

## Hoofdstuk 5. Statistische besluitvorming

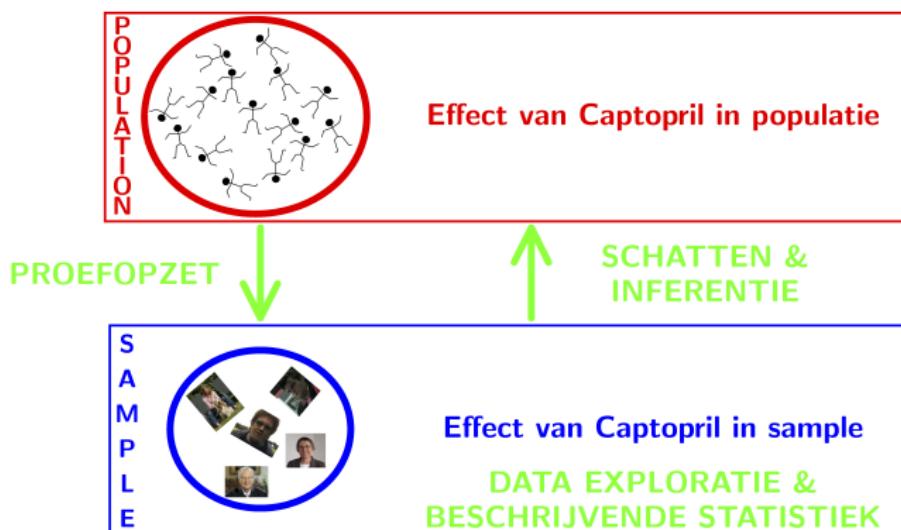
Lieven Clement

2<sup>de</sup> bach. in de Biologie, Chemie, Biochemie en Biotechnologie en Biomedische  
Wetenschappen

## 5.1. Inleiding



Onderzoekers wensen na te gaan of medicijn Captopril een bloeddruk verlagend effect heeft

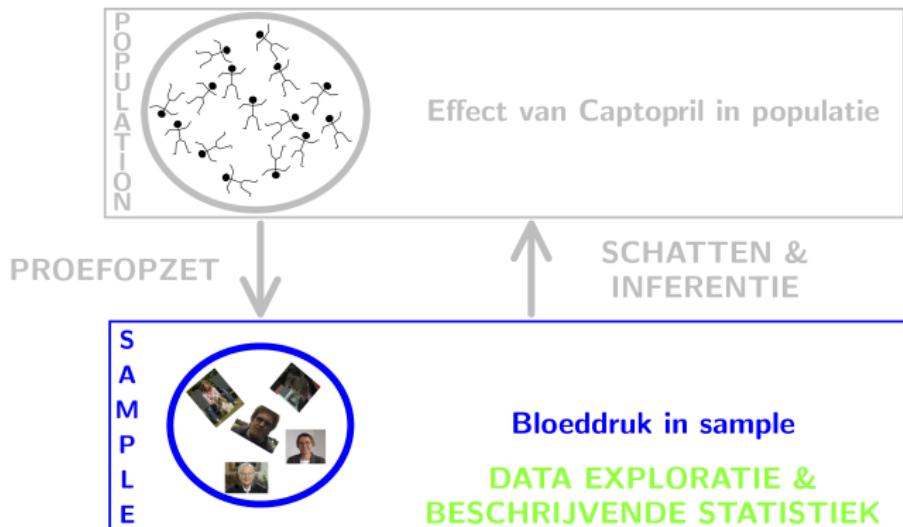


## Overzicht

- Proefopzet/Data Exploratie
- Puntschatters (Schatten)
- Intervalschatters (Statistische Besluitvorming)
- Hypothese testen (Statistische Besluitvorming)

### 5.2.1 Proefopzet

- 15 patiënten werden at random gekozen uit de populatie.
- pre-test/post-test design
- De systolische en diasystolische bloeddruk wordt gemeten voor en na het toedienen van captopril.
- Voordeel: effect van het toedienen van captopril op de bloeddruk kunnen meten voor elke patiënt.
- Nadeel?



## 5.2.2. Data Exploratie & Beschrijvende Statistiek

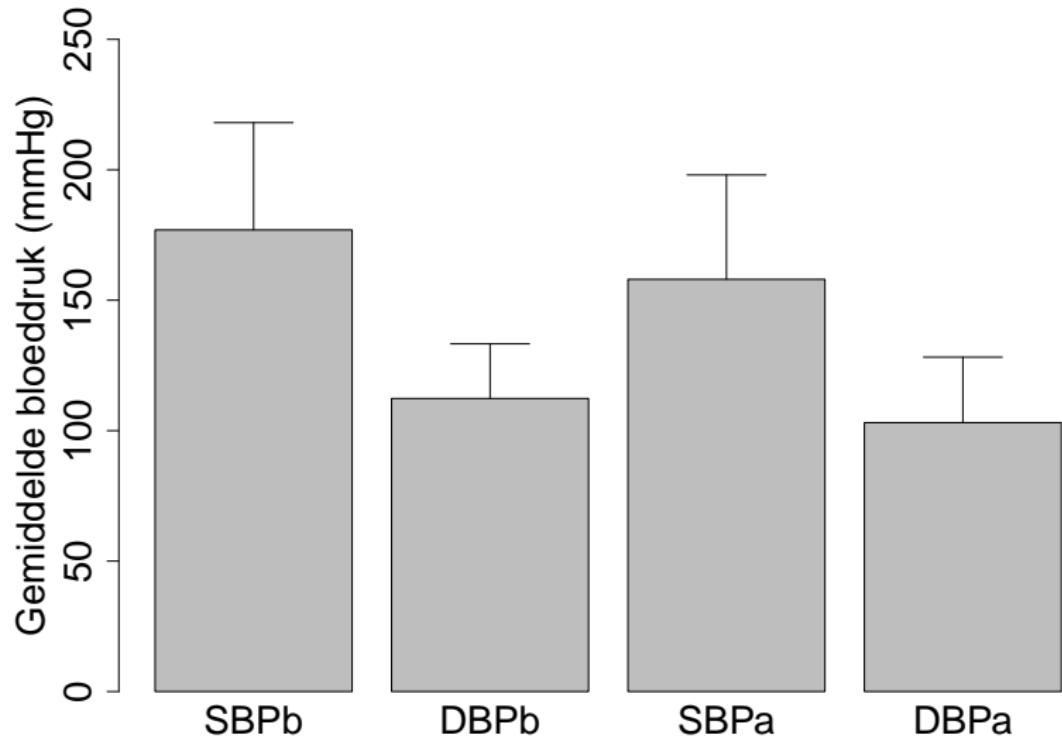
```
captopril <- read.table("dataset/captopril.txt", header=TRUE, sep="")  
head(captopril)
```

```
##   id SBPb DBPb SBPa DBPa  
## 1  1  210  130  201  125  
## 2  2  169  122  165  121  
## 3  3  187  124  166  121  
## 4  4  160  104  157  106  
## 5  5  167  112  147  101  
## 6  6  176  101  145   85
```

```
mm<-apply(captopril[,2:5], MARGIN=2, FUN=mean)  
hh<-apply(captopril[,2:5], MARGIN=2, FUN=sd)  
mp <- barplot(mm, ylim=c(0,250), ylab="Gemiddelde bloeddruk (mmHg)  
#fouten vlaggen  
segments(mp,mm,mp,mm+2*hh)  
segments(mp-.2,mm+2*hh,mp+.2,mm+2*hh)
```

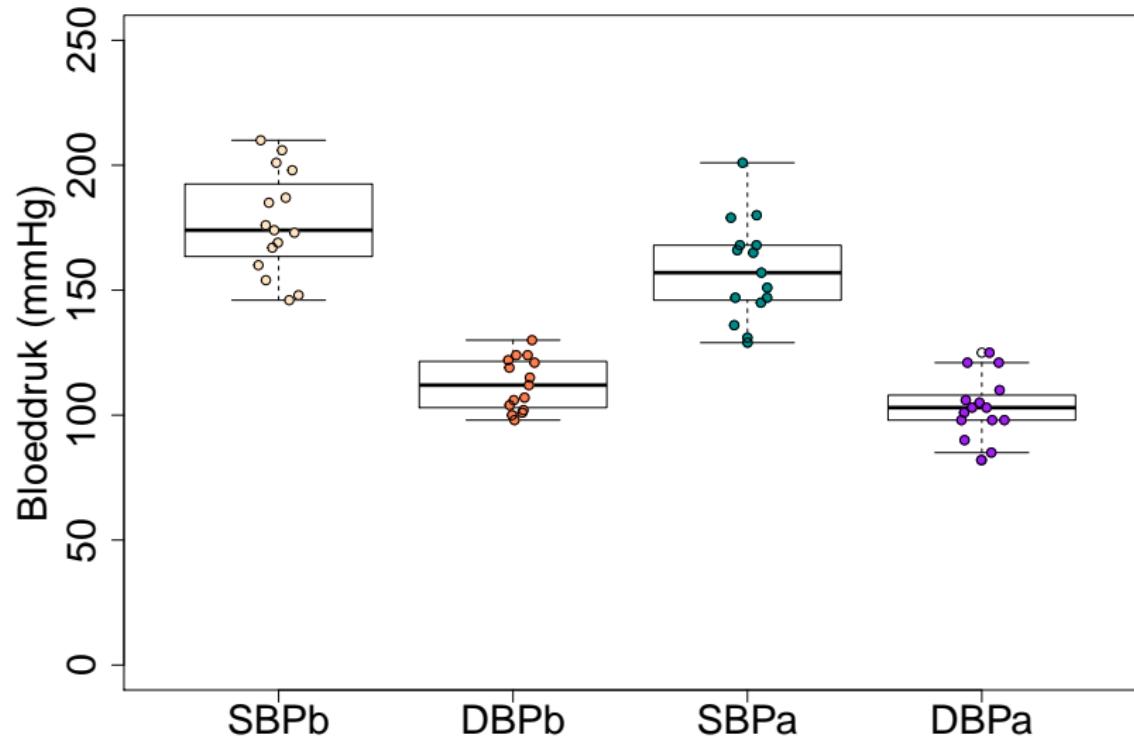


Is deze figuur informatief?



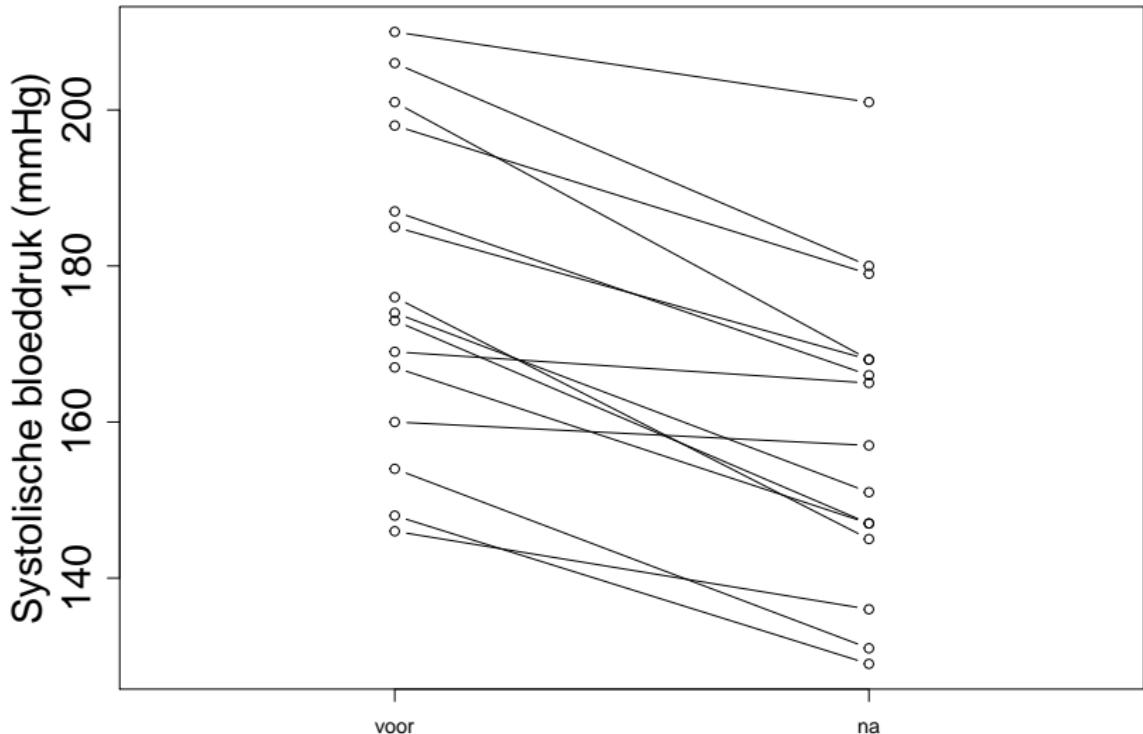
```
boxplot(captopril[,2:5],ylim=c(0,250),ylab="Bloeddruk (mmHg)",ma
set.seed(19)
stripchart(captopril[,2:5],
           vertical = TRUE, method = "jitter",
           pch = 19, col =c("bisque","coral","darkcyan","purple",
           add = TRUE)
set.seed(19)
stripchart(captopril[,2:5],
           vertical = TRUE, method = "jitter",
           pch = 1, col =1,
           add = TRUE)
```

Is deze figuur informatief?



```
matplot(t(captopril[,c("SBPb","SBPa")]),pch=1,lty=1,col="black",  
axis(1,c(1,2),labels=c("voor","na"))
```

Is deze figuur informatief?



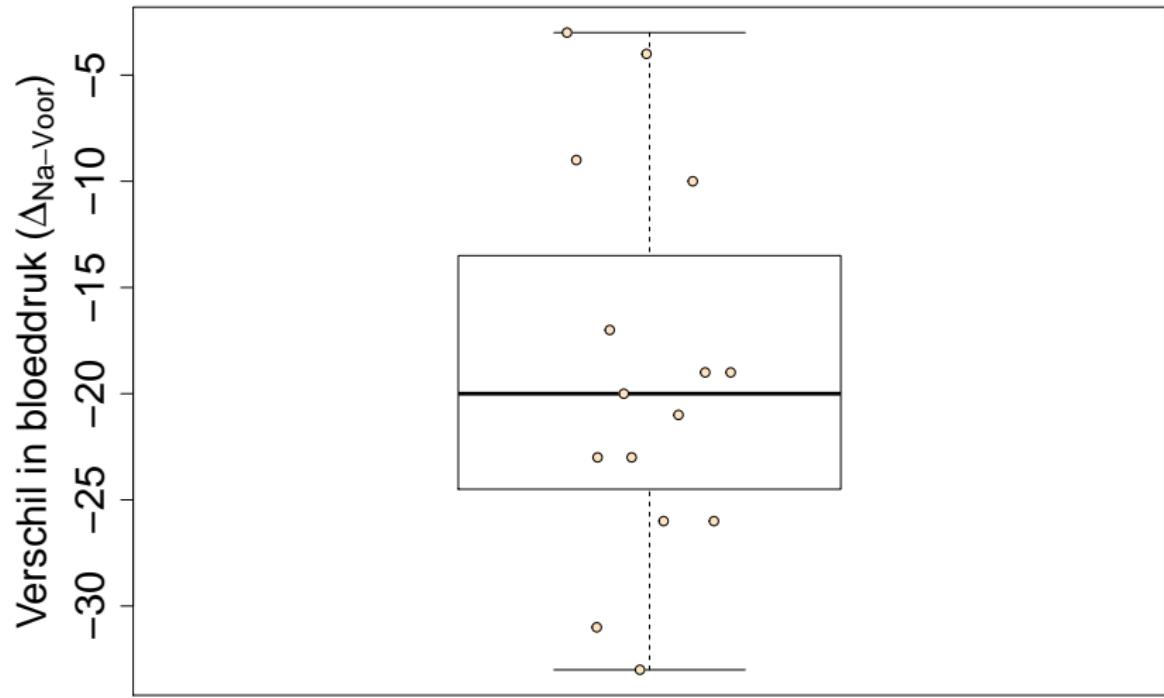
```
delta <- captopril$SBPa - captopril$SBPb
boxplot(delta,
         ylab=expression(paste("Verschil in bloeddruk (", Delta[Na - V
set.seed(19)
stripchart(delta,
            vertical = TRUE, method = "jitter",
            pch = 19, col =c("bisque"),
            add = TRUE)
```

```
summary(delta)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## -33.00 -24.50 -20.00 -18.93 -13.50 -3.00
```

```
sd(delta)
```

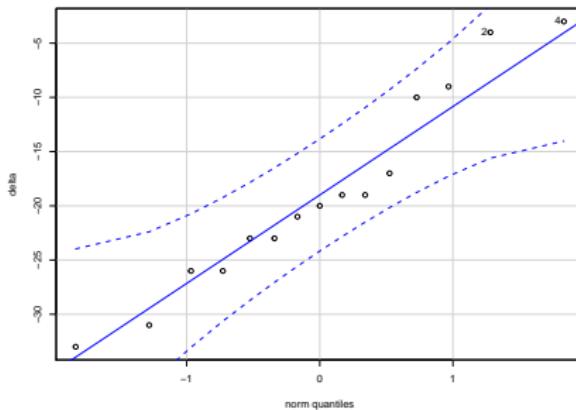
```
## [1] 9.027471
```



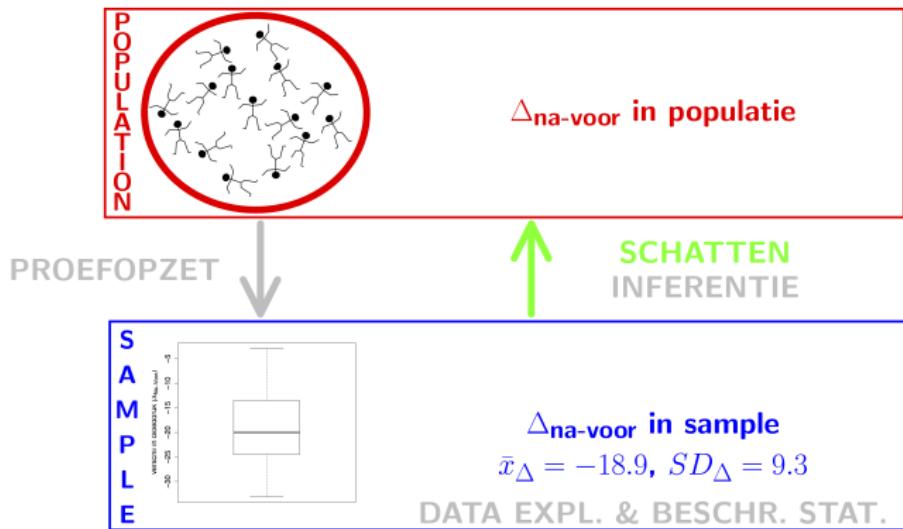
- Pre-test/post-test design: Effect van captopril in steekproef bestuderen via  $X = \Delta_{\text{na-voor}}$ !
- Hoe  $X = \Delta_{\text{na-voor}}$  modelleren en effect van captopril schatten?

```
library(car)
qqPlot(delta,main=expression(paste("Verschil in Systolische Bloed
## [1] 4 2
```

