Les 4 – Basis toetsen in R (1)

Emile Apol

Hanze University Groningen
Applied Sciences
Institute for
Life Science & Technology

Les 3

- Meetschalen
- Beschrijvende statistiek in R
- Stappenplan (verschil)toetsen
- Overzicht statistische toetsen
- *t*-toetsen
- *F*-toets

MEETSCHALEN

- Variabelen kunnen worden ingedeeld in 1 van de 4 onderstaande meetschalen
  - Nominaal
  - Ordinaal
  - Interval
  - Ratio
- R heeft functies voor verschillende meetschalen

Nominale schaal

- Kan gebruikt worden voor labels (nomen = “naam”)
- Bijvoorbeeld subtypen kanker in micro array experiment
- Alleen modus kan hierop toegepast worden
  - Meest voorkomende waarde
- Een nominale variabele heet in R een factor
- Verschillende “labels” heten levels
  - Bijv. de factor geslacht heeft 2 levels: man en vrouw

ORDINALE SCHAAL

- Ordinale variabele = nominale variabele maar met duidelijke volgorde van de levels
- Ordinale variabelen beschrijven alleen een volgorde, niet hoe groot het verschil is
  - Bijv. de ordinale variabele kwaliteit heeft als levels Beste, Goed, Redelijk, Slecht
- Je kan geen waarde geven aan het verschil tussen Goed en Beste.
- Naast de modus kan je ook de mediaan uitrekenen
  - Mediaan: middelste waarde in een gesorteerde lijst

INTERVAL SCHAAL

- De verschillende waarden hebben betekenis
- Er is geen echt nulpunt
- Bijvoorbeeld °C
- Verschil tussen 90°C en 100°C is gelijk aan verschil tussen 10°C en 20°C.
- Je kan optellen en aftrekken
- Je kan ook gemiddelde en SD uitrekenen.
- Je kan niet zeggen dat 20°C twee keer zo groot is als 10°C

Ratio schaal

- Er is wel een echt nulpunt
- Bijvoorbeeld graden Kelvin
  - Heeft echt nulpunt
  - 20K is 2 keer 10K
- Afstanden, energie, microarray meting

Meetschalen - overzicht

Niveau	Kenmerkend	Volgorde	Verschillen	Nulpunt
Nominaal	●			
Ordinaal	●	●		
Interval	●	●	●	
Ratio	●	●	●	●

Factor

- Factors in R kunnen gebruikt worden om variabelen in nominale of ordinale schaal op te slaan
- R kan er dan rekening mee houden dat ze in deze schalen staan
- Functies in R:
  - **`factor( )`**
  - **`as.factor( )`**

Factor

- ```
  geneExp <- c(
     rnorm(100, mean = 5, sd = 1),
     rnorm(100, mean = 6, sd = 2)
  )
  ```
- ```
  samples <- factor( c(
     rep("healthy", 100),
     rep("sick", 100)
  ) )
  ```
- `plot(geneExp ~ samples)`
- `plot(samples, geneExp)`

random getallen
$N(\mu, \sigma)$

~ betekent “als functie van”

