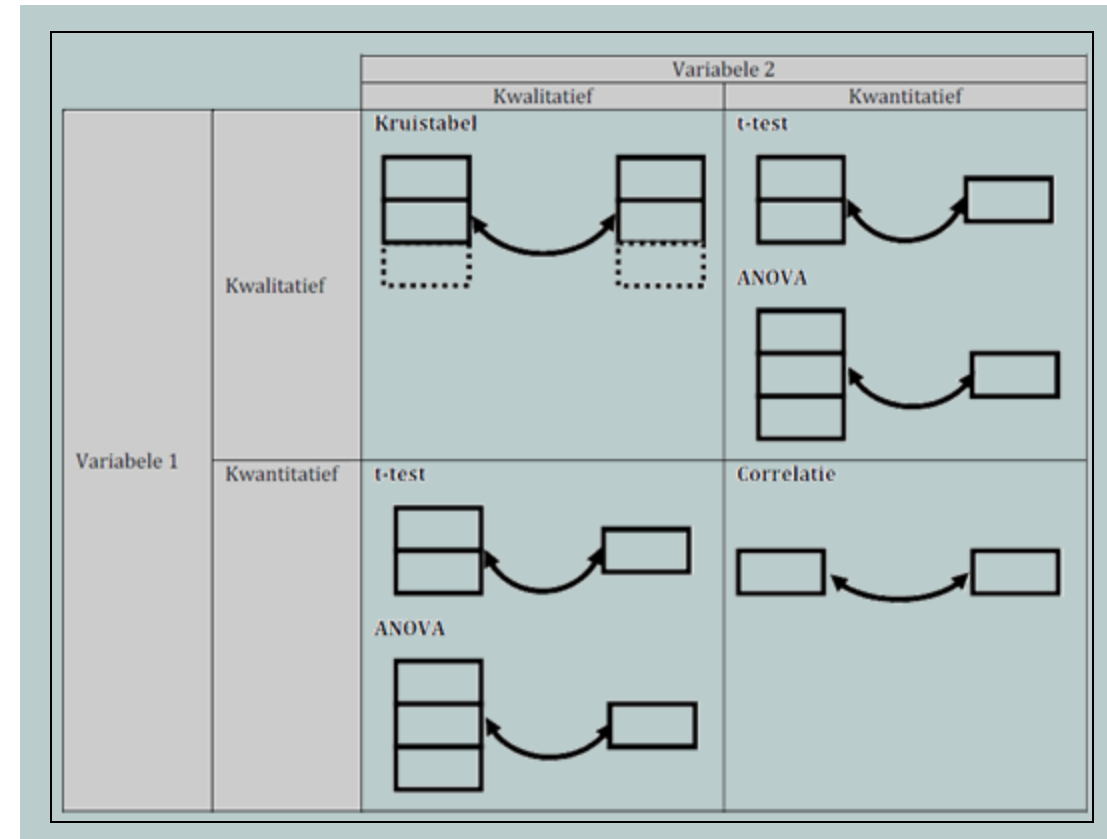
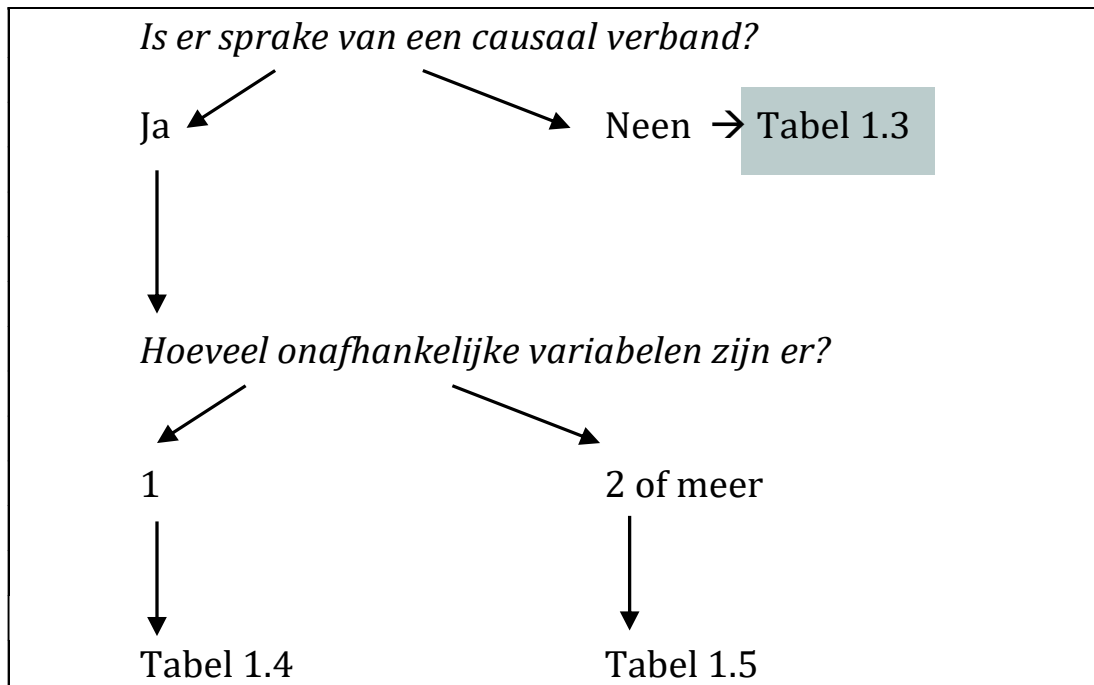
*Herkennen in OV? Is het effect van X op Y... afhankelijk van Z;
hetzelfde/identiek voor ... als voor ...?*

Tekenen?