Verschil in Systolische Bloeddruk ( $\Delta_{Na-Voor}$ )

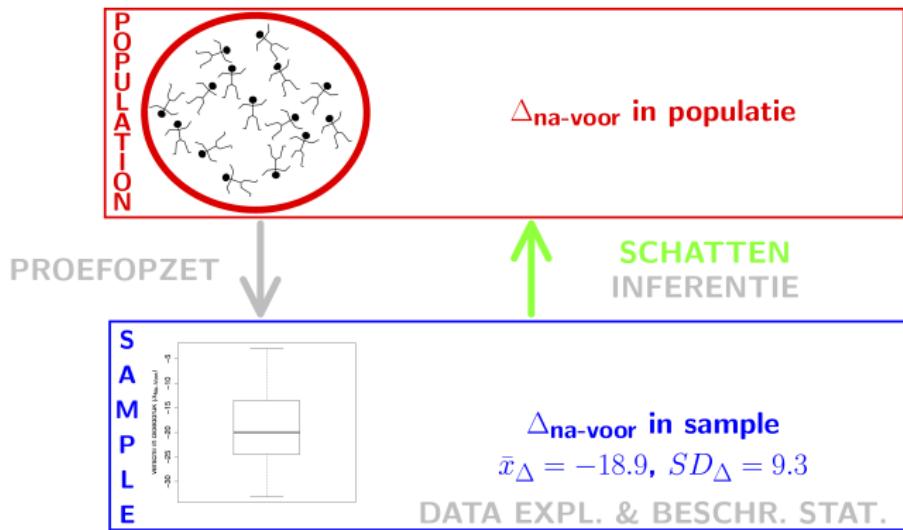


```
## [1] 4 2
```



### 5.2.3. Schatten

- Geen substantiële afwijkingen van Normaliteit
- We kunnen dus veronderstellen dat  $X \sim N(\mu, \sigma^2)$ .
- Effect van captopril in de populatie beschrijven a.d.h.v. de gemiddelde bloeddrukverschil  $\mu$ .
- Het gemiddeld bloeddrukverschil  $\mu$  in de populatie kan worden geschat a.d.h.v. het steekproefgemiddelde  $\bar{x}=-18.93$
- De standaard afwijking  $\sigma$  a.d.h.v. de steekproefstandaarddeviatie  $SD=9.03$ .
- Is het effect dat we observeren in de steekproef groot genoeg is om te kunnen spreken van een effect van captopril in de populatie?
- Schatting zal immers van steekproef tot steekproef variëren



### 5.3. Puntschatters: het steekproefgemiddelde

- Stel dat het bloeddrukverschil  $X$  een lukrake trekking is uit de populatie en onderstel bv  $X \sim N(\mu, \sigma^2)$
- $\mu$  schatten o.b.v. steekproef,  $X_1, \dots, X_n$ , a.d.h.v. het *steekproefgemiddelde*

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$$

van de toevalsveranderlijken  $X_1, X_2, \dots, X_n$ .

- Steekproefgemiddelde  $\bar{X}$  is dus een toevallig veranderlijk die zal variëren van steekproef tot steekproef.
  - Daarom zullen we theoretische verdeling van het steekproefgemiddelde bestuderen
- 
- ➊ Inzicht in welke mate het resultaat van de studie zou variëren indien men een nieuwe, gelijkaardige studie zou opzetten
  - ➋ Leren hoe ver  $\bar{X}$  van  $\mu$  kan afwijken.

## Overzicht

- ① Onvertekendheid van steekproefgemiddelde
- ② Precisie van steekproefgemiddelde
- ③ Verdeling van steekproefgemiddelde

### 5.3.1. Het steekproefgemiddelde is onvertekend

- Resultaten uit steekproef veralgemeenbaar naar studiepopulatie als schatting goed populatiewaarde benadert.
- Veralgemenen van steekproef naar populatie vereist representatieve steekproef
- Dit vermindert dat resultaten vertekend zijn (systematisch te hoog/laag).
- Belangrijk te rapporteren hoe steekproef gekomen werd.

## Intermezzo: Ecstasy bij Stanford studenten

- Variabele: X: Geen ecstasy gebruik  $X=0$ , ecstasy gebruik  $X=1$
- $n = 369$  respondenten
- De prevalentie bedraagt 39%

## Intermezzo: Ecstasy bij Stanford studenten

- Variabele: X: Geen ecstasy gebruik X=0, ecstasy gebruik X=1
- n = 369 respondenten
- De prevalentie bedraagt 39%
- Enkel studenten die zich overdag buiten op de campus bevinden
- Niet representatief.

- Andere voorbeeld
- Gemiddelde gewicht van veldmuizen
- moeilijk: vallen zetten, hongerige muizen eerst en zijn mss ook magerst (haphazard sampling)

- Hoe representativiteit garanderen
- Randomizatie!
- Subjecten lukraak trekken uit populatie (eenvoudige lukrake steekproef): elk subject evenveel kans, niemand 'bevoordeeld'
- Patiënten met hypertensie in captopril studie zijn at random gekozen uit de populatie

- Eenvoudige lukrake steekproef:  $X_1, \dots, X_n$  voor een karakteristiek  $X$
- $X_1, \dots, X_n$  allen dezelfde verdeling
- Ze hebben allen gemiddelde  $\mu$  en variantie  $\sigma^2$
- $E(X_1) = \dots = E(X_n) = \mu$  en  $\text{Var}(X_1) = \dots = \text{Var}(X_n) = \sigma^2$

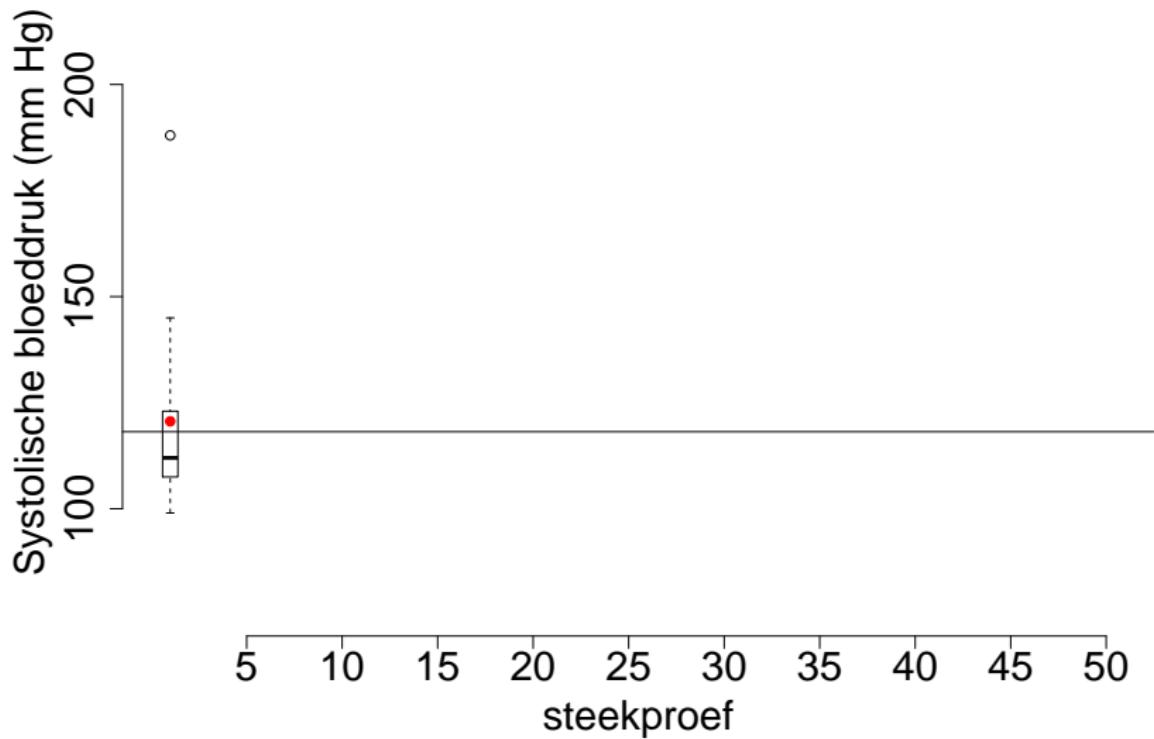
$$\begin{aligned}
 E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\
 &= \frac{E(X_1) + E(X_2) + \dots + E(X_n)}{n} \\
 &= \frac{\mu + \mu + \dots + \mu}{n} \\
 &= \mu
 \end{aligned}$$

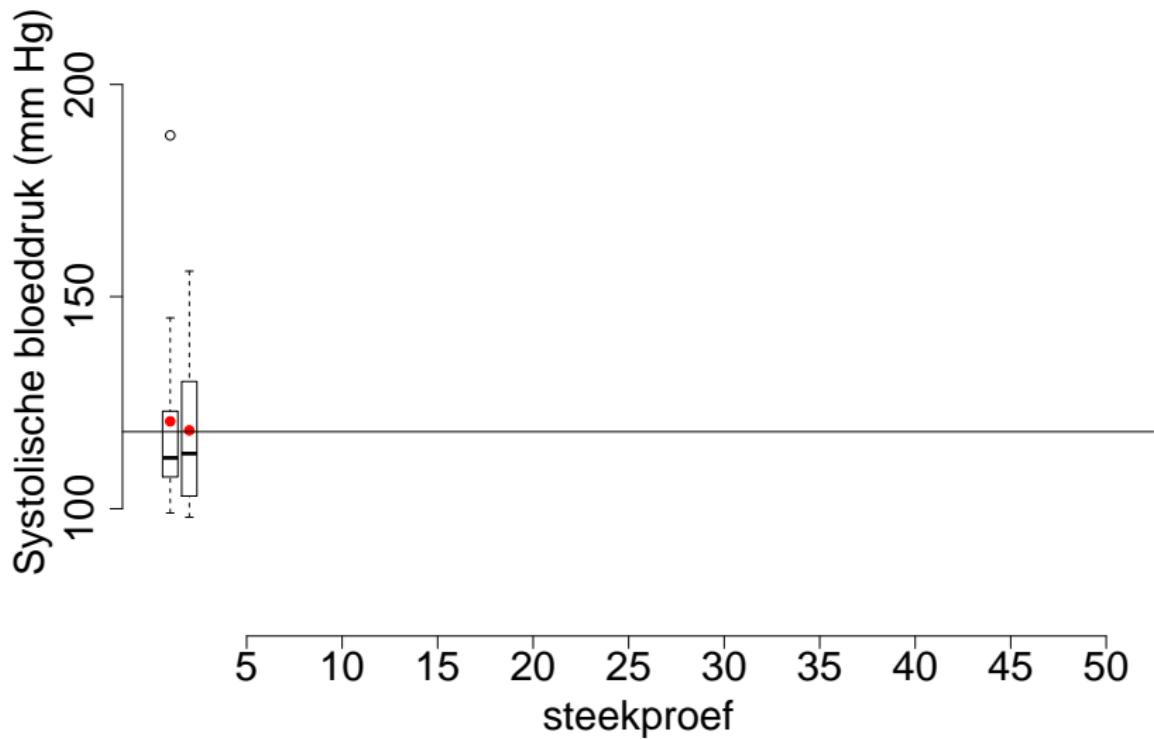
- $\bar{X}$  een *onvertekende schatter* is voor  $\mu$

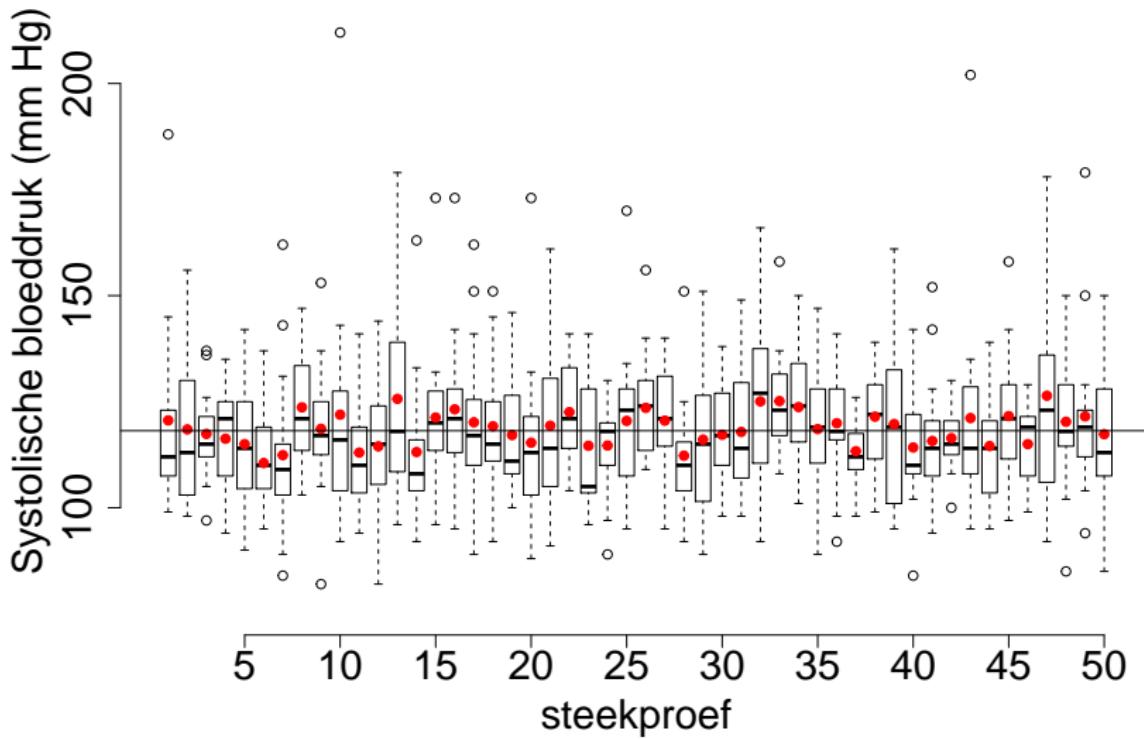
### 5.3.2. Imprecisie/standard error

- Ook voor representatieve steekproeven zijn resultaten imprecies.
- Verschillende steekproeven uit dezelfde populatie leveren immers verschillende resultaten op.
- We illustreren dit door gebruik te maken van de NHANES studie
- Stel dat we uit de Amerikaanse populatie een steekproef van 15 mensen zouden nemen om de gemiddelde bloeddruk te schatten
- We kunnen dit simuleren door uit de NHANES studie 15 personen te trekken
- We zullen dit 50 keer opnieuw herhalen en kunnen dan naar de variatie kijken van steekproef tot steekproef
- We plotten telkens een boxplot voor elke steekproef waarop het gemiddelde is aangeduid.

```
library(NHANES)
bpSamp<-matrix(NA,nrow=15,ncol=50)
for (j in 1:50)
{
  bpSamp[,j] <- sample(na.exclude(NHANES$BPSysAve),15)
  boxplot(bpSamp,ylim=range(na.exclude(NHANES$BPSysAve)))
  points(apply(bpSamp,2,mean),col=2,pch=19)
  abline(h=mean(na.exclude(NHANES$BPSysAve)))
}
```







### 5.3.2. Imprecisie/standard error

- Inzicht in hoe dicht we  $\bar{X}$  bij  $\mu$  mogen verwachten?
- Hoe zal het steekproef gemiddelde variëren van steekproef tot steekproef?
- Nood aan kennis van variabiliteit van  $\bar{X}$
- We moeten dit bepalen o.b.v. 1 steekproef!
- Daarom zullen we veronderstellingen moeten doen.
- We zullen er in het algemeen van uitgaan dat de metingen  $X_1, X_2, \dots, X_n$  afkomstig zijn van *n onafhankelijke observationele eenheden*.