FACTOR

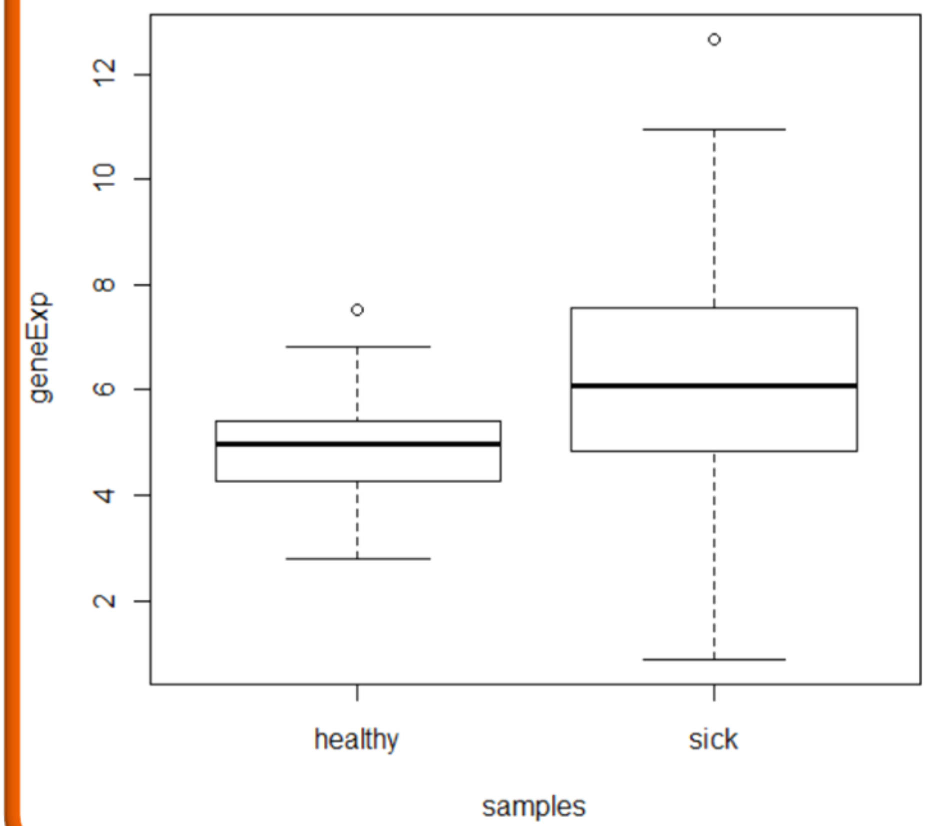
- Functies in R:
  - `factor( )`
  - `as.factor( )`
- De levels van de factor worden door R automatisch in alfabetische volgorde gezet (en dus ook in plot in die volgorde gezet!)
- Zelf aanpassen via optie levels binnen functie factor():
  - `g <- factor(c("B","B","A","A","B"))`

```
[1] B B A A B
Levels: A B
```

  - `g <- factor(c("B","B","A","A","B"), levels = c("B","A"))`

```
[1] B B A A B
Levels: B A
```

PLOT



STATISTIEK 1: BESCHRIJVENDE STATISTIEK

- Handige R functies voor beschrijvende statistiek:
  - `mean( )     # gemiddelde` $\overline{y}$
  - `median( )   # mediaan`
  - `var( )      # variantie` s^2
  - `sd( )       # standaarddeviatie` s
  - `min( )      # minimum`
  - `max( )      # maximum`
  - `quantile( )# kwantielen`
  - `IQR( )      # interquantile range`
  - `summary( ) # nuttige samenvatting`
  - `hist( )     # histogram`
  - `boxplot( ) # boxplot`