Hoe gepaste techniek kiezen?

1. Teken je onderzoeksvraag
2. Gebruik de flowcharts in het OLP



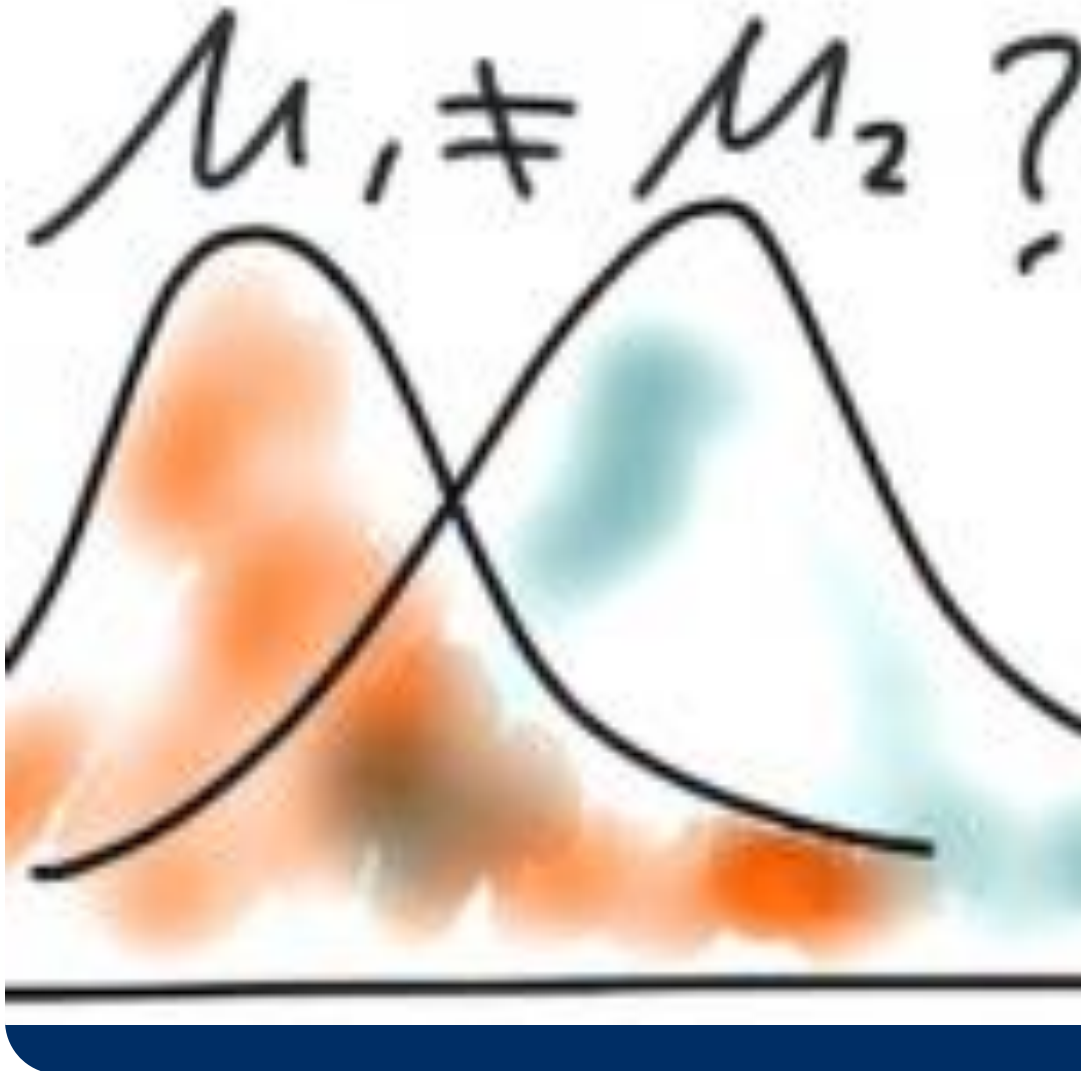
Oefening

- *OV1) Scoren meisjes anders op wiskunde dan jongens?*
- *OV2) In welke mate hangt sociaal-economische status ('laag', 'midden', 'hoog') samen met al dan niet bijscholing krijgen?*
- *OV3) Is de invloed van zelfeffectiviteit op wiskundescores (schaalscore) anders naargelang studiemotivatie (schaalscore) van de leerling, ongeacht het geslacht?*

- 1. Teken de onderzoeksvraag**
- 2. Kies de gepaste techniek**



t-test





Studie 1

EFFECT VAN OPLEIDING?

Vergelijking van hoeveelheid huishoudelijk afval (in kg) geproduceerd door 100 gezinnen die training volgden en 100 willekeurig gekozen gezinnen uit dezelfde populatie



Studie 2

EFFECT VAN EDUCATIEVE CAMPAGNE?