- Afhankelijke gegevens worden ondermeer ook verzameld in pre-test/post-test designs en cross-over studies.
- We hadden voor de captopril studie bijvoorbeeld twee systolische bloeddrukmetingen per persoon.
- We hebben onafhankelijkheid geïntroduceerd door eerst het effect van captopril te schatten voor elke persoon door het verschil  $\Delta_{Na\text{-voor}}$  te bepalen.
- De verschillen zijn wel onafhankelijk gezien de personen zijn echter wel onafhankelijk van elkaar
- We komen hier later nog op terug

## Eigenschap

Als  $X$  en  $Y$  onafhankelijke toevalsveranderlijken zijn, dan geldt

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Algemeen (d.i. voor mogelijks afhankelijke toevalsveranderlijken  $X$  en  $Y$ ) geldt voor constanten  $a$  en  $b$ :

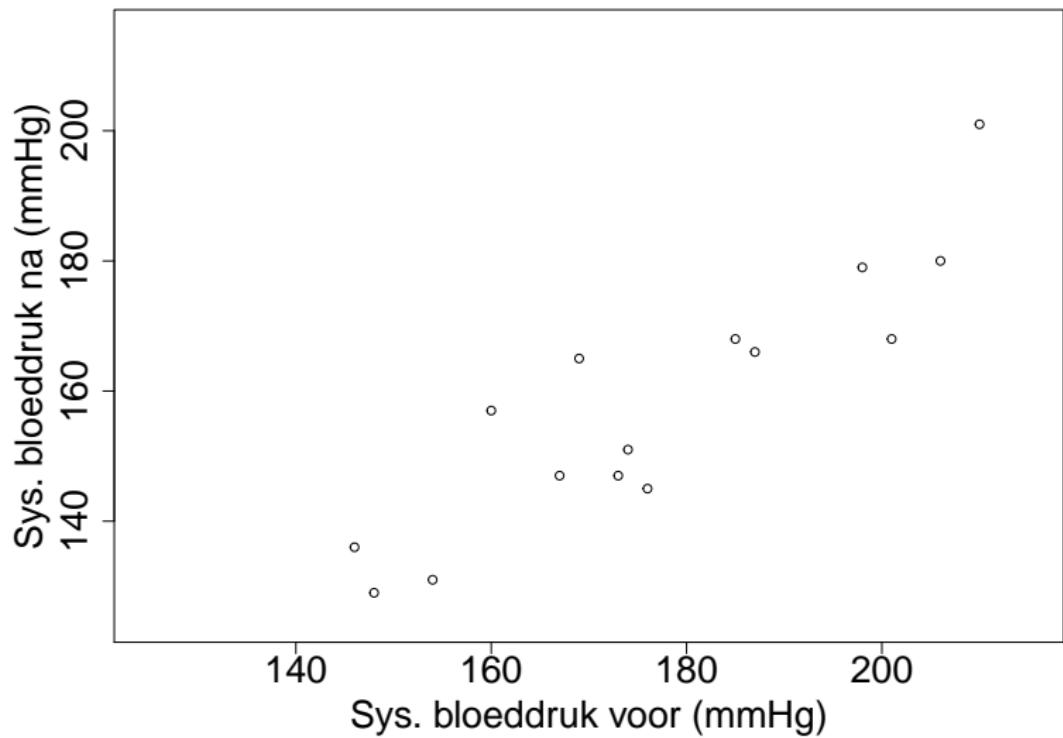
$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cor}(X, Y)\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}$$

## Einde Eigenschap

- Veel gemaakte fout:  $\text{Var}(X - Y) = \text{Var}(X) - \text{Var}(Y)$ . Niets is minder waar!
- Stel bijvoorbeeld lengte  $X$  van moeders en lengte  $Y$  van vaders evenveel variëren:  $\text{Var}(X) = \text{Var}(Y)$ .
- Variatie op verschil  $X - Y$  heeft uiteraard geen variantie van nul heeft

Bovenstaande formules geven inderdaad integendeel aan dat:

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cor}(X, Y)\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}.$$



- Berekening van variantie op  $\bar{X}$
- O.b.v. rekenregels en onafhankelijkheid (deerde overgang)

$$\begin{aligned}
 \text{Var}(\bar{X}) &= \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\
 &= \frac{\text{Var}(X_1 + X_2 + \dots + X_n)}{n^2} \\
 &\stackrel{*}{=} \frac{\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n)}{n^2} \\
 &= \frac{\sigma^2 + \sigma^2 + \dots + \sigma^2}{n^2} \\
 &= \frac{\sigma^2}{n}.
 \end{aligned}$$

- Steekproefgemiddelde heeft dus een spreiding (standaarddeviatie) rond haar gemiddelde  $\mu$  die  $\sqrt{n}$  keer kleiner is dan de deviatie op de oorspronkelijke observaties.
- We leren meer over  $\mu$  door het steekproefgemiddelde  $\bar{X}$  te observeren dan door individuele waarde  $X$  te observeren.
- Hoe meer observaties, hoe preciezer de schatting  $\bar{X}$ .

## Definitie: standaard error

De standaarddeviatie van  $\bar{X}$  is  $\sigma/\sqrt{n}$  en krijgt in de literatuur de speciale naam **standard error** van het gemiddelde. Algemeen noemt men de standaarddeviatie van een schatter voor een bepaalde parameter  $\theta$ , de **standard error** van die schatter. Men noteert dit als  $SE$ .

Stel: -  $n = 15$  systolische bloeddrukobservaties - standaarddeviatie van bloeddrukverschillen in populatie  $\sigma = 9.0$  mmHg - standard error (SE) van de systolische bloeddrukveranderingen  $\bar{X}$ :

$$SE = \frac{9.0}{\sqrt{15}} = 2.32\text{mmHg.}$$

- Meestal is  $\sigma$ , en bijgevolg de standard error van het steekproefgemiddelde, onbekend.
- Men moet dan de standard error schatten.
- schatter:  $S/\sqrt{n}$ ,
- Met  $S^2$  de *steekproefvariantie* van  $X_1, \dots, X_n$  is en  $S$  de *steekproef standaarddeviatie*.

Voor captopril:

```
n=length(delta)
se=sd(delta)/sqrt(n)
se
```

```
## [1] 2.330883
```



### 5.3.2.1. Standaarddeviatie vs standard error

#### Standaard error (s.e.)

- Vaak verwarring tussen tussen standard error en standaarddeviatie.
- Standard error (s.e.) : spreiding op geschatte parameter zoals  $\bar{X}$ .
- s.e. is m.a.w. de standaarddeviatie van het steekproefgemiddelde
- In grotere steekproeven ( $n \uparrow$ ) wordt de schatter meer precies en s.e.  $\downarrow$ .
- De standard error van schatter sterk beïnvloed door  $n$

#### Standaarddeviatie (s.d.)

- Meestal verwijst de term standaarddeviatie naar de standaarddeviatie van individuele metingen.
- De s.d. op individuele metingen neemt niet af als  $n \uparrow$ .
- Het is een parameter van de populatie, een maat voor variabiliteit van een random variabele tussen individuen
- Geeft weer hoe individuele metingen variëren rond pop. gemiddelde  $\mu$
- $n \uparrow$  s.d. meer nauwkeurig, maar blijft inzelfde grootteorde liggen.

## Illustratie via simulatie studie

- Verschillende steekproefgroottes: 15, 100, 1000
- 100 simulaties per steekproefgrootte
- Elke steekproef
  - Observaties trekken uit een normale verdeling  $N(\mu = -18.9, \sigma^2 = 9^2)$
  - Standaarddeviatie schatten:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

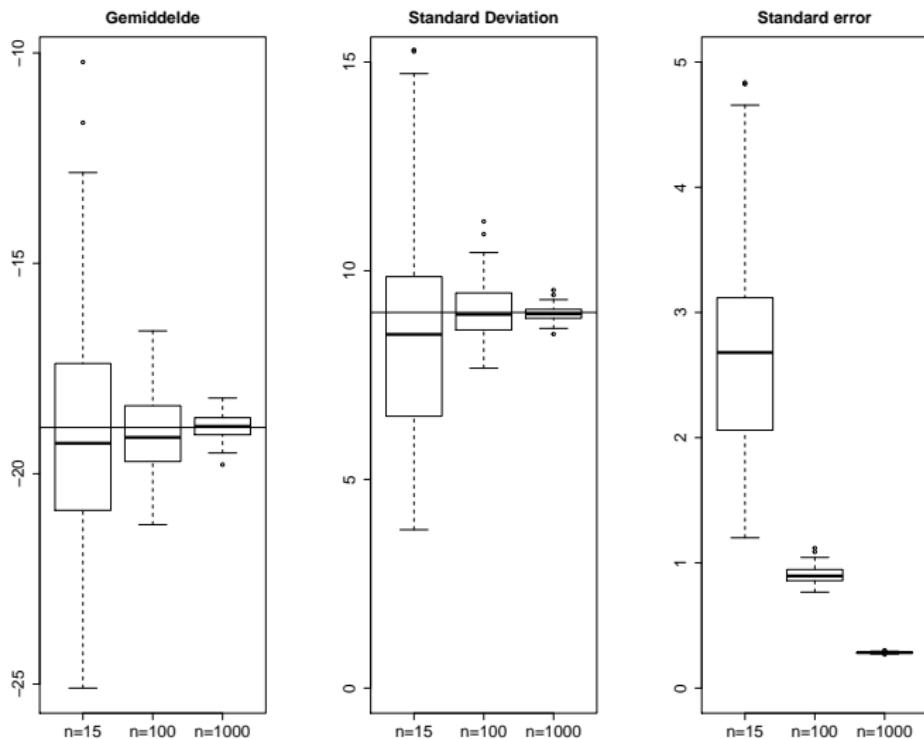
- Standaard error op gemiddelde schatten:

$$s.e = \frac{s}{\sqrt{n}}$$

- We plotten beide schatters voor alle simulaties

```
seMatrix<-matrix(0,nrow=100,ncol=3)
colnames(seMatrix)<-c("n=15","n=100","n=1000")
sdMatrix<-meanMatrix<-seMatrix

for (i in 1:100)
{
x10<-rnorm(10,mean=-18.9,sd=9)
x100<-rnorm(100,mean=-18.9,sd=9)
x1000<-rnorm(1000,mean=-18.9,sd=9)
sdMatrix[i,]<-c(sd(x10),sd(x100),sd(x1000))
seMatrix[i,]<-sdMatrix[i,]/sqrt(c(10,100,1000))
meanMatrix[i,]<-c(mean(x10),mean(x100),mean(x1000))
}
par(mfrow=c(1,3))
boxplot(meanMatrix,main="Gemiddelde")
abline(h=-18.9)
boxplot(sdMatrix,ylim=c(0,15),main="Standard Deviation")
abline(h=9)
boxplot(seMatrix,ylim=c(0,5),main="Standard error")
```



### 5.3.2.2. Geclusterde metingen

- Data in studies zijn niet altijd onafhankelijk.
- Metingen in de captopril voorbeeld zijn geclusterd: 2 systolische bloeddrukmetingen per patiënt: 1 meting voor en 1 meting na het toedienen van captopril
- Gemiddelde bloeddrukverandering  $\mu$  schatten a.d.h.v.

$$(Y_{i1}, Y_{i2}),$$

voor subjecten  $i = 1, \dots, n$ .

- Schatting

$$\bar{X} = \sum_{i=1}^n \frac{Y_{i2} - Y_{i1}}{n}$$

Uit de rekenregels voor de variantie weten we dat

$$\begin{aligned}\text{Var} [\bar{X}] &= \sum_{i=1}^n \frac{\text{Var} [Y_{i1} - Y_{i2}]}{n^2} \\ &= \sum_{i=1}^n \frac{\sigma_1^2 + \sigma_2^2 - 2\text{Cor} [Y_{i1}, Y_{i2}] \sigma_1 \sigma_2}{n^2} \\ &= \frac{\sigma_1^2 + \sigma_2^2 - 2\text{Cor} [Y_{i1}, Y_{i2}] \sigma_1 \sigma_2}{n},\end{aligned}$$

In R kunnen we dit als volgt berekenen:

```
vars=var(captopril[,c("SBPb","SBPa")])
```

```
vars
```

```
##           SBPb      SBPa
## SBPb 422.9238 370.7857
## SBPa 370.7857 400.1429
```

```
cor(captopril$SBPa,captopril$SBPb)
```

```
## [1] 0.9013312
```

```
varXbarDelta=(vars[1,1]+vars[2,2]-2*vars[1,2])/15
```

```
sqrt(varXbarDelta)
```

```
## [1] 2.330883
```

- Metingen zijn heel sterk gecorreleerd.
- Daardoor zal variantie op verschil veel lager.

## Alternatieve methode voor standard error

- Twee gecorreleerde metingen per patient tot 1 meting te reduceren.
- Merk op dat we dit enkel kunnen doen voor gepaarde metingen.
- Alle resulterende metingen zijn dan onafhankelijk. Concreet bloeddrukverschil berekenen voor elke patiënt  $i$ :

$$X_i = Y_{ai} - Y_{bi}$$

en vervolgens standard error op  $\bar{X}$ . In het captopril voorbeeld :

```
sd(delta) / sqrt(15)
```

```
## [1] 2.330883
```

- Exactzelfdeschatting voor de standard error
- Groot voordeel van design:
- Bloeddrukmetingen voor en na het toedienen van captopril sterk positief gecorreleerd → variantie op verschil veel lager dan deze op de originele bloeddrukmetingen.
- Iedere patiënt in de studie dient immers als zijn eigen controle
- op die manier kunnen we de variabiliteit in de bloeddrukmetingen tussen patiënten uit de analyse verwijderen!

### 5.3.2.3. Normaal verdeelde gegevens

- Voor Normaal verdeelde data meerdere onvertekende schatters voor het populatiegemiddelde  $\mu$  b.v.:
  - steekproefgemiddelde
  - mediaan.
- Maar steekproefgemiddelde  $\bar{X}$  is onvertekende schatter voor  $\mu$  met kleinste standard error.
- $\bar{X}$  wijkt gemiddeld minder af van  $\mu$  dan mediaan, die veel meer varieert van steekproef tot steekproef.
- Het steekproefgemiddelde is schatter die onvertekend en meest precies is (kleinste standaarddeviatie).

### 5.3.3. Verdeling van het steekproefgemiddelde

- Hoe varieert  $\bar{X}$  van steekproef tot steekproef?
- Verdeling van  $\bar{X}$  kennen
- Als  $\bar{X}$  Normaal verdeeld zijn, dan is betekenis van de standard error goed te bevatten:
  - s.e. is immers standaarddeviatie van steekproefgemiddelde.

$$X_i \sim N(\mu, \sigma^2) \rightarrow \bar{X} \sim N(\mu, \sigma^2/n)$$

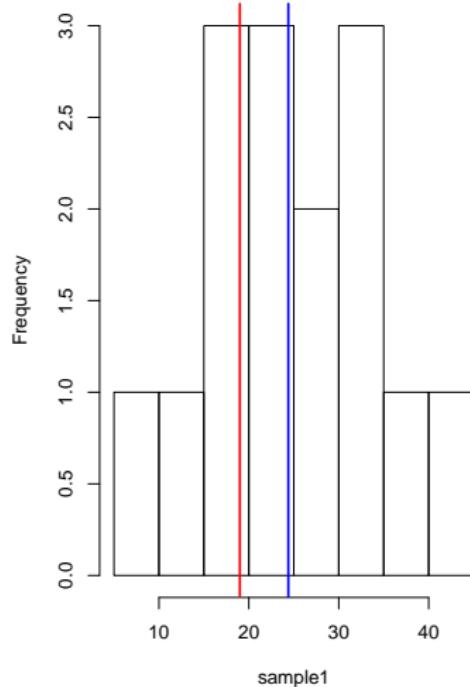
## We tonen dit opnieuw via simulatie

- Steekproeven met  $n=15$  uit  $N(\mu = -19, \sigma^2 = 9^2)$

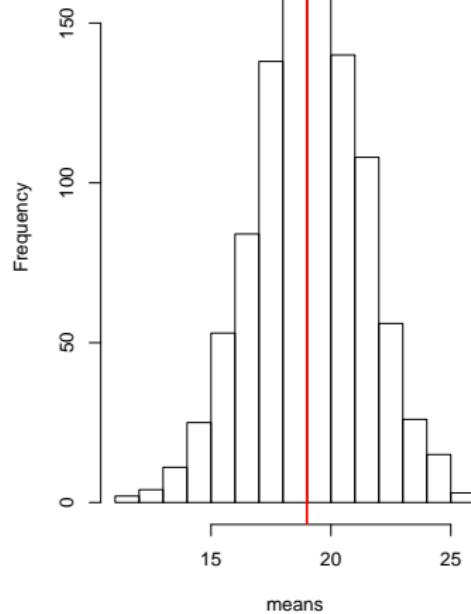
```
par(mfrow=c(1,2))
sample1 <- rnorm(n=15,mean=19,sd=9)
hist(sample1)
abline(v=mean(sample1),col="blue",lwd=2)
abline(v=19,col="red",lwd=2)

#1000 simulations
nSim <- 1000
means <- array(1000)
for (i in 1:1000)
{
  sample1 <- rnorm(n=15,mean=19,sd=9)
  means[i]=mean(sample1)
}
hist(means)
abline(v=19,col="red",lwd=2)
```

Histogram of sample1



Histogram of means



- In captopril is systolische bloeddrukverandering approximatif Normaal verdeeld.
- $s.e. = 2.32 \text{ mm Hg}$
- Op 100 studies met  $n = 15$  subjecten verwachten we dat de geschatte gemiddelde systolische bloeddrukafwijking ( $\bar{X}$ ) op minder dan  $2 \times 2.32 = 4.64 \text{ mm Hg}$  van het werkelijke populatiegemiddelde ( $\mu$ ) ligt in 95 studies.

- Wanneer individuele observaties  $X_i$  geen Normale verdeling hebben, is  $\bar{X}$  bij benadering toch nog Normaal verdeeld zodra het aantal observaties groot genoeg is.
- Hoe groot de steekproef moet zijn voor Normale benadering werkt?
- Hangt af van hoe scheef verdeling is

## De Centrale Limietstelling (CLT)

Stel dat  $X_1, X_2, \dots, X_n$ ,  $n$  onafhankelijke lukrake trekkingen van de toevalsveranderlike  $X$  voorstellen, met allen dezelfde theoretische verdeling. Laat  $X$  gemiddelde  $\mu$  en variantie  $\sigma^2$  hebben maar verder een ongespecificeerde verdeling, dan wordt de verdeling van het steekproefgemiddelde  $\bar{X}_n = \sum_{i=1}^n X_i/n$  naarmate  $n$  groter wordt steeds beter benaderd door de Normale verdeling met gemiddelde  $\mu$  en variantie  $\sigma^2/n$ .

## Einde Stelling

- De CLT zal toelaten om meeste technieken van deze cursus toe te passen op zeer uitgebreid spectrum van experimenten!