Herhaalde meting

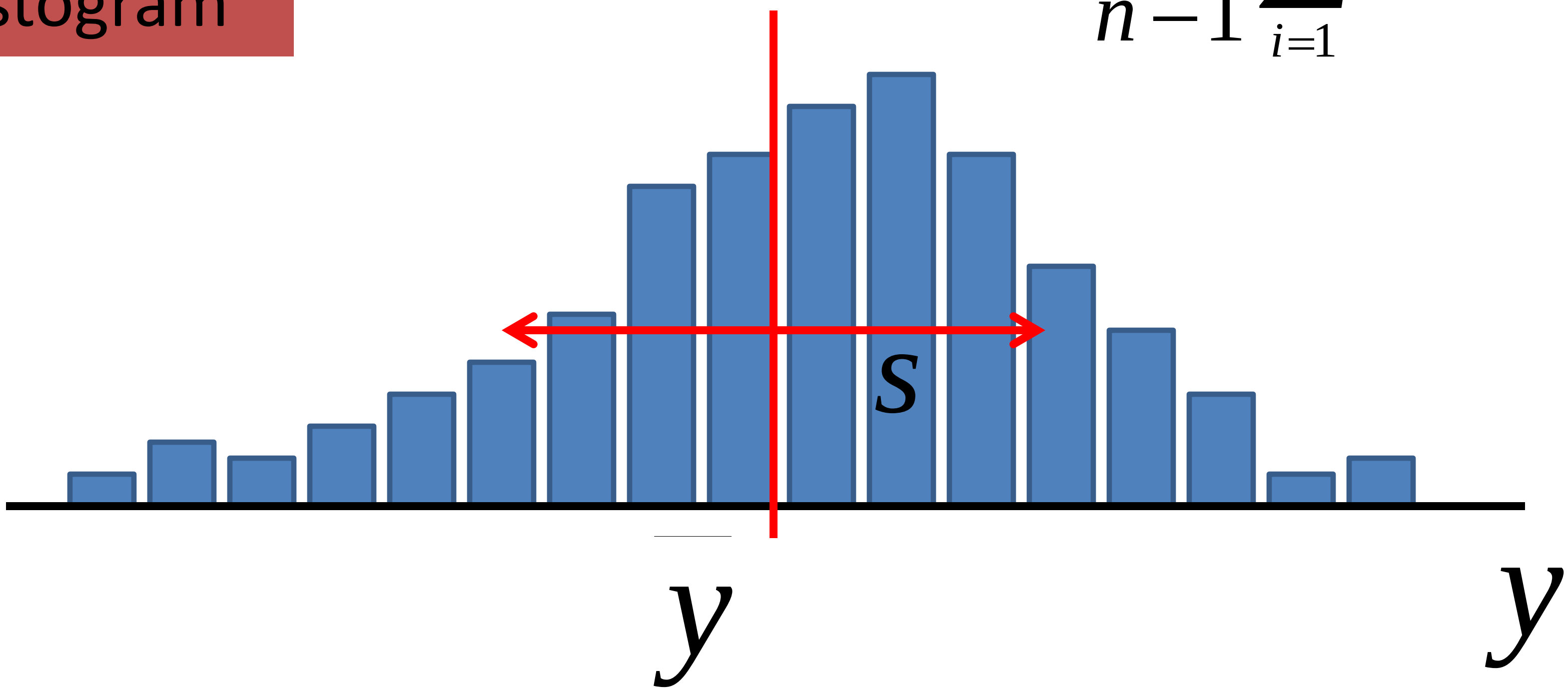
- Gemiddelde ($\rightarrow$ bias)
- Spreiding ($\rightarrow$ precisie)