Meting van hoeveelheid huishoudelijk afval (in kg) bij 100 gezinnen voor de campagne en meting bij dezelfde 100 gezinnen nadat de campagne een maand gelopen heeft



Studie 1



Studie 2



Deze zou je eventueel zo kunnen visualiseren, maar eigenlijk gaat het 2 maal om kg afval (kg voor en kg na)

2 vormen van t-test

- **Independent samples design**

- Respondenten Conditie 1 \neq respondenten Conditie 2
=> Independent samples t-test



- **Repeated measures design**

- Respondenten Conditie 1 = respondenten Conditie 2
=> Paired-samples t-test



Hoe nagaan of er een effect is?



Hoe nagaan of er een effect is?

- Gemiddeldes vergelijken
- Betrouwbaarheidsintervallen rond gemiddelde!

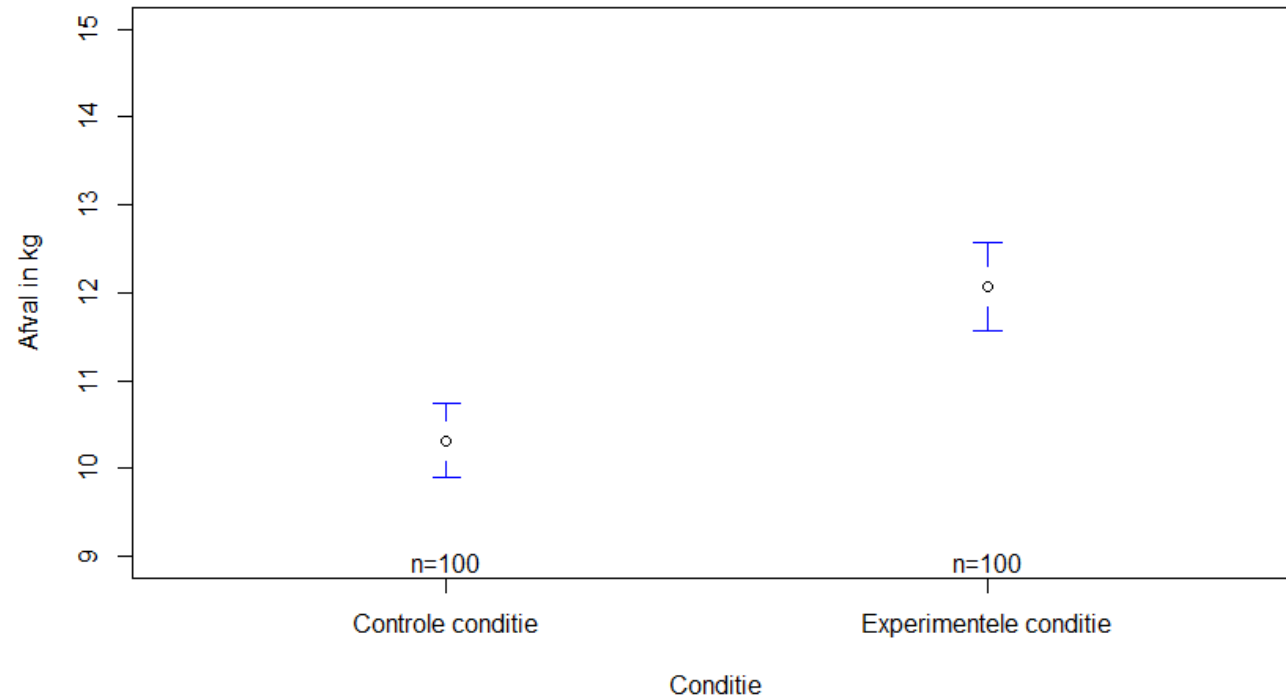


Hoe nagaan of er een effect is?

```
> tapply(Afval$Afval.in.kg, Afval$Conditie, mean)
  Controle conditie  Experimentele conditie 
      10.32127         12.06773 

> errorbar(Afval$Afval.in.kg ~ Afval$Conditie,
  xlab = "Conditie",
  ylab = "Afval in kg",
  ylim = c(9,15))
```

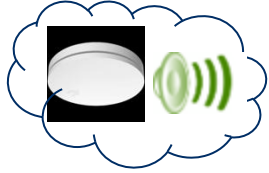
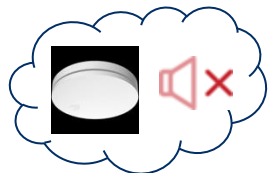
DUS?



Wat kan er aan de hand zijn?



De ene fout is de andere niet...



Brandalarm gaat af		
		° Niets aan de hand °
Realiteit		
		Er is brand °
Realiteit		

Van brandalarm naar statistiek?



Van brandalarm naar statistiek?



Er is niets aan de hand ...

~ Er is geen verschil tussen beide condities in de populatie

NULHYPOTHESE (H_0)







Het brandt ...