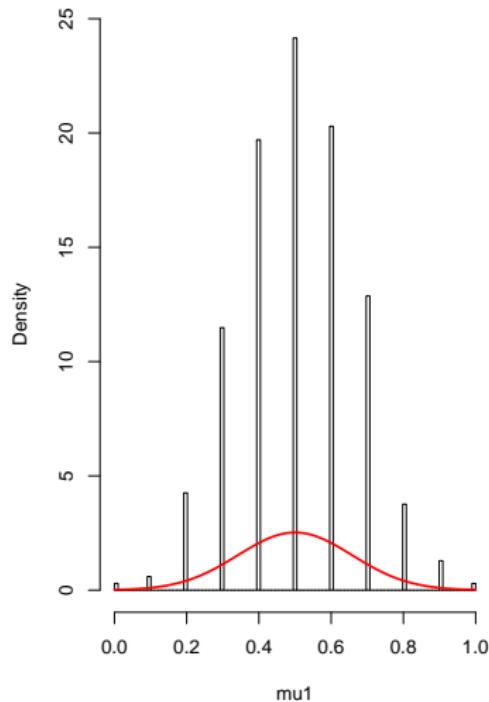
## Illustratie van de CLT a.d.h.v. simulatie studie

- Simulatie van experiment waarbij we een munt opwerpen
- Data zijn Bernoulli verdeeld:  $X = 0$  (munt) of  $X = 1$  (kop)
- Bernoulli heeft 1 parameter: de kans op succes ( $X = 1$ )  $\pi = .5$  voor niet gebiased muntstuk.
- Heel duidelijk dat data niet normaal verdeel zijn: ze zijn discreet!
- Simulatie van duizend experimenten met steekproefgrootte 10 en steekproefgrootte 100

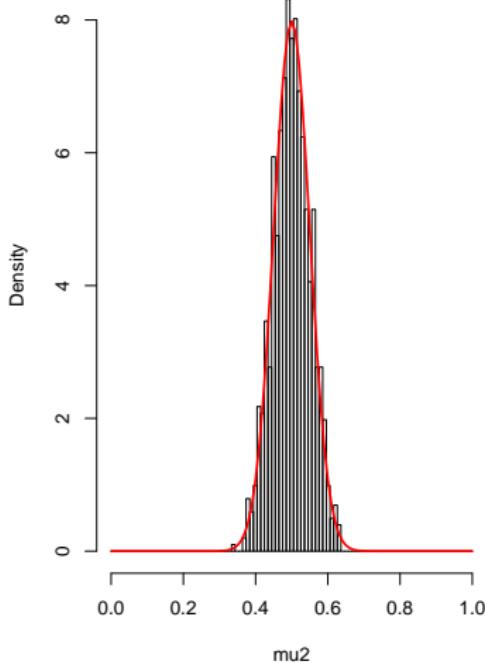
```
par(mfrow=c(1,2))
set.seed(113)
nObs <- 10
Nsim <- 1000
pi <- 0.5
mu1 <- array(dim=Nsim)
nBin <- 100
for (i in 1:Nsim)
{
  x <- rbinom(n=nObs,pi,size=1)
  mu1[i] <- mean(x)
}
hist(mu1, breaks=seq(0,1,length=nBin), main= paste("Steekproef groot"))
grid <- seq(0,1,.01)
lines(grid,dnorm(grid,mean=.5, sd=.5/sqrt(nObs)), col=2, lwd=2)

nObs <- 100
mu2 <- array(dim=Nsim)
for (i in 1:Nsim)
{
```

Steekproefgrootte 10



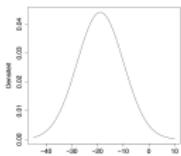
Steekproefgrootte 100



## Overzicht

- Proefopzet/Data Exploratie
- Puntschatters (Schatten)
- Intervalschatters (Statistische Besluitvorming)
- Hypothese testen (Statistische Besluitvorming)

## POPULATIE

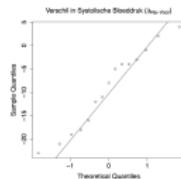


$\Delta_{\text{na-voor in populatie}}$

$$\hat{\mu}_{\Delta} = -18.9$$

PROEFOPZET

## SAMPLE



SCHATTEN &  
INFERENTIE

$\Delta_{\text{na-voor in sample}}$

$$\bar{x}_{\Delta} = -18.9, SD_{\Delta} = 9.3, SE_{\Delta} = 2.4$$

DATA EXPL. & BESCHR. STAT.

## 5.4. Intervalschatters

- $\bar{X}$  varieert rond  $\mu$  die we wensen te schatten
- In deze sectie: interval rond het  $\bar{X}$  waarbinnen we o.b.v. de data  $\mu$  met gegeven kans (bvb. 95% kans) kunnen verwachten.
- Eerst werken we dit uit in veronderstelling dat  $\sigma^2$  gekend is
- Daarna zullen we van deze onrealistische veronderstelling afstappen gezien de populatievariantie bijna is nooit gekend

### 5.4.1. Gekende variantie op de metingen

- $X \sim N(\mu, \sigma^2) \rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- 95% referentie-interval voor het steekproefgemiddelde:

$$\left[ \mu - 1.96 \frac{\sigma}{\sqrt{n}}, \mu + 1.96 \frac{\sigma}{\sqrt{n}} \right]$$

- Het bevat met 95% kans het steekproefgemiddelde van een lukrake steekproef.
- Kan niet worden berekend want  $\mu$  is onbekend.
- Schatten door  $\mu$  te vervangen door  $\bar{X}$ .

$$\left[ \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right]$$

$\left[ \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right]$  is referentie-interval voor het steekproefgemiddelde

- Nuttigere interpretatie:
- ongelijkheid  $\mu - 1.96 \sigma / \sqrt{n} < \bar{X}$  herschrijven als  $\mu < \bar{X} + 1.96 \sigma / \sqrt{n}$ .

Hieruit volgt:

$$\begin{aligned} 95\% &= P(\mu - 1.96 \sigma / \sqrt{n} < \bar{X} < \mu + 1.96 \sigma / \sqrt{n}) \\ &= P(\bar{X} - 1.96 \sigma / \sqrt{n} < \mu < \bar{X} + 1.96 \sigma / \sqrt{n}) \end{aligned}$$

## Definitie 95% betrouwbaarheidsinterval voor populatiegemiddelde

### Het interval

$$[\bar{X} - 1.96 \sigma / \sqrt{n}, \bar{X} + 1.96 \sigma / \sqrt{n}], (\#eq : bi) \quad (1)$$

bevat met 95% kans het populatiegemiddelde  $\mu$ . Het wordt een **95% betrouwbaarheidsinterval** (in het Engels: *95% confidence interval*) voor het populatiegemiddelde  $\mu$  genoemd. De kans dat het de populatieparameter  $\mu$  bevat, d.i. 95%, wordt het *betrouwbaarheidsniveau* genoemd.

Een 95% betrouwbaarheidsinterval bepaalt met andere woorden een reeks waarden waarbinnen de gezochte populatieparameter *waarschijnlijk* (namelijk met 95% kans) valt.

Stel

- een steekproef met een bloeddrukdaling van  $-18.93\text{mmHg}$
- standaarddeviatie van de bloeddrukdalingen is gekend en bedraagt  $9\text{mmHg}$
- $[-18.93 - 1.96 \times 9/\sqrt{15}, -18.93 + 1.96 \times 9/\sqrt{15}] = [-23.48, -14.38]\text{mmHg}$ .

- Merk op dat we over “95% kans” spreken omdat eindpunten van het 95% betrouwbaarheidsinterval toevalsveranderlijken zijn die variëren van steekproef tot steekproef.
- M.a.w verschillende steekproeven leveren telkens andere betrouwbaarheidsintervallen op, vermits die intervallen berekend zijn op basis van de gegevens in de steekproef.
- Het zijn dus *stochastische intervallen*.
- Voor 95% van alle steekproeven zal 95% betrouwbaarheidsinterval de werkelijke waarde van de gezochte populatieparameter bevatten
- voor de overige 5% niet.
- Uit een gegeven betrouwbaarheidsinterval kan js niet besluiten of het de gezochte parameterwaarde bevat of niet, vermits ze onbekend is.
- Maar we werken met een methode die in 95% van de gevallen de gezochte waarde zal bevatten.

- Meestal is variantie niet gekend en kan ze worden geschat a.d.h.v.  $S^2$
- Als  $n$  groot is kan men aantonen dat  $[\bar{X} - 1.96 s/\sqrt{n}, \bar{X} + 1.96 s/\sqrt{n}]$  het populatiegemiddelde met bij benadering 95% kans bevat
- Om ander betrouwbaarheidsniveau te gebruiken vervangt men 1.96 door het relevante kwantiel  $z_{\alpha/2}$ .
- Betrouwbaarheidsinterval wordt niet alleen gebruik voor gemiddelde maar ook voor andere populatieparameters

**Definitie Betrouwbaarheidsinterval** Een  $(1 - \alpha)100\%$  **betrouwbaarheidsinterval** voor een populatieparameter  $\theta$  is een geschat (en bijgevolg stochastisch) interval dat met  $(1 - \alpha)100\%$  kans de echte waarde van die populatieparameter  $\theta$  bevat.

- Voor het steekproefgemiddelde bekomen we dus dus

$$[\bar{X} - z_{\alpha/2} \sigma / \sqrt{n}, \bar{X} + z_{\alpha/2} \sigma / \sqrt{n}]$$

- Wat beïnvloedt de breedte van het interval?

## 5.4.2. Ongekende variantie op de metingen

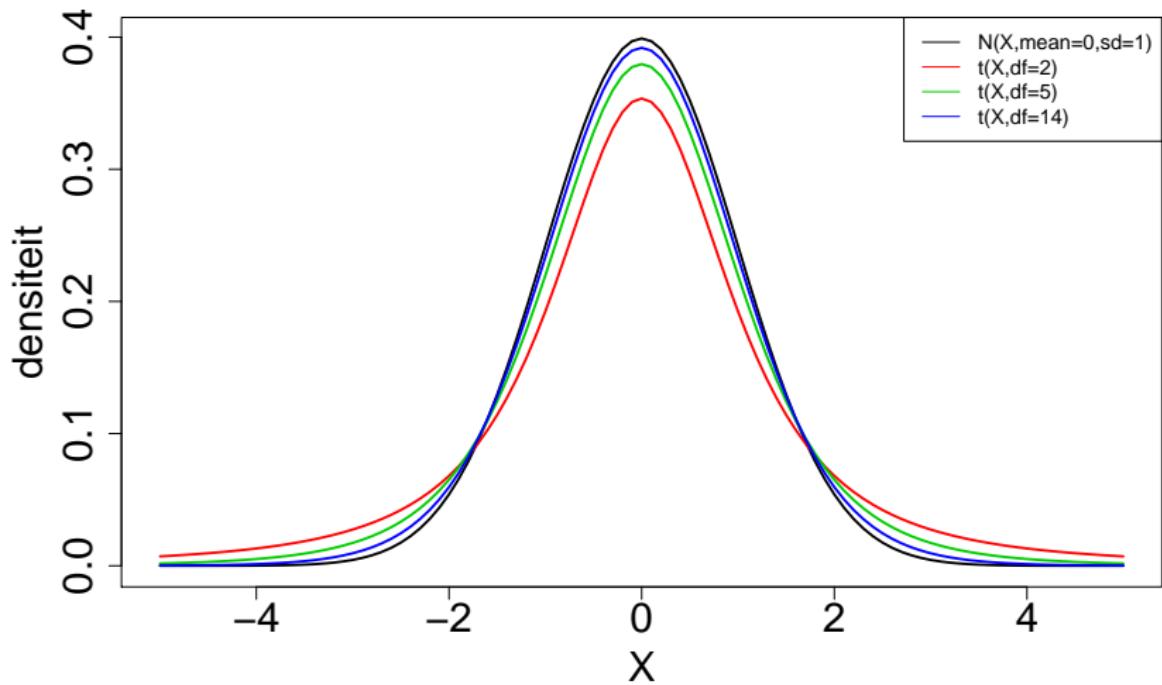
- Tot nog toe werd verondersteld dat de populatievariantie  $\sigma^2$  gekend is.
- In de praktijk komt het quasi nooit voor dat men de populatievariantie  $\sigma^2$  exact kent.
- $\sigma^2$  wordt geschat o.b.v.  $S^2$  a.d.h.v. de steekproef.
- Vorige Intervallen iets te smal (geen rekening met onzekerheid op schatter voor variantie)
- Voor relatief kleine steekproeven hangt verdeling van  $(\bar{X} - \mu)/(S/\sqrt{n})$  af van de grootte  $n$  van de steek
- Wanneer steekproef voldoende groot is, ligt  $S$  voldoende dicht bij  $\sigma$  (zie simulatiestudie voor  $S$  met 1000 observaties)
- $(\bar{X} - \mu)/(S/\sqrt{n})$  volgt bij benadering een standaardnormale verdeling, bijgevolg is

$$\left[ \bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

- een benaderd  $(1 - \alpha)100\%$  betrouwbaarheidsinterval is voor  $\mu$ .
- Voor kleine steekproeven is dit niet langer het geval.

- Bij kleine steekproeven introduceert schatting van  $S$  extra onnauwkeurigheid in de gestandaardiseerde waarde  $(\bar{X} - \mu)/(S/\sqrt{n})$ .
- Deze is nog wel gecentreerd rond nul en symmetrisch, maar niet langer Normaal verdeeld.
- De echte verdeling voor eindige steekproefgrootte  $n$  heeft zwaardere staarten dan de Normale.
- Hoeveel zwaarder de staarten zijn, hangt van de steekproefgrootte  $n$  af.
- Ze krijgt de naam (Student)  $t$ -verdeling met  $n - 1$  vrijheidsgraden (in het Engels: *degrees of freedom*).

```
grid <- seq(-5,5,.1)
plot(grid,dnorm(grid),ylab="densiteit",xlab="X",type="l",lwd=2)
dfs <- c(2,5,14)
for (i in 1:length(dfs))
  lines(grid,dt(grid,dfs[i]),col=i+1,lwd=2)
legend("topright",lty=1,col=1:4,legend=c("N(X,mean=0,sd=1)",past
```



- t-verdelingen hebben bredere staarten dan Normale  $\rightarrow$  grotere percentielwaarden en dus bredere intervallen voor zelfde betrouwbaarheidsniveau.
- Bouwt extra onzekerheid in gerelateerd aan schatting van  $S$

**Definitie t-verdeling** Als  $X_1, X_2, \dots, X_n$  een steekproef vormen uit de Normale verdeling  $N(\mu, \sigma^2)$ , dan is  $(\bar{X} - \mu)/(S/\sqrt{n})$  verdeeld als een  $t$ -verdeling met  $n - 1$  vrijheidsgraden.

- Percentielen van  $t$ -verdeling in Tabellen of berekenen in R.
- Onderstaande code: 95%, 97.5%, 99.5% percentiel voor  $t$ -verdeling met 14 vrijheidsgraden
- bruikbaar voor berekening van 90%, 95% en 99% betrouwbaarheidsintervallen.

```
qt(.975,df=14)
```

```
## [1] 2.144787
```

```
qt(c(.95,.975,.995),df=14)
```

```
## [1] 1.761310 2.144787 2.976843
```

- 97.5% percentiel 2.14 voor een  $t$ -verdeling met  $n - 1 = 14$  vrijheidsgraden inderdaad groter dan kwantiel uit de Normaal verdeling 1.96.

Analoog als bij Normaal verdeling met gekende variatie wordt  $100\%(1 - \alpha)$  betrouwbaarheidsinterval voor het gemiddelde  $\mu$  van een Normaal verdeelde veranderlijke  $X$  met onbekende variantie

$$\left[ \bar{X} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \right]$$

$(1 - \alpha/2)100\%$  percentiel van de Normale verdeling wordt vervangen door  $(1 - \alpha/2)100\%$  percentiel van de t-verdeling met  $n - 1$  vrijheidsgraden.

95% BI voor gemiddelde bloeddrukverandering

```
mean(delta) - qt(.975,df=14)*sd(delta)/sqrt(n)
```

```
## [1] -23.93258
```

```
mean(delta) + qt(.975,df=14)*sd(delta)/sqrt(n)
```

```
## [1] -13.93409
```

99% BI voor gemiddelde bloeddrukverandering

```
mean(delta) - qt(.995,df=14)*sd(delta)/sqrt(n)
```

```
## [1] -25.87201
```

```
mean(delta) + qt(.995,df=14)*sd(delta)/sqrt(n)
```

```
## [1] -11.99466
```



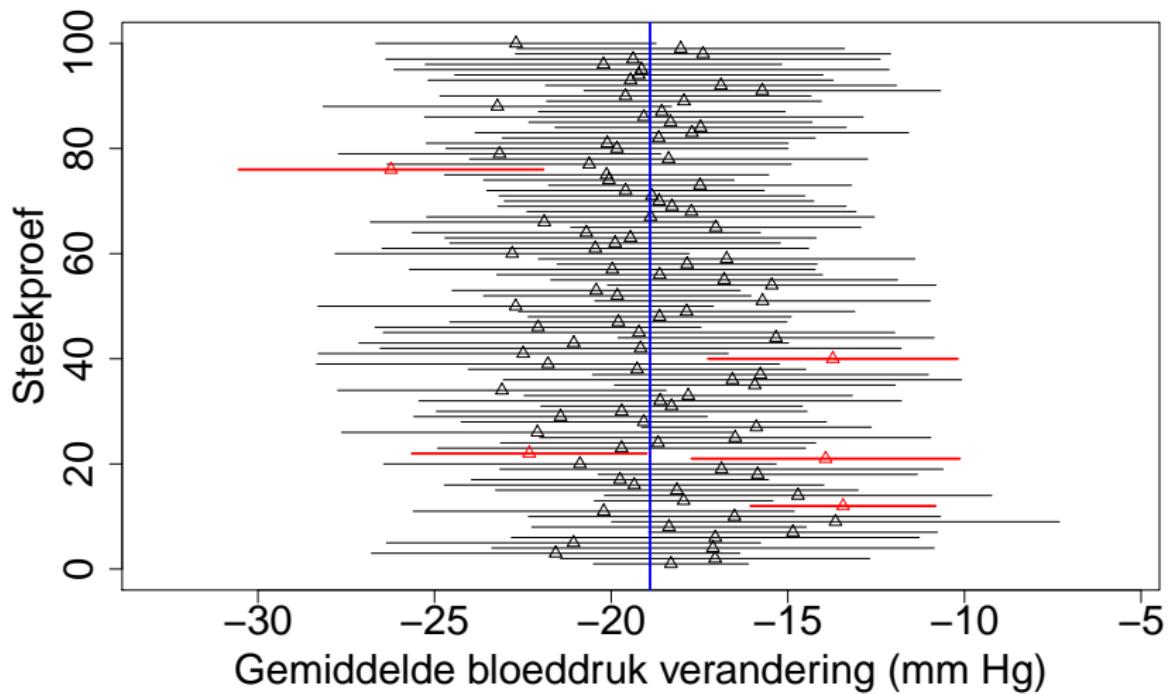
### 5.4.3. Interpretatie van betrouwbaarheidsintervallen

- Via simulatie
- Stel dat gemiddelde bloeddrukdaling in de populatie  $X \sim N(\mu = -18.9, \sigma^2 = 9.0^2)$ .
- 1000 simulaties met steekproef  $n = 15$
- We houden het gemiddelde, de ondergrens en bovengrens van het BI en of het BI het werkelijke gemiddelde bevat bij.

```
set.seed(115)
mu <- -18.9
sigma <- 9.0
nSim <- 1000
alpha <- 0.05
n <- 15
muHat <- sigmaHat <- BI.ondergrens <- BI.bovengrens <- omvat <-
cnt<-0
for(i in 1:nSim) {
  y<-rnorm(n,mean=mu,sd=sigma)
  muHat[i]<-mean(y)
  sigmaHat[i]<-sd(y)/sqrt(n)
  BI.ondergrens[i]<-muHat[i]-qt(1-alpha/2,df=n-1)*sigmaHat[i]
  BI.bovengrens[i]<-muHat[i]+qt(1-alpha/2,df=n-1)*sigmaHat[i]
  omvat[i]<-(mu<BI.bovengrens[i])&(BI.ondergrens[i]<mu)
  cnt<-cnt+as.numeric(omvat[i])
}
cnt/nSim
```

```
## [1] 0.951
```