~ Er is wel een verschil tussen beide condities in de populatie

ALTERNATIEVE HYPOTHESE (H_1)

Type I en type II fouten

		Steekproef (onze data) zegt ...	
		$H_0 = \text{juist}$	$H_0 = \text{fout}$
Realiteit	$H_0 = \text{juist}$		Type I fout (vals positief) α 
	$H_0 = \text{fout}$	Type II fout (vals negatief) β 	

Opleiding heeft GEEN effect in steekproef én ook NIET in populatie

Opleiding heeft WEL een effect in steekproef maar NIET in populatie

Opleiding heeft GEEN effect in steekproef maar WEL in populatie

Opleiding heeft WEL een effect in steekproef én OOK in populatie

Type 1 en 2 fouten bij brandalarm:

FOCUS t-test

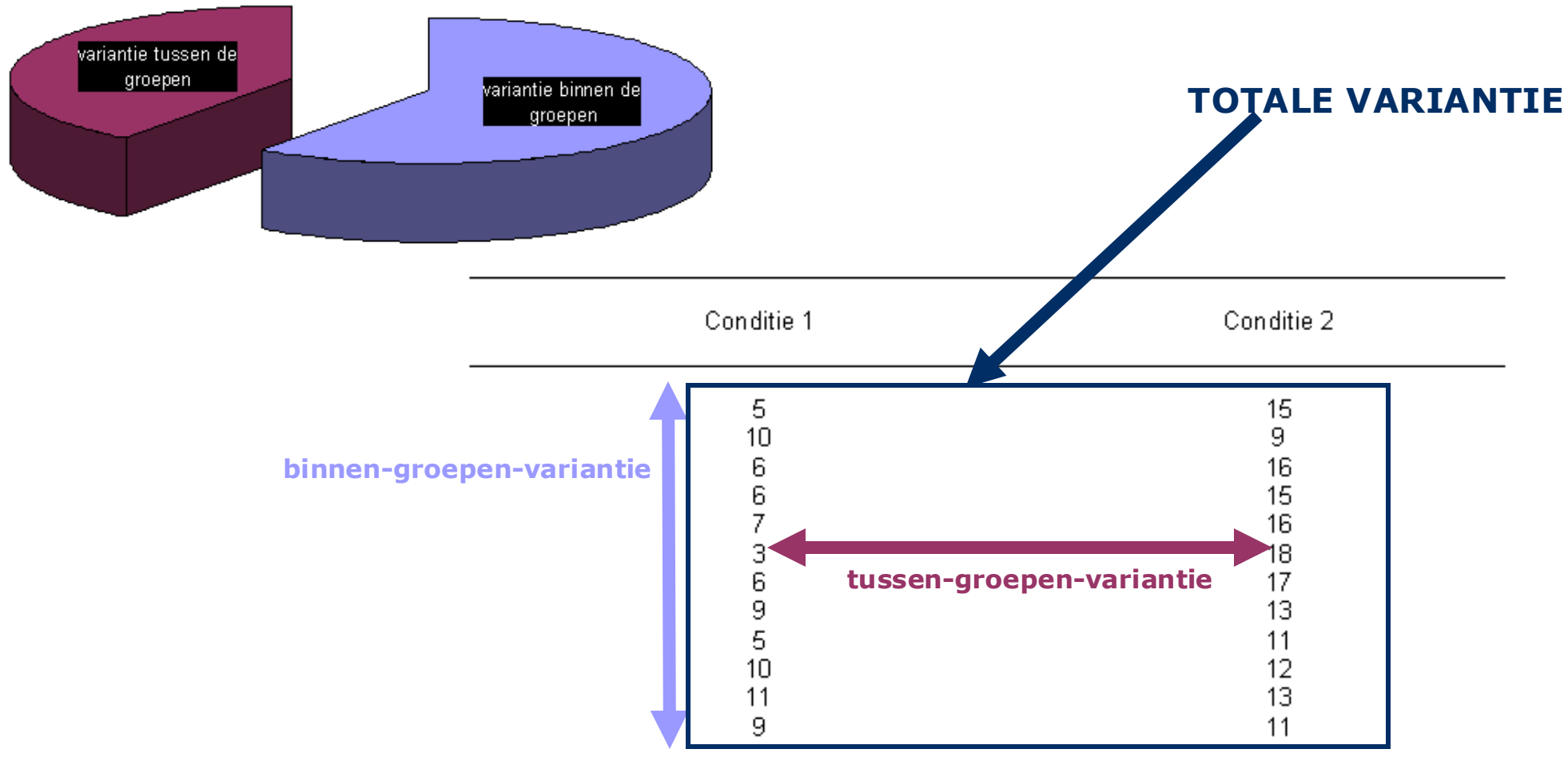


Brandalarm gaat niet af, maar er is brand...
Toch denken dat er geen brand is, is een
type 2 fout (**vals negatief**)



Brandalarm gaat af, maar er is geen
brand... Het is gewoon warm!
Toch denken dat er brand is, is een type
1 fout (**vals positief**)

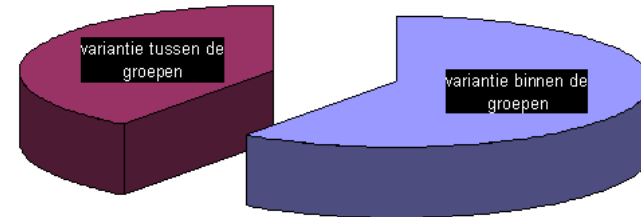
t-test: twee soorten variantie



t-test: berekening

- Toets gebaseerd op verhouding ‘tussen-groepen-variantie’ en ‘binnen-groepen-variantie’:

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S^2_1}{n_1} + \frac{S^2_2}{n_2}}}$$



- Indicator ‘tussen-groepen-variantie’: verschil in gemiddelde score ($\bar{x}_1 - \bar{x}_2$)
- Indicator ‘binnen-groepen-variantie’: moeilijker te berekenen ($\sqrt{\frac{S^2_1}{n_1} + \frac{S^2_2}{n_2}}$)



t-waarde

Betekenis t-waarde

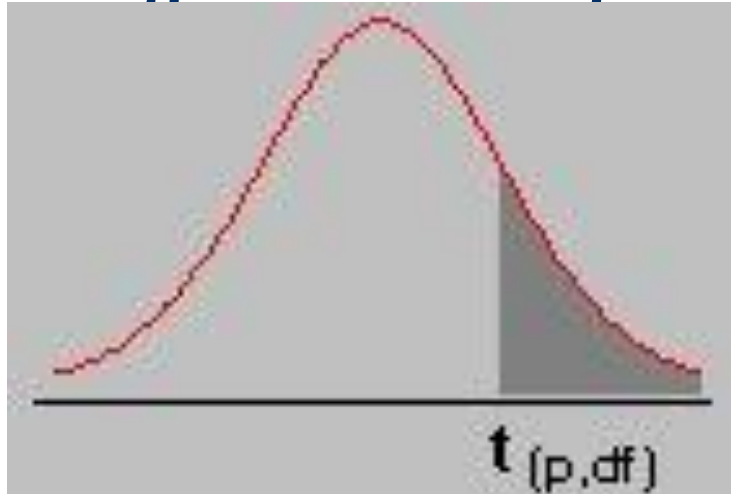
- t-waarde is hoog → groot verschil tussen beide groepen
- t-waarde is laag → klein verschil tussen beide groepen

t-test: t-verdeling

■ Theoretische kansverdeling

- Wat is de kans (p) om een t-waarde te bekomen in een steekproef indien H_0 geldig is in de populatie?
- *In hoeveel % van gelijkaardige steekproeven zouden we zo'n t-waarde of hoger tegenkomen als er in de gehele populatie geen verschil is?*

- **Afhankelijk van een andere grootte: 'vrijheidsgraden' (df)** (= aantal respondenten – aantal condities)



t-test: vuistregel

- **$p \leq 0,05$ wijst op een statistisch significant verschil**

“Kans dat we dit verschil in onze steekproef zouden vaststellen indien de nulhypothese opgaat voor de populatie is kleiner dan of gelijk aan 0,05 (~5%)” of

“We hebben 5% kans of minder om een type-I fout te maken (= ten onrechte de nulhypothese verwerpen)”

- **Dus?**

- H_0 geldig? Bijzonder kleine kans!
- H_1 aannemen: er is een verschil tussen gemiddeldes in de populatie



Ander voorbeeld...

“Heb je je tanden gepoetst?”

Observatie: tandenborstel is vochtig

H_0 : niet de tanden gepoetst

p-waarde: kans dat de tandenborstel zo vochtig is
(\sim t-waarde) indien niet de tanden gepoetst (H_0)

t-test in R

```
t.test(Afval$Afval.in.kg ~ Afval$Conditie, var.equal = TRUE)
```

Two Sample t-test

data: Afval\$Afval.in.kg by Afval\$Conditie

t = -5.251, df = 198, p-value = 3.886e-07

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

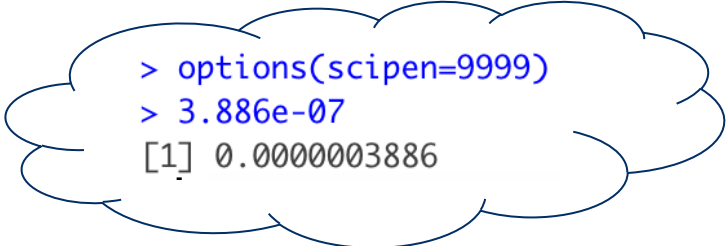
-2.402343 -1.090571

sample estimates:

mean in group Controle conditie mean in group Experimentele conditie

10.32127

12.06773



```
> options(scipen=9999)  
> 3.886e-07  
[1] 0.0000003886
```


t-test: assumptions (1)

- **Voorwaarden om t-test te mogen gebruiken?**

- Afhankelijke variabele is normaal verdeeld
- Binnengroepvarianties zijn gelijk

- **Assumptie m.b.t. binnengroepvarianties checken in R** (package car laden)