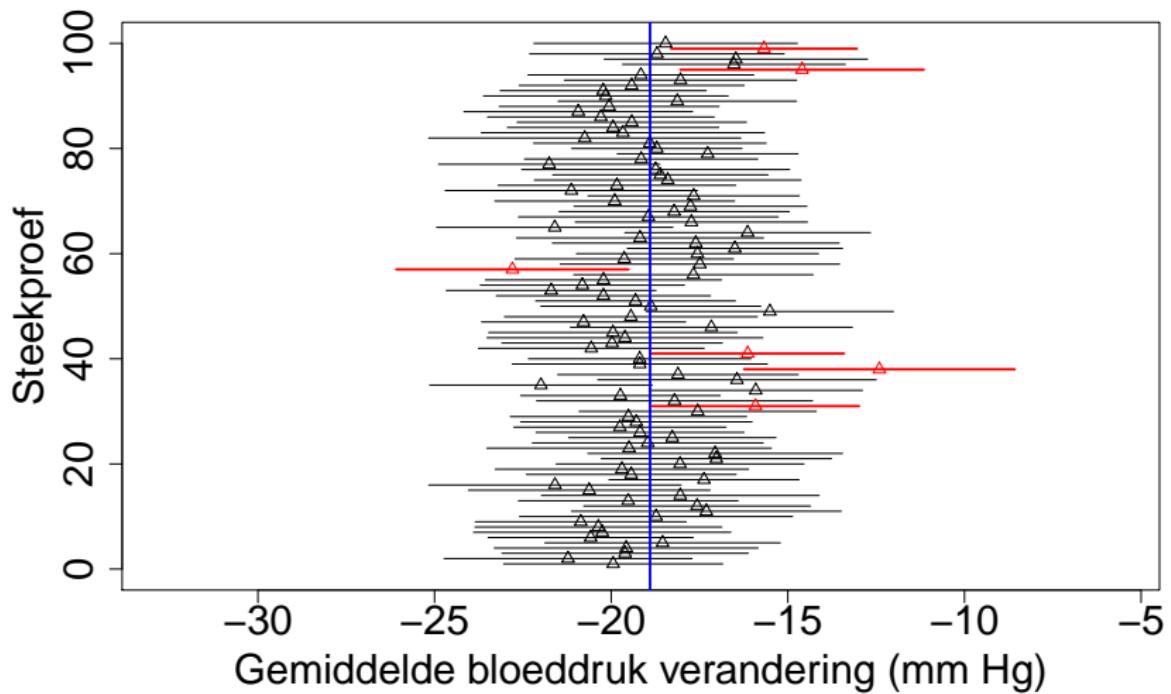




## Simulatie opnieuw maar $n = 30$

```
n <- 30
muHat <- sigmaHat <- BI.ondergrens <- BI.bovengrens <- omvat <-
cnt<-0
for(i in 1:nSim) {
  y<-rnorm(n,mean=mu,sd=sigma)
  muHat[i]<-mean(y)
  sigmaHat[i]<-sd(y)/sqrt(n)
  BI.ondergrens[i]<-muHat[i]-qt(1-alpha/2,df=n-1)*sigmaHat[i]
  BI.bovengrens[i]<-muHat[i]+qt(1-alpha/2,df=n-1)*sigmaHat[i]
  omvat[i]<-(mu<BI.bovengrens[i])&(BI.ondergrens[i]<mu)
  cnt<-cnt+as.numeric(omvat[i])
}
cnt/nSim

## [1] 0.949
```

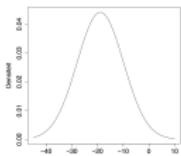


#### 5.4.4. Wat rapporteren?

- Rapporteer dus zeker steeds onzekerheid op de resultaten!
- Conclusies trekken o.b.v. 1 schatting kan zeer misleidend zijn!
- In statistische analyses rapporteert men daarom systematisch betrouwbaarheidsintervallen.
- Betrouwbaarheidsintervallen vormen een goed compromis
- ze zijn smal genoeg om informatief te zijn, maar haast nooit zeer misleidend.
- Wetenschappelijke relevantie en statistische significantie
- We besluiten dat de parameter die ons interesseert in het 95% betrouwbaarheidsinterval zit, en weten dat die uitspraak met 95% kans correct is.
- In de statistiek trekt men dus nooit absolute conclusies.

Op basis van de data-analyse voor het captopril voorbeeld kunnen we dus besluiten dat de gemiddelde bloeddrukdaling 18.9mmHg bedraagt na het toedienen van captopril. Met een 95% betrouwbaarheidsinterval op het gemiddelde van [-22.3,-15.6]mmHg. Op basis van het betrouwbaarheidsinterval is het duidelijk dat het toedienen van captopril resulteert in een sterke bloeddrukdaling bij patiënten met hypertensie.

## POPULATIE

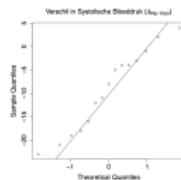


$\Delta_{\text{na-voor in populatie}}$

$$\hat{\mu}_{\Delta} = -18.9, 95\% \text{ CI: } [-23.9, -13.9]$$

PROEFOPZET

## SAMPLE



$\Delta_{\text{na-voor in sample}}$

$$\bar{x}_{\Delta} = -18.9, SD_{\Delta} = 9.3, SE_{\Delta} = 2.4$$

DATA EXPL. & BESCHR. STAT.



SCHATTEN &  
INFERENTIE

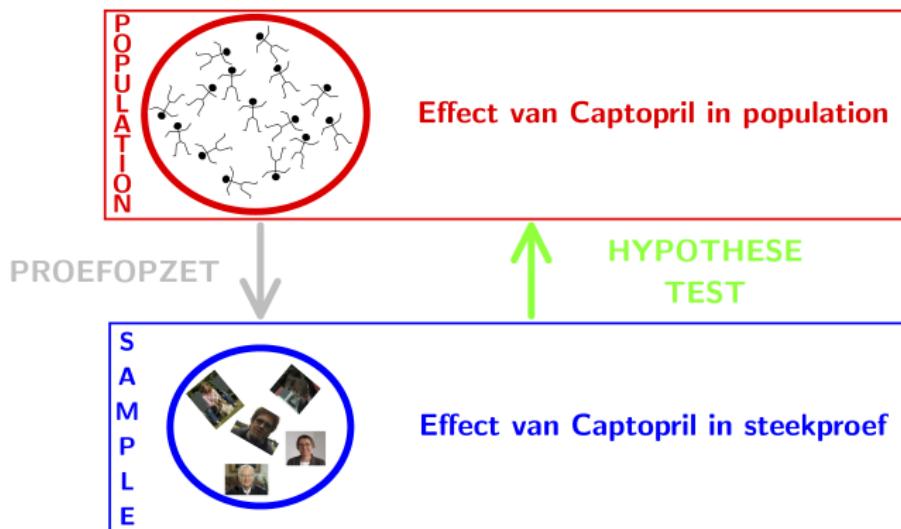
## Overzicht

- Proefopzet/Data Exploratie
- Puntschatters (Schatten)
- Intervalschatters (Statistische Besluitvorming)
- Hypothese testen (Statistische Besluitvorming)

## 5.5. Principe van Hypothesetoetsen (via one sample t-test)

Captopril voorbeeld: Onderzoekers wensen na te gaan of medicijn

Captopril een bloeddruk verlagend effect heeft

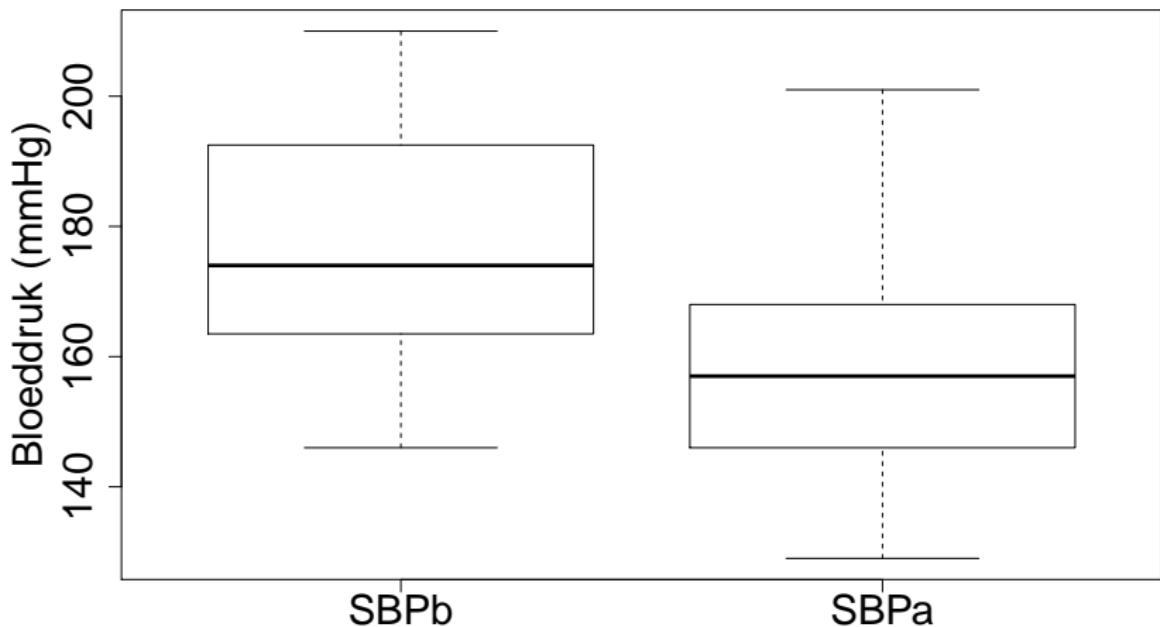


In wetenschappelijk onderzoek wenst men vaak op basis van empirische data een ja/nee antwoord te geven.

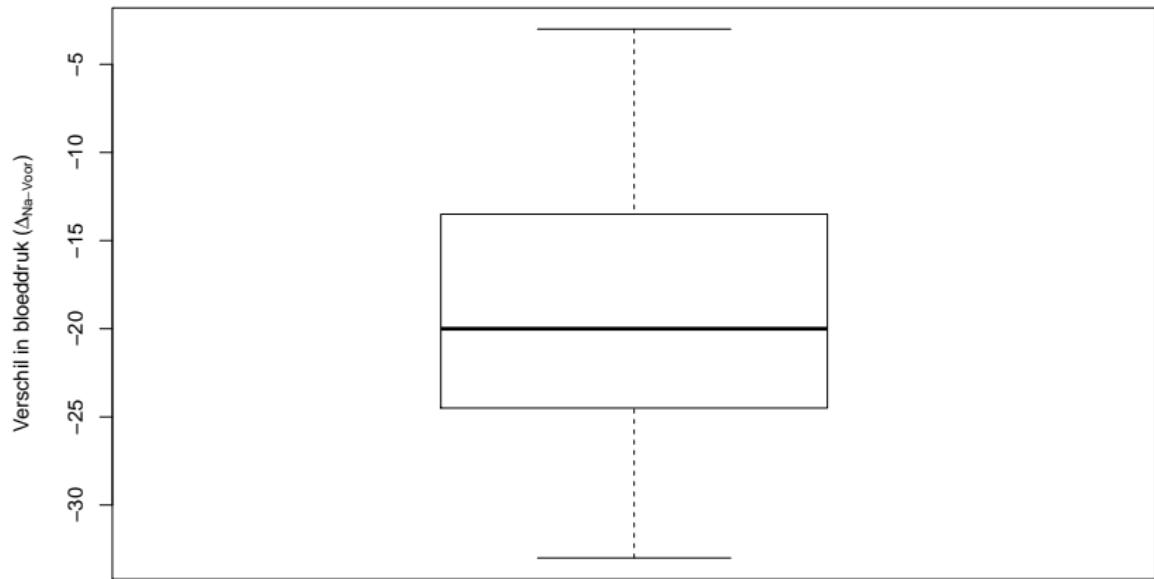
Voorbeeld: Captopril

- Is er geen/wel een effect van het toedienen van Captopril op de systolische bloeddruk?
- Beslissen op basis van gegevens is niet evident.
- Er is immers onzekerheid of de bevindingen uit de steekproef generaliseerbaar zijn.
- Is een schijnbaar gunstig effect systematisch of toevallig?

## Captopril: Systolische Bloeddruk



## Captopril: Verschil in systolische bloeddruk



- Een natuurlijke beslissingsbasis is het gemiddeld verschil  $X$  in de systolische bloeddruk:

$$\bar{x} = -18.93 \text{ mmHg} \ (s = 9.03, SE = 2.33).$$

- Volstaat niet dat  $\bar{x} < 0$  om te beslissen dat gemiddelde systolische bloeddruk lager is na het toedienen van captopril *op het niveau van de volledige populatie*.
- Om het effect die we in de steekproef observeren te kunnen *veralgemenen* naar de populatie moet de bloeddrukverlaging voldoende groot zijn.
- Maar hoe groot moet dit effect nu zijn?

- Om uitsluitsel te geven werden *statistische toetsen* ontwikkeld.
- Deze leveren een ja/nee antwoord
- Tegenwoordig is het haast onmogelijk om een wetenschappelijk onderzoeksartikel te lezen zonder de resultaten van dergelijke toetsen te ontmoeten.

- Volgens het *falsificatieprincipe* van Popper kan je nooit een hypothese bewijzen op basis van data.
- Daarom zullen we twee hypotheses introduceren: een nulhypothese en een alternatieve hypothese.
- We zullen dan later a.d.h.v. de toets de nulhypothese trachten te ontkrachten.

## 5.5.1 Hypotheses

- Vertaling van wetenschappelijke vraagstelling naar een nulhypothese ( $H_0$ ) en een alternatieve hypothese ( $H_1$ )
- Eerst moet de probleemstelling vertaald worden naar een geparametriserd statistisch model.
- Uit de proefopzet volgt dat

$$X_1, \dots, X_n \text{i.i.d} f(X),$$

met  $f(X)$  de dichtheidsfunctie van de bloeddrukverschillen.

- **Vereenvoudiging:** veronderstel dat  $f(X)$  gekend is op een eindig-dimensionale set van parameters  $\theta$  na (parametrisch statistisch model).

- Voor het captopril  $X \sim N(\mu, \sigma^2)$  met parameters  $\theta = (\mu, \sigma^2)$ , het gemiddelde  $\mu$  en variantie  $\sigma^2$ .
- De vraagstelling is geformuleerd in termen van de gemiddelde bloeddrukdaling:  $\mu = E_f[X]$ .
- De **alternatieve hypothese** wordt geformuleerd in termen van een parameter van  $f(X)$  en dient uit te drukken wat de onderzoekers wensen te bewijzen aan de hand van de studie.
- Hier:

$$H_1 : \mu < 0.$$

Gemiddeld gezien daalt de bloeddruk bij patiënten met hypertensie na toediening van captopril.

- De **nulhypothese** is meestal een uitdrukking van de nultoestand, i.e. de omstandigheden waarin niets bijzonders aan de hand is.
- De onderzoekers wensen meestal te bewijzen via empirisch onderzoek dat de nulhypothese niet waar is: **Falsificatie principe**.
- De **nulhypothese wordt veelal uitgedrukt door gebruik te maken van dezelfde parameter als deze die in  $H_1$  gebruikt is.**
- Hier:

$$H_0 : \mu = 0$$

m.a.w. gemiddeld gezien blijft de systolische bloeddruk na toediening van captoril onveranderd.

## 5.5.2. Test-statistiek

Eens de populatie, de parameters en de nulhypothese en alternatieve hypothese bepaald zijn, is de basisgedachte van een hypothesetest als volgt :

Construeer een teststatistiek zodanig dat deze

- ➊ de evidentie meet die aanwezig is in de steekproef,
- ➋ tegen de gestelde nulhypothese,
- ➌ ten voordele van de alternatieve hypothese.

Een teststatistiek is dus noodzakelijk een functie van de steekproefobservaties.

Voor het captopril voorbeeld drukt de statistiek

$$T = \bar{X} - \mu_0$$

uit hoever het steekproefgemiddelde van de bloeddrukalding ligt van het gemiddelde  $\mu_0 = 0$  in de populatie onder de nulhypothese.

- Als  $H_0$  waar is en er geen effect is van captopril in de populatie, dan verwachten we dat  $T = 0$
- Als  $H_1$  waar is, dan verwachten we dat  $T < 0$ .

In de praktijk niet alleen de grootte van het effect in rekening brengen maar ook de onzekerheid op het effect.

Balanceer effectgrootte t.o.v. de standard error.

$$T = \frac{\bar{X} - 0}{\text{SE}_{\bar{X}}}$$

Waarbij  $\mu_0 = 0$  voor het captopril voorbeeld.

Opnieuw geldt dat

- Als  $H_0$  waar is en er dus geen effect is van captopril in de populatie, dan verwachten we dat de teststatistiek  $T$  dicht ligt bij  $T = 0$
- Als  $H_1$  waar is, dan verwachten we dat  $T < 0$ .
- Voor het captopril voorbeeld vinden we  $t = (-18.93 - 0)/2.33 = -8.12$ .
- Is  $t = -8.12$  groot genoeg in absolute waarde om te kunnen besluiten dat  $\mu < 0$  en met welke zekerheid kunnen we dit besluiten?

- Teststatistiek  $T$  daarom verder bestuderen
- $T$  is een toevalsveranderlike
- verdeling van  $T$  hangt af van de verdeling van de steekproefobservaties
- die verdeling is ongekend!
- Normaliteit verondersteld, maar het gemiddelde en de variantie onbepaald.
- Bovendien is hypothesetest geconstrueerd om uitspraak te doen over het gemiddelde  $\mu$ !
- Oplossing:  $H_0 : \mu = 0$ : geen effect van captopril.
- Onder  $H_0$  is het gemiddelde van de normale distributie gekend!
- Als bloeddrukverschillen  $X_1, \dots, X_{15}$  onafhankelijk en i.i.d. zijn:

$$\bar{X} \stackrel{H_0}{\sim} N(0, \sigma^2/n)$$

- Gezien we  $\sigma^2$  niet kennen kunnen we deze vervangen door de steekproef variantie.

$$T = \frac{\bar{X} - 0}{\text{SE}_{\bar{X}}} \stackrel{H_0}{\sim} t(n-1)$$

- T volgt dus een t-verdeling met  $n-1$  vrijheidsgraden onder de **nulhypothese**.
- Als  $H_1$  waar is  $\rightarrow$  meer kans op een kleine waarde (sterk negatief, sterke daling) dan onder  $H_0$ .
- t-verdeling onder  $H_0$  gebruiken om na te gaan of de geobserveerde test-statistiek  $t = -8.12$  klein genoeg is om te kunnen besluiten dat  $\mu < 0$ .
- Is de geobserveerde teststatistiekwaarde ( $t = -8.12$ ) een waarde die we verwachten als  $H_0$  waar is, of is het een waarde die onwaarschijnlijk klein is als  $H_0$  waar is?**
- In het laatste geval deduceren we dat  $H_0$  niet waar is, en concluderen we  $H_1$ .