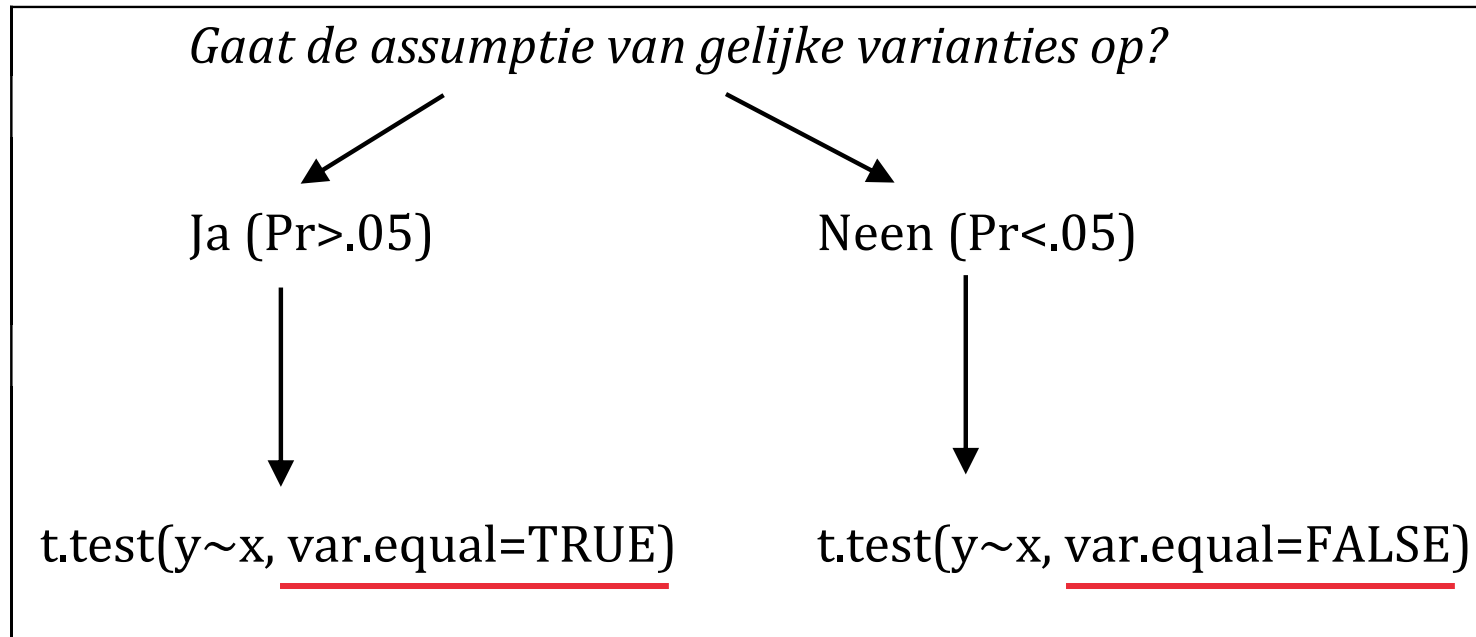
```
> library(car)
> leveneTest(Afval$Afval.in.kg, Afval$Conditie)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  1.0953 0.2966
      198
```

H_0 : er is geen verschil in binnengroepvarianties
 H_1 : er is wel een verschil in binnengroepvarianties

DUS...

t-test: assumptions (2)

- Wat als binnengroepvarianties wel verschillen?



t-test: hoe groot is het effect?

- **Cohen's d:**
$$d = \frac{\bar{x}_1 - \bar{x}_2}{(s_1 + s_2) / 2}$$

- **Interpretatie?**

Effect Size (d)	Cohen's vuistregel
≥ 0.8	Groot
0.5-0.7	Medium
0.2-0.4	Klein
≤ 0.2	Geen effect

- **In R:**

```
> source(file.choose()) # 0LP2 functies laden  
> d(Afval$Afval.in.kg, Afval$Conditie)  
[1] -0.7449206
```

Significantie en effectgrootte

- **Kan je verklaren waarom...**
 - een miniem verschil tussen beide condities toch een significante t-test oplevert?
 - een verschil niet significant is, maar wel een medium tot sterk effect heeft?