- De vraag blijft:
  - 1 hoe groot moet geobserveerde teststatistiek  $t$  zijn om  $H_0$  te verwerpen zodat
  - 2 we bereid zijn om  $H_1$  te besluiten en
  - 3 hoe zeker zijn we van deze beslissing?
- Het antwoord hangt samen met de interpretatie van de kansen die berekend kunnen worden op basis van de distributie van  $t$  onder de nulhypothese, de nuldistribution

### 5.5.3. De p-waarde

- Kans waarop de keuze tussen  $H_0$  en  $H_1$  gebaseerd wordt.
- berekeningswijze is context-afhankelijk
- Voor capitol voorbeeld wordt *p*-waarde gegeven door

$$p = P [T \leq t \mid H_0] = P_0 [T \leq t],$$

- waar de index “0” in  $P_0 [.]$  aangeeft dat de kans onder de nulhypothese berekend wordt.
- m.a.w. de kans om in een willekeurige steekproef onder  $H_0$  een waarde voor de teststatistiek  $T$  te bekomen die lager of gelijk is aan de waarde die in de huidige steekproef werd geobserveerd (meer extreem in de richting van  $H_1$ ).

- De  $p$ -waarde voor het captopril voorbeeld wordt berekend als

$$p = P_0 [T \leq -8.12] = F_t(-8.12; 14) = 0.6 \cdot 10^{-6}.$$

waarbij  $F_t(\cdot; 14)$  de cumulatieve distributie functie is van een t-verdeling met 14 vrijheidsgraden,

$$F_t(x; 14) = \int_{-\infty}^x f_t(x; 14).$$

Waarbij  $f_t(\cdot; 14)$  de densiteitsfunctie is van de t-verdeling.

De oppervlakte onder de densiteitsfunctie is opnieuw een kans.

- Deze kans kan berekend worden in R m.b.v. de functie `pt(x,df)` die twee argumenten heeft
- de waarde van de test-statistiek `x` en
- het aantal vrijheidsgraden van de t-verdeling `df`.
- `pt(x,df)` berekent de kans om een waarde te observeren die kleiner of gelijk is aan `x` wanneer men een willekeurige observatie trekt uit een t-verdeling met `df` vrijheidsgraden.

```
n <- length(delta)
stat<-(mean(delta)-0)/(sd(delta)/sqrt(n))
stat
```

```
## [1] -8.122816
```

```
pt(stat,n-1)
```

```
## [1] 5.731936e-07
```

## Definitie *p*-waarde

De **p-waarde** (ook wel **geobserveerd significantieniveau** genoemd) is de kans om onder de nulhypothese een even of meer “extreme” toetsingsgrootheid waar te nemen (in de richting van het alternatief) dan de waarde  $t$  die geobserveerd werd o.b.v. de steekproef. Hoe kleiner die kans is, hoe sterker het bewijs tegen de nulhypothese.

Merk op dat de p-waarde de kans **niet** uitdrukt dat de nulhypothese waar is!

- Het woord “extreem” duidt op de richting waarvoor de teststatistiek onder de alternatieve hypothese meer waarschijnlijk is.
- In het voorbeeld is  $H_1 : \mu < 0$  en verwachten we dus kleinere waarden van  $t$  onder  $H_1$ .
- Vandaar de kans op  $T \leq t$ .
- Uit de definitie van de  $p$ -waarde volgt dat een kleine  $p$ -waarde betekent dat de geobserveerde teststatistiek eerder onwaarschijnlijk is als aangenomen wordt dat  $H_0$  correct is.
- Dus een voldoende kleine  $p$ -waarde noopt ons tot het **verwerpen van  $H_0$**  ten voordele van  $H_1$ .
- De drempelwaarde waarmee de  $p$ -waarde vergeleken wordt, wordt het **significantieniveau** genoemd en wordt voorgesteld door  $\alpha$ .

## Definitie significantieniveau

De drempelwaarde  $\alpha$  staat gekend als het **significantieniveau** van de statistische test. Een statistische test uitgevoerd op het  $\alpha$  significantieniveau wordt een **niveau- $\alpha$  test** genoemd (Engels: *level- $\alpha$  test*).

Een toetsingsresultaat wordt *statistisch significant* genoemd wanneer bijhorende p-waarde kleiner is dan  $\alpha$

- $\alpha$  wordt meestal gelijk aan 5% wordt genomen.
- Hoe kleiner de p-waarde hoe meer 'significant' het testresultaat afwijkt van de verwachting onder de nulhypothese.
- Het aangeven van een p-waarde voor een toets geeft bijgevolg meer informatie over het resultaat dan een eenvoudig ja/nee antwoord of de nulhypothese wordt verworpen op een vast gekozen  $\alpha$ -niveau.
- Het geeft immers niet alleen aan of de nulhypothese verworpen wordt op een gegeven significantieniveau, maar ook op welke significantieniveaus de nulhypothese verworpen wordt.
- Ze vat dus de bewijskracht tegen de nulhypothese samen

$> 0.10$	niet significant (zwak bewijs)
$0.05 - 0.10$	marginaal significant, suggestief
$0.01 - 0.05$	significant
$0.001 - 0.01$	sterk significant
$< 0.001$	extreem significant

## 5.5.4. Kritieke waarde

- **alternatieve wijze voor de formulering van de beslissingsregel** via een kritieke waarde.
- Beslissingsregels schrijven a.d.h.v. de teststatistiek.
- Bij gebruik van  $p$ -waarden bepaalt  $p = \alpha$  de grens.
- Een  $p$ -waarde van  $\alpha$  schrijven we als

$$p = P_0 [T \leq t] = \alpha.$$

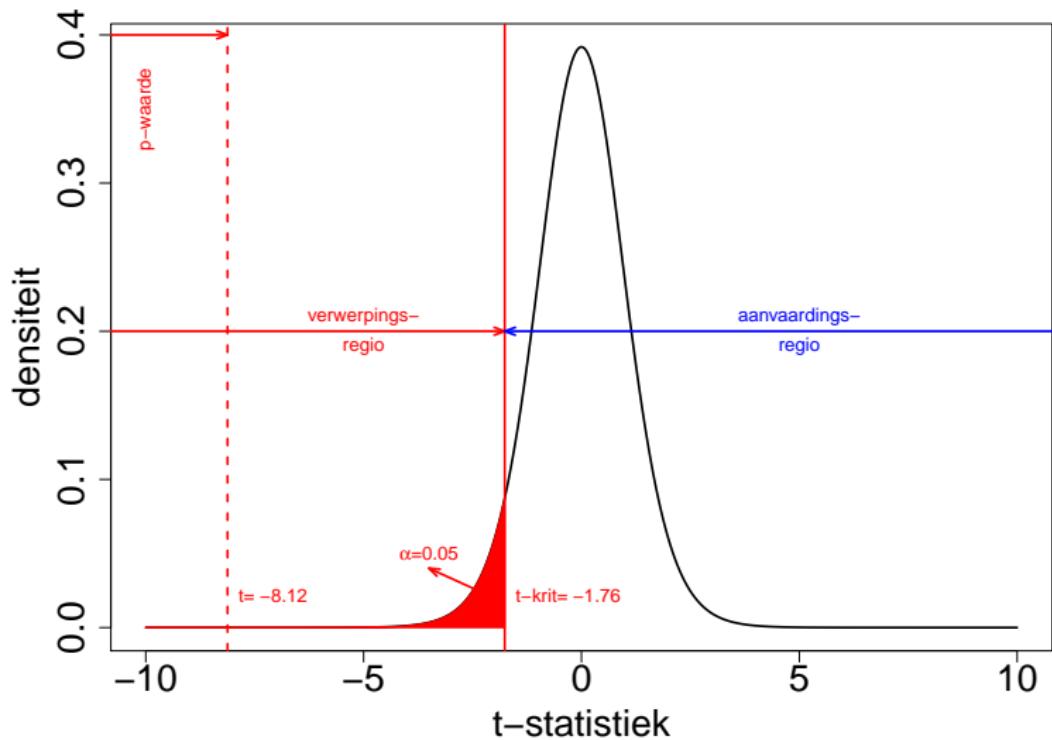
- Dat is exact de definitie van het het  $\alpha$ -percentiel van de distributie van  $T$ .
- In het voorbeeld is de nuldistributie  $t_{n-1}$ .
- Dus,

$$P_0 [T \leq -t_{n-1;\alpha}] = \alpha.$$

De beslissingsregel mag dus ook geschreven worden als

als  $t < -t_{n-1;\alpha}$  dan verwerp  $H_0$  en besluit  $H_1$   
als  $t \geq -t_{n-1;\alpha}$  dan aanvaard  $H_0$ .

- Het percentiel  $t_{n-1;\alpha}$  dat de drempelwaarde vormt wordt **kritieke waarde** op het 5% significantieniveau genoemd.



## 5.4. Beslissingsfouten

Beslissing om  $H_0$  al dan niet te verwerpen bepaald door slechts een steekproef te observeren. Foute beslissing kan genomen worden.

Besluit		Werkelijkheid	
		$H_0$	$H_1$
Aanvaard $H_0$	OK	Type II ( $\beta$ )	
	Type I ( $\alpha$ )	OK	

- Type I fout: verkeerdelijk de nulhypothese verwerpen
- Type II fout: verkeerdelijk de nulhypothese niet verwerpen.

## Voorbeeld: captopril

- $H_0$ : toedienen van captopril heeft geen effect op syst. bloeddruk
- $H_a$ : toedienen van captopril zorgt gemiddeld gezien voor een verlaging van de syst. bloeddruk
- **Type I fout**: er is gemiddeld geen bloeddrukvermindering na toedienen van captopril, maar men besluit tegendeel.
- **Type II fout**: er is gemiddeld wel een bloeddrukvermindering na toedienen van captopril, maar de statistische test detecteert dit niet.

- De beslissing is gebaseerd op een teststatistiek  $T$  die een toevalsveranderlijke is.
- De beslissing is dus ook stochastisch en aan de vier mogelijke situaties uit bovenstaand schema kunnen dus probabiliteiten toegekend worden.
- Net zoals voor het afleiden van de steekproefdistributie van de teststatistiek, moeten we de distributie van de steekproefobservaties kennen alvorens het stochastisch gedrag van de beslissingen te kunnen beschrijven.
- Indien we aannemen dat  $H_0$  waar is, dan is de distributie van  $T$  gekend en kunnen ook de kansen op de beslissingen bepaald worden voor de eerste kolom van de tabel.

We starten met de kans op een type I fout (hier uitgewerkt voor het captoril voor beeld):

$$P[\text{type I fout}] = P[\text{verwerp } H_0 \mid H_0] = P_0[T < t_{n-1;1-\alpha}] = \alpha.$$

- Het significantieniveau  $\alpha$  is dus de kans op het maken van een type I fout.
- Statistische test garandeert dus dat kans op type I fout gecontroleerd wordt op het significantieniveau  $\alpha$ .
- De kans op het correct aanvaarden van  $H_0$  is dus  $1 - \alpha$ .
- We kunnen aantonen dat p-waarde onder  $H_0$  uniform verdeeld is
- Het leidt dus tot een uniforme beslissingsstrategie.

- Bepalen van de kans op een type II fout is minder evident
- De alternatieve hypothese minder éénduidig is als de nulhypothese.
- In het captopril voorbeeld is  $H_1 : \mu < 0$
- Met deze informatie wordt de distributie van de steekproefobservaties niet volledig gespecificeerd
- Dus ook niet de distributie van de teststatistiek.
- Dit impliceert dat we eigenlijk de kans op een type II fout niet kunnen berekenen.

- Klassieke *work-around* bestaat erin om één specifieke distributie te kiezen die voldoet aan  $H_1$ .

$$H_1(\delta) : \mu = 0 - \delta \text{ voor een } \delta > 0.$$

- De parameter  $\delta$  kwantificeert de afwijking van de nulhypothese.
- De **kracht** van een test (Engels: *power*) is een kans die meer frequent gebruikt wordt dan de kans op een type II fout  $\beta$ .
- De kracht wordt gedefinieerd als

$$\pi(\delta) = 1 - \beta(\delta) = P_\delta [T > t_{n-1;1-\alpha}] = P_\delta [P < \alpha].$$

- Kracht van niveau- $\alpha$  test voor detectie van afwijking  $\delta$  van het gemiddelde onder de nulhypothese  $\mu_0 = 0$  is dus de kans dat de niveau- $\alpha$  test detecteert wanneer de afwijking in werkelijkheid  $\delta$  is.
- Merk op dat  $\pi(0) = \alpha$  en de kracht van een test toeneemt als de afwijking van de nulhypothese toeneemt.

De **kracht** van de test (d.i. de kans om Type II fouten te vermijden) wordt typisch niet gecontroleerd, tenzij d.m.v. studiedesign en steekproefgrootte.



## Interpretatie

Stel dat we voor een gegeven dataset bekomen dat  $p < \alpha$ , m.a.w.  $H_0$  wordt verworpen.

- Volgens het schema slechts twee mogelijkheden (zie onderste rij van schema):
  - ofwel is de beslissing correct,
  - ofwel hebben we een type I fout gemaakt.
- Over de type I fout weten we echter dat ze slechts voorkomt met een kleine kans.

Anderzijds, indien  $p \geq \alpha$  en we  $H_0$  niet verwerpen, dan zijn er ook twee mogelijkheden:

- ofwel is de beslissing correct,
- ofwel hebben we een type II fout gemaakt.

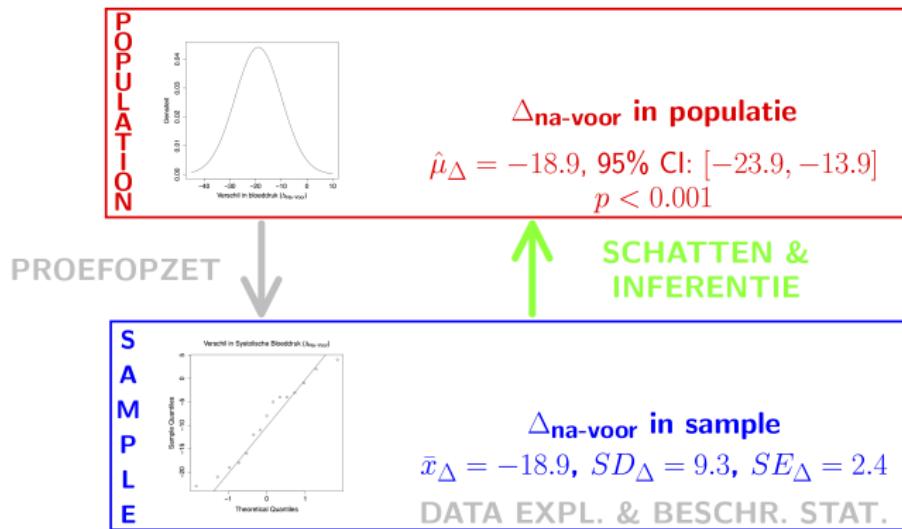
De kans op een type II fout ( $\beta$ ) is echter niet gecontroleerd op een gespecifieerde waarde.

De statistische test is zodanig geconstrueerd dat ze enkel de kans op een type I fout controleert (op  $\alpha$ ).

Om wetenschappelijk eerlijk te zijn, moeten we een pessimistische houding aannemen en er rekening mee houden dat  $\beta$  groot zou kunnen zijn (i.e. een kleine kracht).

- Bij  $p < \alpha$  wordt de nulhypothese verworpen
  - We mogen hieruit concluderen dat  $H_1$  waarschijnlijk juist is.
  - Dit noemen we een sterke conclusie.
- Bij  $p \geq \alpha$  wordt de nulhypothese aanvaard
  - Maar dat impliceert niet dat we concluderen dat  $H_0$  juist is.
  - We kunnen enkel besluiten dat de data onvoldoende bewijskracht tegen  $H_0$  ten gunste van  $H_1$  bevatten.
  - Dit noemen we een daarom zwakke conclusie.

## 5.5.6. Conclusies Captopril voorbeeld



De test die we hebben uitgevoerd is in de literatuur ook bekend als

- de **one sample t-test** op het verschil of
- als een **gepaarde t-test**
- we beschikken immers over gepaarde gegevens per patiënt.

De test is eenzijdig uitgevoerd. We testen tegen het alternatief dat er een bloeddrukdaling is.

- Beide testen (one sample t-test op het verschil en de gepaarde t-test) geven ons inderdaad dezelfde resultaten:

```
t.test(delta,alternative="less")
```

```
##  
##  One Sample t-test  
##  
## data:  delta  
## t = -8.1228, df = 14, p-value = 5.732e-07  
## alternative hypothesis: true mean is less than 0  
## 95 percent confidence interval:  
##       -Inf -14.82793  
## sample estimates:  
## mean of x  
## -18.93333
```

```
t.test(captopril$SBPa, captopril$SBPb, paired=TRUE, alternative="le
```

```
##  
##  Paired t-test  
##  
## data:  captopril$SBPa and captopril$SBPb  
## t = -8.1228, df = 14, p-value = 5.732e-07  
## alternative hypothesis: true difference in means is less than  
## 95 percent confidence interval:  
##       -Inf -14.82793  
## sample estimates:  
## mean of the differences  
##                      -18.93333
```

## Conclusie

Na toediening van captopril is er een extreem significante verlaging van de systolische bloeddruk bij patiënten met hypertensie ( $p << 0.001$ ). De systolische bloeddruk neemt gemiddeld met 18.9 mm kwik af na de behandeling met captopril (95% BI  $[-\infty, -14.82]$  mm Hg).

Merk op dat we

- ❶ Een eenzijdig interval rapporteren gezien we enkel geïnteresseerd zijn om aan te tonen dat er een bloeddrukdaling is.
- ❷ Door het pre-test/post-test design geen uitsluitsel kunnen geven of dit te wijten is aan de werking van het middel of aan een placebo effect. Er was geen goede controle! Het gebrek van een goede controle is veelal een probleem bij pre-test/post-test designs.

## 5.5.7. Eenzijdig of tweezijdig toetsen?

De test in het captopril voorbeeld was een eenzijdige test. We wensen immers enkel te detecteren of de captopril behandeling de bloeddruk gemiddeld gezien doet dalen.

In andere gevallen of een andere context wenst men enkel een stijging te detecteren.