Significantie en effectgrootte

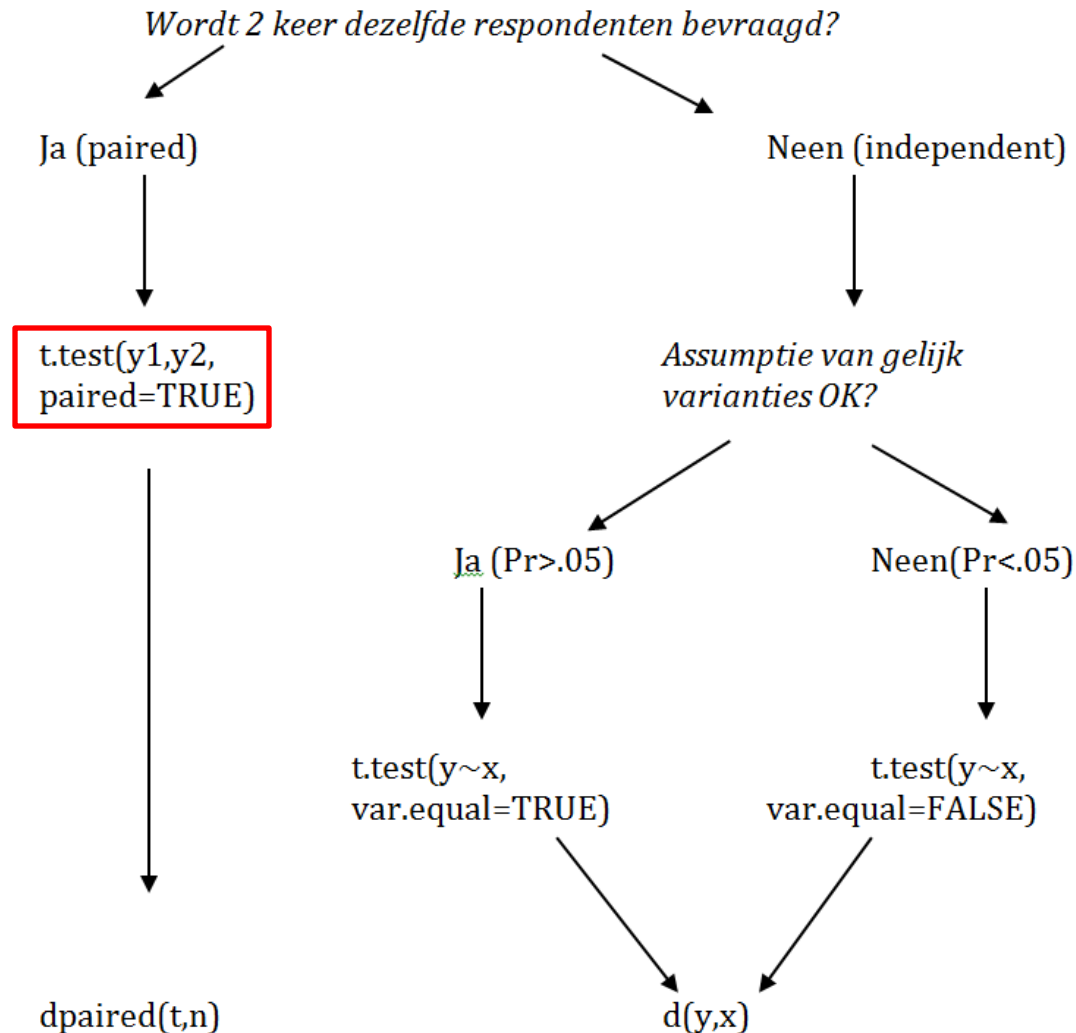
- **Kan je verklaren waarom...**

- een miniem verschil tussen beide condities toch een significante t-test oplevert?
- een verschil niet significant is, maar wel een medium tot sterk effect heeft?

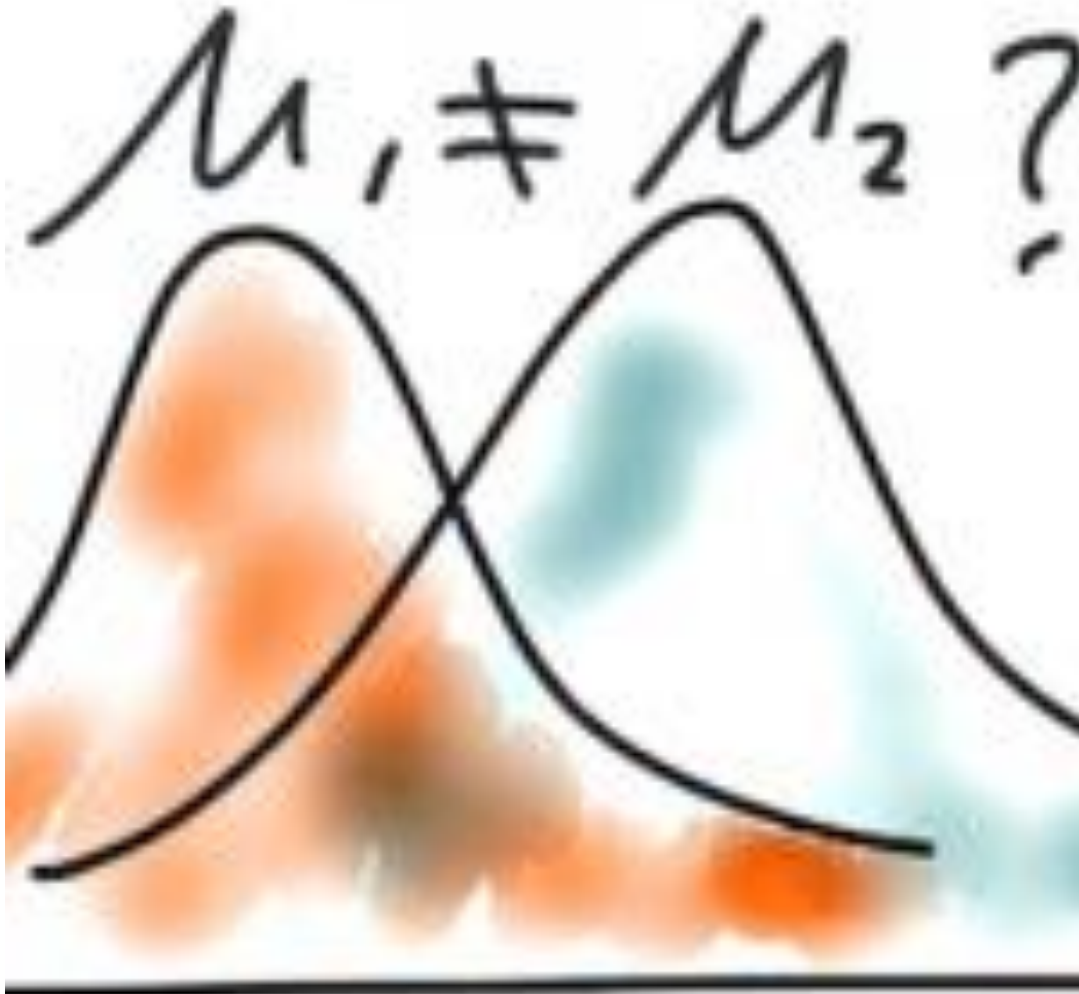
- **Vergelijking gebaseerd op betrouwbaarheidsintervallen**

$$SD(\bar{x}) = \sqrt{\frac{s_x^2}{n}} = \frac{s_x}{\sqrt{n}}$$

Wat met een repeated measures design?



Figuur 3.9 Flowchart t-test procedure



t-test

Gemiddeldes?

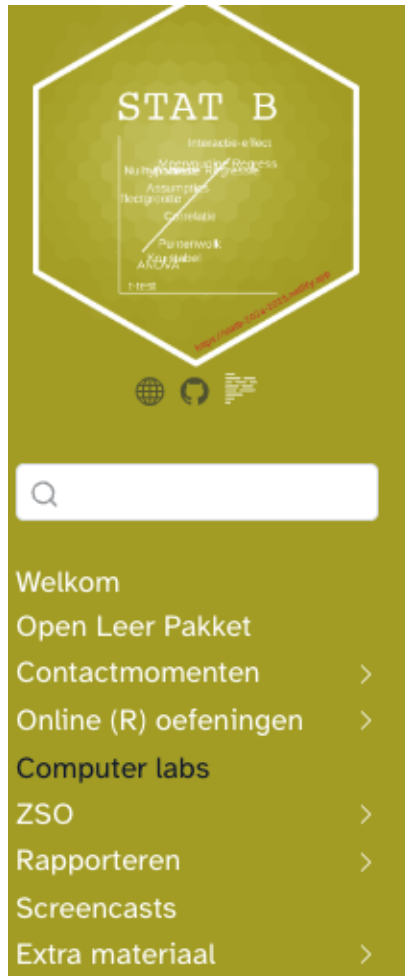
Assumptie binnengroepvariantie?

Statistische significantie? (t-test)

Effectgrootte?













Time for P  actice

https://statb-2024-2025.netlify.app/computer_labs



Computer labs

Tijdens de contactmomenten is er tijd voorzien om een reeks oefeningen uit te voeren in **R** in het bijzijn van de docenten. Dit noemen we de *computer labs*. Hieronder vind je een lijst van de oefeningen en responsen georganiseerd per contactmoment.

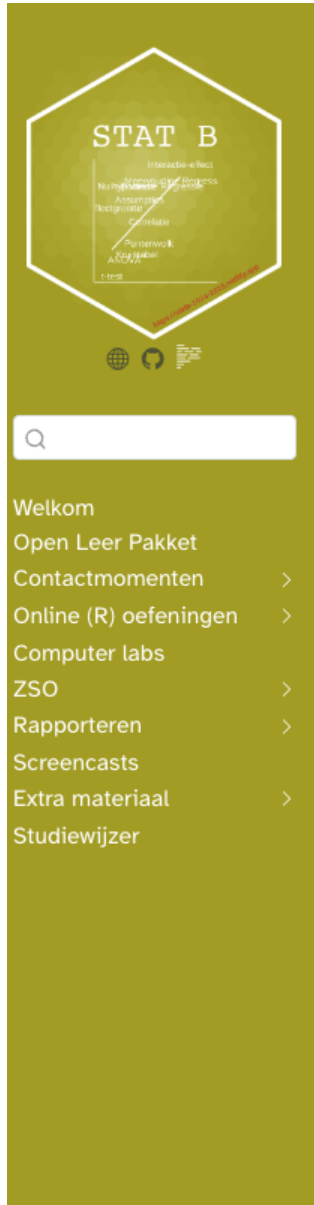
-  [C1: Opdrachten computer lab](#) /  [C1: Respons computer lab](#)
-  [C2: Opdrachten computer lab](#) /  [C2: Respons computer lab](#)
-  [C3: Opdrachten computer lab](#) /  [C3: Respons computer lab](#)
-  [C4: Opdrachten computer lab](#) /  [C4: Respons computer lab](#)
-  [C5: Opdrachten computer lab](#) /  [C5: Respons computer lab](#)
-  [C6: Opdrachten computer lab](#) /  [C6: Respons computer lab](#)

Ter info

Doorheen het vak maken we oefeningen in de computerlokalen. De zogenaamde *computer labs* op deze website. Al deze oefeningen maken gebruik van dezelfde dataset: **Techniek.RData**.

De dataset kan je downloaden via Blackboard of via deze link: [download Techniek.RData](#)

https://statb-2024-2025.netlify.app/contactmomenten/c1/opdrachten_c1



Contactmoment 1: Opdrachten computer lab

Oefening 1

- Is er een verschil in interesse voor techniek ([Interest.voor](#)) tussen leerlingen met een vader die wel of niet een diploma hoger onderwijs kehaalde ([Dipvader](#))?
- Hoe groot is het effect van de variabele [Dipvader](#) op [Interest.voor](#) ?
- Geef het effect visueel weer.

Oefening 2

- Is er een verschil in interesse voor techniek ([Interesse.voor](#)) tussen eerste- en tweedejaars leerlingen ([Studiejaar](#))?
- Hoe groot is het effect van [Studiejaar](#) op [Interesse.voor](#) ?
- Geef het effect visueel weer.

Oefening 3

In het onderzoek achter de dataset is de interesse van leerlingen voor techniek op twee meetmomenten gemeten: aan het begin van een trimester ([Interest.voor](#)) en aan het einde van datzelfde trimester ([Interest.na](#)). Daarom kunnen we de volgende onderzoeksvraag nagaan:

- Verschilt de interesse voor techniek van leerlingen in het begin van een trimester ('Interest.voor') van de interesse van leerlingen voor techniek aan het einde van een trimester ('Interest.na')?
- Hoe groot is het effect van een trimester technieklessen?
- Geef het effect visueel weer.