Stel dat bloeddrukverschil was gedifinieerd als  $X'_i = Y_i^{\text{voor}} - Y_i^{\text{na}}$

- positieve waarden geven dan aan dat er een bloeddrukdaling was na de behandeling van captopril
- de bloeddruk bij aanvang is dan immers groter dan na de behandeling.
- De gemiddelde bloeddrukverandering in de populatie noteren we nu als  $\mu' = E[X']$ .
- In dat geval een eenzijdige test om  $H_0 : \mu' = 0$  te testen tegen  $H_1 : \mu' > 0$ .
- Voor deze test kunnen we de p-waarde als volgt berekenen:

$$p = P_0 [T \geq t].$$

Analyse o.b.v. toevallige veranderlike  $X'$ . Argument  
alternative="greater" zodat we  $H_1 : \mu' > 0$  toetsen:

```
delta2 <- captopril$SBPb - captopril$SBPa  
t.test(delta2, alternative="greater")
```

```
##  
##  One Sample t-test  
##  
## data: delta2  
## t = 8.1228, df = 14, p-value = 5.732e-07  
## alternative hypothesis: true mean is greater than 0  
## 95 percent confidence interval:  
##  14.82793      Inf  
## sample estimates:  
## mean of x  
## 18.93333
```

Uiteraard bekomen we met deze analyse exact dezelfde p-waarde en hetzelfde betrouwbaarheidsinterval. Enkel het teken is omgewisseld.

Naast eenzijdige testen kunnen eveneens tweezijdige testen worden uitgevoerd.

Het had gekund dat de onderzoekers de werking van het nieuwe medicijn captopril wensten te testen, maar het werkingsmechanisme nog niet kenden in de ontwerp fase.

In dat geval zou het eveneens interessant geweest zijn om zowel een stijging als een daling van de bloeddruk te kunnen detecteren.

Hiervoor zou men een tweezijdige toetsstrategie moeten gebruiken waarbij men de nulhypothese

$$H_0 : \mu = 0$$

gaat testen versus het alternatieve hypothese

$$H_1 : \mu \neq 0,$$

zodat het gemiddelde onder de alternatieve hypothese verschillend is van 0.

Het kan zowel een positieve of negatieve afwijking zijn en men weet niet bij aanvang van de studie in welke richting het werkelijk gemiddelde zal afwijken onder  $H_1$ .

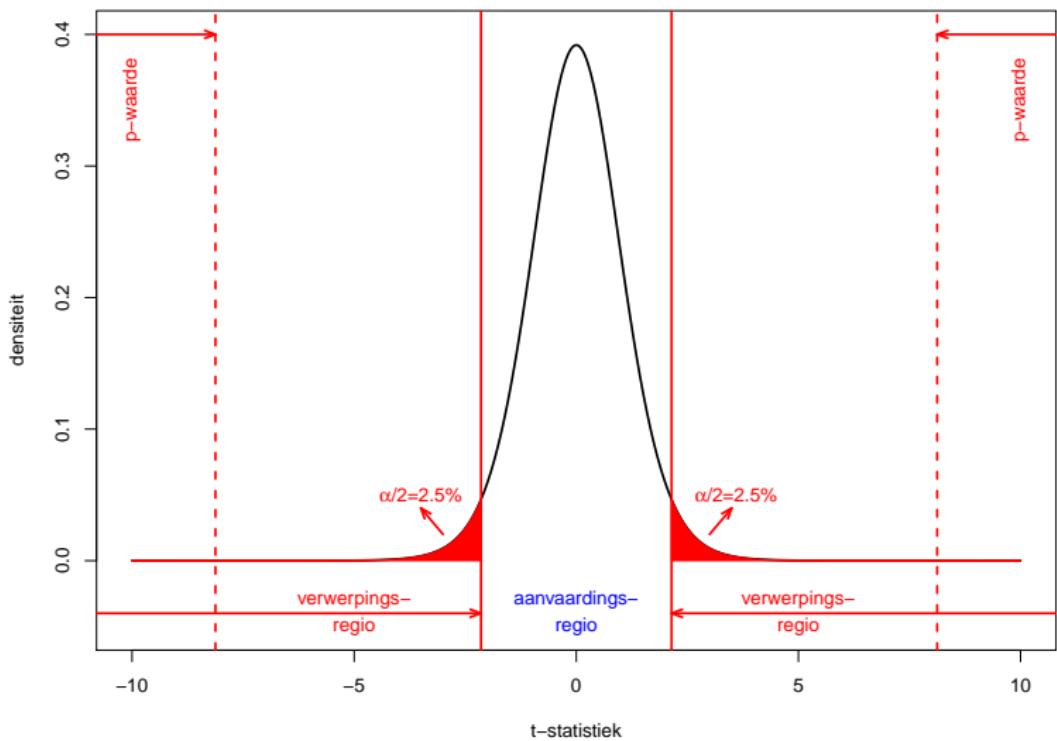
We kunnen tweezijdig testen op het  $\alpha = 5\%$  significantieniveau door

1 een kritieke waarde af te leiden:

- Bij tweezijdige test kan effect onder  $H_1$  zowel positief of negatief zijn.
- Onder  $H_0$  kans berekenen om een effect te observeren dat meer extreem is dan het resultaat dat werd geobserveerd in de steekproef.
- "Meer extreem" betekent nu dat de statistiek groter is in absolute waarde dan het geobserveerde resultaat, want zowel grote (sterk positieve) als kleine (sterk negatieve) waarden zijn een indicatie van een afwijking van de nulhypothese.
- Kritieke waarde af leiden waarbij we significatie-niveau  $\alpha$  verdelen over de linker en rechter staart van de verdeling onder  $H_0$ .
- t-verdeling symmetrisch dus kritieke waarde  $c$  kiezen zodat  $\alpha/2 = 2.5\%$  dat  $T \geq c$  en  $\alpha/2 = 2.5\%$  kans is dat  $T \leq -c$ .
  - Anders geformuleerd onder  $H_0$   $\alpha = 5\%$  kans dat  $|T| \geq c$

2 We kunnen ook gebruik maken van een tweezijdige p-waarde:

$$\begin{aligned} p &= P_0 [T \leq -|t|] + P_0 [T \geq |t|] \\ &= P_0 [|T| \geq |t|] \\ &= P_0 [T \geq |t|] \times 2. \end{aligned}$$



Als niet vooraf gedefineerd dat men enkel een bloeddrukdaling wenst te detecteren was twee-zijdige test vereist.

Argument alternative van de `t.test` functie is default  
`alternative="two.sided"` → twee-zijdig.

```
t.test(delta)
```

```
##  
##  One Sample t-test  
##  
## data:  delta  
## t = -8.1228, df = 14, p-value = 1.146e-06  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## -23.93258 -13.93409  
## sample estimates:  
## mean of x  
## -18.93333
```

- We bekomen nog steeds een extreem significant resultaat.
- De p-waarde is echter dubbel zo groot omdat we tweezijdig testen.
- We verkrijgen eveneens een tweezijdig betrouwbaarheidsinterval.

## Wanneer eenzijdig of tweezijdig toetsen?

- Met een eenzijdige toets kan men gemakkelijker een alternatieve hypothese aantonen (op voorwaarde dat ze waar is) dan met een tweezijdige toets.
- Alle informatie kan immers worden aangewend om in 1 enkele richting te zoeken.
- Precies daarom vergt de eenzijdige toets een extra beschouwing vóór de aanvang van de studie.
- Ook al hebben we sterke a priori vermoedens, vaak kunnen we niet zeker zijn dat dat zo is.
- Anders was er immers geen reden om dit te willen toetsen.

Als men een eenzijdige test voorstelt, maar men vindt een resultaat in de andere richting dat formeel statistisch significant is, dan is het niet geschikt om dit te zien als bewijs voor een werkelijk effect in die richting.

- Dat is omdat de onderzoekers die mogelijkheid uitgesloten hebben bij de planning van de studie en het resultaat daarom zó onverwacht is dat het als een vals positief resultaat kan gezien worden.
- Een eenzijdige test is om die reden niet aanbevolen.

Een tweezijdige toets is altijd verdedigbaar omdat ze in principe toelaat om elke afwijking van de nulhypothese te detecteren. Ze worden daarom het meest gebruikt en ten zeerste aangeraden.

**Het is nooit toegelaten om een tweezijdige toets in een eenzijdige toets om te zetten **op basis van wat men observeert in de gegevens!****

Anders wordt de type I fout van de toetsingsstrategie niet correct gecontroleerd.

Dat wordt geïllustreerd in de onderstaande simulatie. We evalueren twee strategieën:

- de correcte tweezijdige test en
- een test waar we eenzijdig toetsen op basis van het teken van het geobserveerde effect.

```
set.seed(115)
mu <- 0
sigma <- 9.0
nSim <- 1000
alpha <- 0.05
n <- 15
pvalsCor <- pvalsInCor<-array(0,nSim)
for (i in 1:nSim)
{
  x <- rnorm(n,mean=mu,sd=sigma)
  pvalsCor[i] <- t.test(x)$p.value
  if (mean(x)<0)
    pvalsInCor[i] <- t.test(x,alternative="less")$p.value else
    pvalsInCor[i] <- t.test(x,alternative="greater")$p.value
}
```

```
mean(pvalsCor<0.05)
```

```
## [1] 0.049
```

```
mean(pvalsInCor<0.05)
```

```
## [1] 0.106
```

- Type I fout correct gecontroleerd op  $\alpha$  bij tweezijdige test
- Type I fout niet correct bij eenzijdige toetsen op basis van het teken van het geobserveerde effect

## 5.7. Two-sample t-test

## ZWEETANALYSE

## Moeder, waarom stinken wij?

Hoe erg je zweet stinkt, hangt vooral af van de bacteriëngemeenschap onder je oksels. En van het soort sportshirt dat je draagt, zo blijkt uit onderzoek van de Universiteit Gent. Dries Clijns CALLEWAERT



Fig. 1

**S**teriel zenuw sticht niet. Het zijn de bacteriën op de huid die een onfelle geur kunnen veroorzaken. De geur is een waarschuwing voor de kleiner, vluchtere bacteriën die kunnen stikken.

Op onze handen leven een miljoen bacteriën, maar er zijn ook 10 miljoen op de handen. 1,8 miljoen meer dan het gezonde organisme van het menselijk lichaam. De hand heeft verschillende schilden met hun eigen karakteristieke bacteriemengeling. De掌 is een mengeling van *Staphylococcus epidermidis* en *Staphylococcus aureus* bacteriën per vierkante centimeter. De oksel lijkt achter aan op een trappenhuisje, maar is veel minder hondig van bacteriën per vierkante centimeter. De oksel heeft een mengeling van *Staphylococcus epidermidis* en *Staphylococcus aureus*.

Maar niet alleen het aantal bacteriën is van belang, ook de soorten bacteriën. Wel kunnen ook, bijvoorbeeld, de giftige sporen van de *Clostridium botulinum* bacterie de bacterieën groeien en doodslaan.

De meeste mensen (rond de 60%) van de proefpersonen overleden aan staphylocokken, bij een aantal mensen was de oorzaak een *Clostridium botulinum* spore, gegeven dat staphylocokken.

1

10

1

ter componenten, maar de productiecapaciteit daar viel heel veel in de staatsbegroting. De productiecapaciteit van de landbouwsector daalde in grote mate. De landbouwsector had steeds te maken met beperkende factoren, zoals arbeid, land, voedsel, energie, hygiëne, gezondheid en gezag. De landbouwsector was de enige sector die een goede haaklijn had bij de criminale factor.

Verder endemisch was dat er een tekort aan arbeid bestond. Onder deze term was de tekort van de productiecapaciteit bedoeld, weliswaar klein, maar wel duidelijk ten opzichte tussen de landbouwsector en de industrie. De arbeidsmarkt was verder overgegaan wegens de oorlog de tijd dat de manen productieve harsen in de industrie niet waren teruggekeerd met zeep. Er was een tekort aan arbeid, maar er was een achterstand op de arbeidsmarkt die de arbeidsmarkt in de overheid.

*Polyester shirts stinken  
‘Stankbacteriën’ gedijen*

woaden: van een anti-transpirant kan je soms net minder gaan stinken. Let's wat de fabrikanten van deze producten maar beter in de gaten kunnen houden.

**KLEUR**  
En hoe zit het met kleidij en zweefgeurkjes? Kleine absorberen ontzettend goed en zorgen voor een warme omgeving waarin bacteriën goed gedijen. Wij wilden onderzoeken waarom sommige kleedijstukken sterker stinken dan andere. We lieten 25 volwassenen, goede mannen en vrouwen een uur lang doelsturen aan een intensieve sessie spanning. Na afloop ver-

en meer dan katoen.  
en er goed

e geur. Een anti-transpirant kan bewerken om een aanzetting van bacteriën veroorza-  
ken die nog meer onaangename geuren produceren.  
De bacteriën in de kledij waren anders dan die

de hand. De geur van het T-shirt kan dus ook verschillen van de lichaamsgroei van de drager. Daardoor hangt de microbiologie van het T-shirt nauw samen met het type voed. We werken nu aan oplossingen om de bacteriële geschiedenis te snuiven naar een betere geur. ■

Two-sample t-test werd ontwikkeld om verschillen in gemiddelde te detecteren tussen twee onafhankelijke groepen.

Oksel dataset:

- Men vermoedt dat hinderlijke geur onder de oksels (bromhidrosis) wordt veroorzaakt door specifieke microorganismen die behoren tot de groep van de *Corynebacterium spp.*
- Het is immers niet het zweet dat de geur veroorzaakt, maar de geur is het resultaat van specifieke bacteriën die het zweet metaboliseren.
- Een andere sterk abundante groep wordt gevormd door de *Staphylococcus spp.*
- In de CMET-groep van de Universiteit Gent onderzoek naar microbiële transplanties in oksels om mensen van hinderlijke okselgeur te verlossen. - Therapie:
  - 1 oksel-microbiom verwijderen door lokale antibiotica behandeling
  - 2 Via microbiële transplantatie de populatie te beïnvloeden. (zie: <https://youtu.be/9RIFyqLXdVw> )

## De primaire onderzoeks vraag

Leidt microbiële transplantatie na zes weken tot een verandering in de relatieve abundantie van *Staphylococcus spp.* in het oksel microbioom i.v.m. placebo behandeling die enkel bestaat uit een antibiotica behandeling?

## Design

Twintig personen met een hinderlijke okselgeur worden willekeurig toegekend aan twee behandelingsgroepen

- placebo (enkel antibiotica)
- transplantie (antibiotica, gevolgd door microbiële transplantatie).

Staalname 6 weken na start van behandeling

Relatieve abundancies van *Staphylococcus spp.* en *Corynebacterium spp.* in het microbioom gemeten via DGGE (*Denaturing Gradient Gel Electrophoresis*).



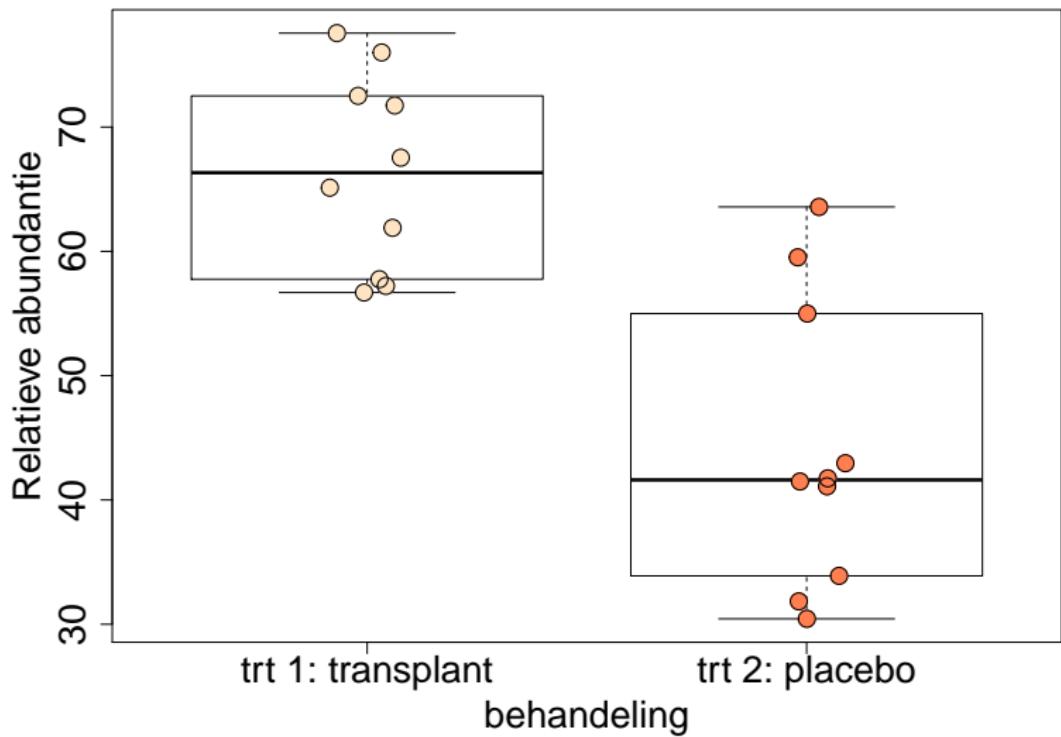
Dataset bevat variabelen Staph en Cor die de relatieve abundancies (%) weergeven van *Staphylococcus spp.* en *Corynebacterium spp.*

De variabele Rel werd berekend als

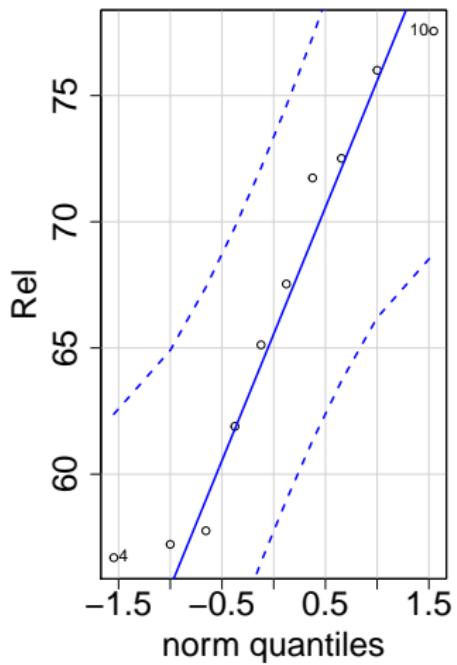
$$\text{Rel} = \frac{\text{Staph}}{\text{Staph} + \text{Cor}}.$$

Deze variabele is het relatief aandeel van *Staphylococcus spp.* op het totaal aantal *Staphylococcus spp.* en *Corynebacterium spp.*.

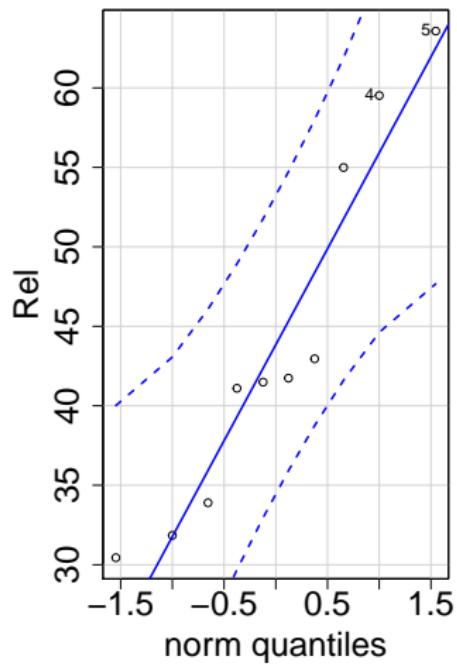
```
##           trt Staph  Cor      Rel
## 1 trt 2: placebo  34.7 28.4 54.99208
## 2 trt 2: placebo  16.4 35.1 31.84466
## 3 trt 2: placebo  31.4 45.0 41.09948
## 4 trt 2: placebo  44.7 30.4 59.52064
## 5 trt 2: placebo  45.9 26.3 63.57341
## 6 trt 2: placebo  30.7 43.3 41.48649
```



trt 1: transplant



trt 2: placebo



Notatie:

Stel  $Y_{ij}$  de uitkomst van observatie  $i = 1, \dots, n_j$  uit populatie  $j = 1, 2$ .

Gebruik van de term **behandeling** of **groep** i.p.v. populatie

Hier is behandeling  $j = 1$  de microbiële transplantatie en behandeling  $j = 2$  de placebo behandeling.

We veronderstellen

$$Y_{ij} \text{ i.i.d. } N(\mu_j, \sigma^2) \quad i = 1, \dots, n_j \quad j = 1, 2.$$

Merk op dat dit inhoudt dat gelijke varianties verondersteld worden:  
**homoskedasticiteit**

Ongelijke varianties met **heteroskedasticiteit**.

Testen van

$$H_0 : \mu_1 = \mu_2$$

tegenover de alternatieve hypothese

$$H_1 : \mu_1 \neq \mu_2.$$

$H_1$  drukt dus opnieuw de onderzoeksvraag uit: een verschil in relatieve abundantie van *Staphylococcus spp.* na microbiële transplantatie t.o.v. de placebo behandeling.

$H_0$  en  $H_1$  kunnen ook worden uitgedrukt in termen van de effectgrootte tussen behandeling en placebo groep  $\mu_1 - \mu_2$ :

$$H_0 : \mu_1 - \mu_2 = 0,$$

$$H_1 : \mu_1 - \mu_2 \neq 0.$$

Effectgrootte schatten a.d.h.v. de steekproefgemiddeldes:

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{Y}_1 - \bar{Y}_2.$$

Gezien experimentele eenheden onafhankelijk zijn, zijn de steekproefgemiddeldes dat ook en is de variantie op het verschil:

$$\text{Var}_{\bar{Y}_1 - \bar{Y}_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right).$$

De standard error is bijgevolg:

$$\sigma_{\bar{Y}_1 - \bar{Y}_2} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Variantie kan apart geschat worden in elke groep d.m.v. steekproefvariatie

Als we gelijkheid van variantie kunnen veronderstellen kan variantie meer precies worden geschat door gebruik te maken van alle gegevens in beide groepen.

Deze variatieschatter wordt ook de gepoolde variantieschatter genoemd:  
 $S_p^2$ .

O.b.v. de observaties uit de eerste groep kan  $\sigma_1^2$  geschat worden als

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{i1} - \bar{Y}_1)^2.$$

Analoog geldt voor tweede groep

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_{i2} - \bar{Y}_2)^2.$$

Merk op dat we homoscedasticiteit veronderstellen,  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ .

Dus  $S_1^2$  en  $S_2^2$  zijn schatters zijn voor dezelfde parameter  $\sigma^2$ .

Daarom kunnen ze gezamenlijk gebruikt worden om tot één schatter te komen die alle  $n_1 + n_2$  observaties gebruikt:

$$S_p^2 = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2 = \frac{1}{n_1 + n_2 - 2} \sum_{j=1}^2 \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2.$$

Gepoolde variantieschatter wordt dus geschat door gebruik te maken van de kwadratische afwijkingen tussen de observaties en hun groepsgemiddelde en dat te delen door het aantal vrijheidsgraden  $n_1 + n_2 - 2$

Two-sample  $t$ -teststatistiek:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Als de data onafhankelijk zijn, de steekproefgemiddelden normaal verdeeld zijn en de variantie in beide groepen gelijk zijn, dan kan men aantonen de teststatistiek  $T$  opnieuw een  $t$ -verdeling volgt met  $n_1 + n_2 - 2$  vrijheidsgraden onder de nulhypothese.

Aangezien de alternatieve hypothese  $H_1 : \mu_1 \neq \mu_2$  impliceert dat de probabiliteitsmassa van de distributie van  $T$  onder  $H_1$  verschuift naar hogere of lagere waarden, zullen we  $H_0$  wensen te verwerpen ten gunste van  $H_1$  voor grote absolute waarde van de teststatistiek. De  $p$ -waarde wordt dus

$$\begin{aligned} p &= P_0 [T \leq -|t|] + P_0 [T \geq |t|] \\ &= P_0 [|T| \geq |t|] \\ &= P_0 [T \geq |t|] \times 2 \\ &= 2 \times (1 - F_T(|t|; n_1 + n_2 - 2)), \end{aligned}$$

met  $F_T(\cdot; n_1 + n_2 - 2)$  de cumulatieve distributiefunctie van  $t_{n_1+n_2-2}$ .

### 5.6.1. Oksel-voorbeeld

De onderzoeksraag van het oksels-voorbeeld kan vertaald worden in een nulhypothese en een alternatieve hypothese.

De nulhypothese verwoordt de stelling dat de behandeling geen effect heeft op de gemiddelde relatieve abundantie van *Staphylococcus spp.*.

Indien  $\mu_1$  en  $\mu_2$  de gemiddelde abundanties voorstellen in respectievelijk de transplantatie groep en de placebo groep, dan schrijven we

$$H_0 : \mu_1 = \mu_2.$$

De alternatieve hypothese correspondeert met wat we wensen te bewijzen aan de hand van de experimentele data: een verschil in gemiddelde abundantie van *Staphylococcus spp.* in de transplantatie groep i.v.m. de placebo groep. Dus

$$H_1 : \mu_1 \neq \mu_2.$$

De berekeningen kunnen als volgt in R worden uitgevoerd:

```
ybar1<-mean(oksel$Staph[oksel$trt=="trt 1: transplant"])
ybar1
```

```
## [1] 49.79
```

```
ybar2<-mean(oksel$Staph[oksel$trt=="trt 2: placebo"])
ybar2
```

```
## [1] 31.9
```

```
var1<-var(oksel$Staph[oksel$trt=="trt 1: transplant"])
var1
```

```
## [1] 64.95656
```

```
var2<-var(oksel$Staph[oksel$trt=="trt 2: placebo"])
var2
```

```
## [1] 76.78222
```



```
n1<-sum(oksel$trt=="trt 1: transplant")
```

```
n1
```

```
## [1] 10
```

```
n2<-sum(oksel$trt=="trt 2: placebo")
```

```
n2
```

```
## [1] 10
```

*#gepoelde variantieschatting*

```
sp2<-((n1-1)*var1+(n2-1)*var2)/(n1+n2-2)
```

```
sp2
```

```
## [1] 70.86939
```

```
#geobserveerde t-statistiek
t.obs<-(ybar1-ybar2)/sqrt(sp2/n1+sp2/n2)
t.obs
```

```
## [1] 4.751886
```

```
#p-waarde
p<-(1-pt(abs(t.obs),df=n1+n2-2))*2
p
```

```
## [1] 0.0001592919
```

```
#Berkening o.b.v. probabiliteit in de linker staart
# is vaak stabiever in R
p<-pt(-abs(t.obs),df=n1+n2-2)*2
p
```

```
## [1] 0.0001592919
```

De R software heeft ook een specifieke functie voor het uitvoeren van deze *t*-test.

```
t.test(Staph~trt, data=oksel, var.equal=TRUE)
```

```
##  
##  Two Sample t-test  
##  
## data:  Staph by trt  
## t = 4.7519, df = 18, p-value = 0.0001593  
## alternative hypothesis: true difference in means is not equal to zero  
## 95 percent confidence interval:  
##   9.980404 25.799596  
## sample estimates:  
## mean in group trt 1: transplant      mean in group trt 2: placebo  
##                           49.79                           31.1
```

Uit deze analyse lezen we  $p \approx 0.16 \times 10^{-3} << 0.05$ .

Dus op het 5% significantieniveau verwerpen we de nulhypothese ten voordele van de alternatieve en besluiten we dat de gemiddelde abundantie van *Staphylococcus spp.* extreem significant hoger is in de transplantatie groep dan in de placebo groep.

Indien de transplantatie geen effect heeft op de gemiddelde abundantie van *Staphylococcus spp.*, dan is er slechts een kans van 16 in de 100000 om een teststatistiek te bekomen in een willekeurige steekproef die minstens zo extreem is als deze die wij geobserveerd hebben.

Dit is uiterst zeldzaam onder de hypothese dat  $H_0$  waar is, en het is kleiner dan 5% (het significantieniveau). Indien  $H_1$  waar zou zijn, dan verwachten we grotere absolute waarden van de teststatistiek en verwachten we dus ook kleine  $p$ -waarden. Om deze reden wensen we niet verder te geloven dat  $H_0$  waar is, en besluiten we dat er veel evidentie in de steekproefdata zit om te besluiten dat  $H_1$  waar is op het 5% significantieniveau.

**Good statistical practice** houdt ook in dat niet enkel de  $p$ -waarde van de hypothesetest wordt gerapporteerd, maar dat ook de gemiddelden en een maat voor de betrouwbaarheid van de schattingen (bv. BI) worden gerapporteerd.

## Conclusie

Gemiddeld is de relatieve abundantie van *Staphylococcus spp.* in het microbioom van de oksel in de transplantatie groep extreem significant verschillend van dat in de controle groep ( $p << 0.001$ ). De relatieve abundantie van *Staphylococcus spp.* is gemiddeld 17.9% hoger in de transplantatie groep dan in de controle groep (95% BI [10.0,25.8]).

## 5.7. Aannames

Geldigheid van t-testens hangt af van enkele distributionele veronderstellingen:

- Onafhankelijke gegevens (design)
- One-sample t-test: normaliteit van de steekproefobservaties
- Paired t-test: normaliteit van de verschillen tussen de gepaarde observaties
- Two-sample t-test: normaliteit van de steekproefobservaties in beide groepen, en gelijkheid van varianties.

Indien niet voldaan is aan de veronderstellingen, is de t-distributie niet de correcte nuldistributie

Bijgevolg is er geen garantie dat de p-waarde en kritieke waarden correct zijn.

Ook voor de constructie van het betrouwbaarheidsinterval van het gemiddelde hebben we beroep gedaan op de veronderstelling van normaliteit.

De normaliteitsveronderstelling was nodig om kwantielen uit de t-verdeling te kunnen gebruiken bij het opstellen van de boven- en ondergrens

De correcte probabiliteitsinterpretatie van het betrouwbaarheidsinterval hangt hiervan af.

## 5.7.1. Nagaan van de veronderstelling van Normaliteit

Normaliteit kan via de volgende methoden nagegaan worden.

### **Boxplots en histogrammen**

Beide figuren laten toe om een idee te vormen over de vorm van de distributie: symmetrie, outliers. . .

### **QQ-plots**

Deze plots laten toe om op een grafische wijze na te gaan in welke mate steekproefobservaties zich gedragen als een vooropgestelde distributie.

## Hypothesetesten (goodness-of-fit test)

Goodness-of-fit testen zijn statistische hypothesetesten die ontwikkeld zijn voor het testen van de nulhypothese dat de steekproefobservaties uit een vooropgestelde distributie getrokken zijn (hier: normale distributie). De alternatieve hypothese is meestal de negatie van de nulhypothese (hier: geen normaliteit). Bekende testen zijn: Kolmogorov-Smirnov, Shapiro-Wilk en Anderson-Darling.

Op het eerste zicht lijkt een goodness-of-fit test een gemakkelijke en zinvolle oplossing.

De methode geeft een  $p$ -waarde en deze laat onmiddellijk toe om te besluiten of de data normaal verdeeld zijn.

Er is echter kritiek te leveren op deze aanpak:

- indien  $p \geq \alpha$ , dan is normaliteit niet bewezen! Het zegt enkel dat er onvoldoende evidentie is tegen de veronderstelling van normaliteit. In een kleine steekproef is de kracht van een test meestal klein.
- indien  $p < \alpha$ , dan mag wel besloten worden om de nulhypothese te verwerpen en mag dus besloten worden dat de data niet normaal verdeeld zijn, maar soms is een afwijking van normaliteit niet zo erg.

**Algemeen advies:** Start met een grafische exploratie van de data (boxplots, histogrammen en QQ-plots) en houdt hierbij steeds de steekproefgrootte in het achterhoofd om te vermijden dat je de figuren zou overinterpretieren.

Als je twijfelt kan je gebruik maken van simulaties waarbij je nieuwe steekproeven simuleert met eenzelfde steekproefgrootte en data die uit de Normaal verdeling komt met eenzelfde gemiddelde en variantie als wat in de steekproef werd geobserveerd.

Indien een afwijking van normaliteit wordt vastgesteld, tracht dan na te gaan (bv. via literatuur) of de statistische methode die je wenst toe te passen, gevoelig is voor dergelijke afwijkingen (een t-test is bijvoorbeeld vrij ongevoelig voor afwijkingen van Normaliteit als de afwijkingen symetrisch zijn). Eventueel kan je ook beroep doen op de centrale limietstelling.

## 5.7.2. Nagaan van homoscedasticiteit

Dat kan opnieuw via boxplots.

De grootte van de box is de interkwartiel range (IQR), een robuuste schatter voor de variantie. Als de verschillen in IQR range niet te groot is  
→ homoscedasticiteit aannemen.

Opnieuw kan inzicht gekregen worden in dergelijke plots door gebruik te maken van simulaties (zie Oefeningen).

Men kan eveneens een formele F-test gebruiken om de varianties te vergelijken (zie oefeningen), maar hiervoor geldt dezelfde kritiek als voor het testen van normaliteit (zie vorige sectie).

Als er bij het vergelijken van gemiddelden tussen twee groepen niet aan homoscedasticiteit is voldaan, kan je gebruik maken van de Welch two-sample T-test. Hierbij wordt de gepoolde variantieschatter niet langer gebruikt.

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

waarbij  $S_1^2$  en  $S_2^2$  de steekproefvarianties zijn in beide groepen.

Deze statistiek volgt bij benadering een t-verdeling met een aantal vrijheidsgraden dat ligt tussen het kleinste aantal observaties  $\min(n_1 - 1, n_2 - 1)$  en  $n_1 + n_2 - 2$ .

De vrijheidsgraden worden in R berekend via de Welch–Satterthwaite benadering. Dat kan door in de `t.test` functie het argument `var.equal=FALSE` te zetten.

```
t.test(Staph~trt, data=oksel, var.equal=FALSE)
```

```
##  
##  Welch Two Sample t-test  
##  
## data:  Staph by trt  
## t = 4.7519, df = 17.876, p-value = 0.0001622  
## alternative hypothesis: true difference in means is not equal to zero  
## 95 percent confidence interval:  
##  9.976456 25.803544  
## sample estimates:  
## mean in group trt 1: transplant      mean in group trt 2: placebo  
##                                49.79                                31.1
```

Merk op dat we in de output zien dat een Welch T-test is uitgevoerd aan de titel boven de analyse.



## 5.8 Wat rapporteren?

- In de wetenschappelijke literatuur is er een overdreven aandacht voor p-waarden.
- Nochtans is het interessanter om een schatting te rapporteren samen met een betrouwbaarheidsinterval (dan met een p-waarde).

**Vuistregel:** Rapporteer een schatting steeds samen met een betrouwbaarheidsinterval (en een p-waarde), want

- ❶ Het resultaat van een toets kan veelal uit een betrouwbaarheidsinterval worden afgeleid;
- ❷ Dit laat toe om te oordelen of het resultaat ook **wetenschappelijk van belang** is.

## Reden 1: Relatie toetsen en betrouwbaarheidsintervallen

Stel dat we voor een zekere parameter  $\theta$  (bvb. een populatiegemiddelde, verschil in populatiegemiddelden, odds ratio, regressieparameter) de nulhypothese wensen te toetsen dat  $H_0 : \theta = \theta_0$  versus het alternatief  $H_A : \theta \neq \theta_0$  voor een zeker getal  $\theta_0$ .

Dan kan men aantonen dat men deze tweezijdige toetsingsprocedure kan uitvoeren op het  $\alpha 100\%$  significantieniveau door de nulhypothese te verwerpen als en slechts als het  $(1 - \alpha)100\%$  betrouwbaarheidsinterval voor  $\theta$  het getal  $\theta_0$  niet omvat.

M.a.w. het  $(1 - \alpha)100\%$  betrouwbaarheidsinterval voor  $\theta$  bevat alle getallen  $\theta_0$  zodat de tweezijdige toets van  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$  de nulhypothese niet verwerpt.

## Reden 2: Statistische significantie versus wetenschappelijke relevantie

Een betrouwbaarheidsinterval laat toe om zowel statistische significantie als wetenschappelijk belang van een resultaat te interpreteren.

Stel dat experimentele behandeling *significant betere* respons oplevert dan standaard/placebo. Een associatie is *statistisch significant* als  $P < \alpha$ , de data dragen m.a.w. voldoende bewijskracht om te besluiten dat er een associatie is. Dan blijft het mogelijk dat het effect *wetenschappelijk irrelevant* is. Met betrouwbaarheidsintervallen kunnen we dit wel evalueren.

Maar, dat laat echter nog veel subjectiviteit en manipulatie toe. Onderzoekers hopen in de praktijk immers wetenschappelijk belangrijke vondsten te maken en kunnen daarom geneigd zijn om hun oordeel over wat wetenschappelijk belangrijk is, wijzigen in functie van het bekomen betrouwbaarheidsinterval.

Om dit te vermijden is het wenselijk dat wetenschappers *a priori*, d.i. vooraleer de gegevens verzameld werden, hun oordeel over wetenschappelijke relevantie uitdrukken.