som van i = 1 t/m i = n

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

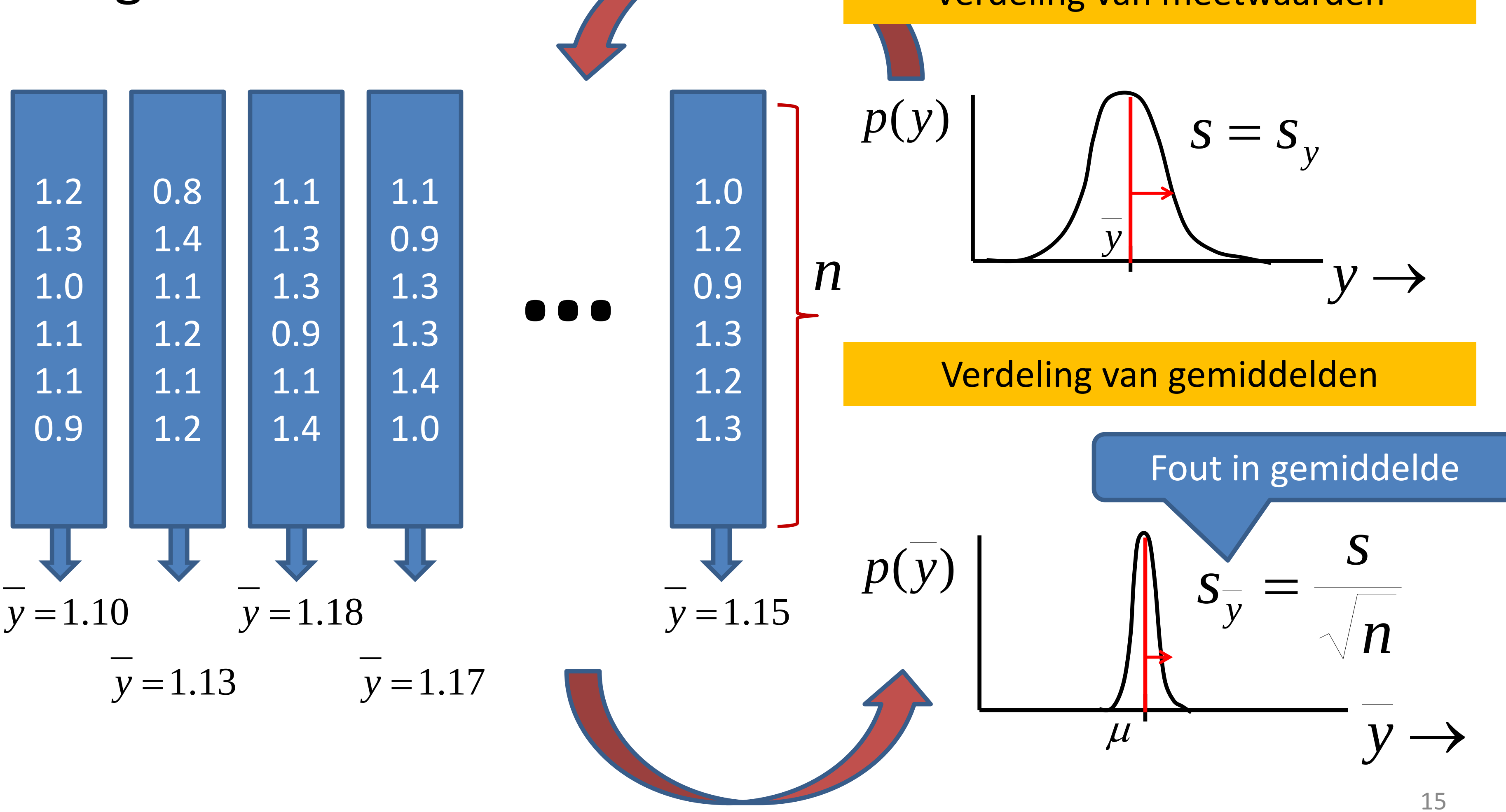
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

histogram



Gemiddelde en fout in gemiddelde

- Meerdere series van n metingen: meerdere gemiddelden



Fout in gemiddelde $s_{\bar{y}}$

- De fout in het gemiddelde = standaarddeviatie van de kansverdeling van gemiddelden…
- Voor de berekening van deze fout hoeven we niet vele meetseries te doen, maar slechts 1 serie met n metingen:

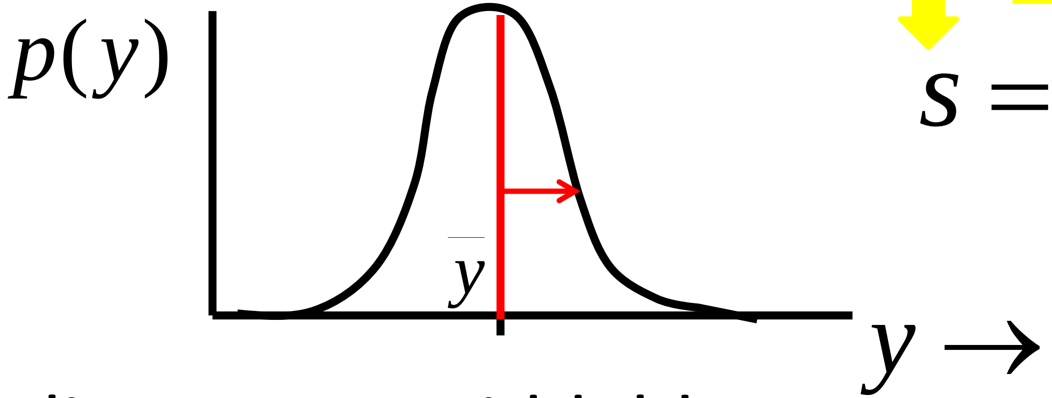
$$s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

Standaarddeviatie van n metingen

Fout in gemiddelde van n metingen

Gemiddelde, fout in gemiddelde

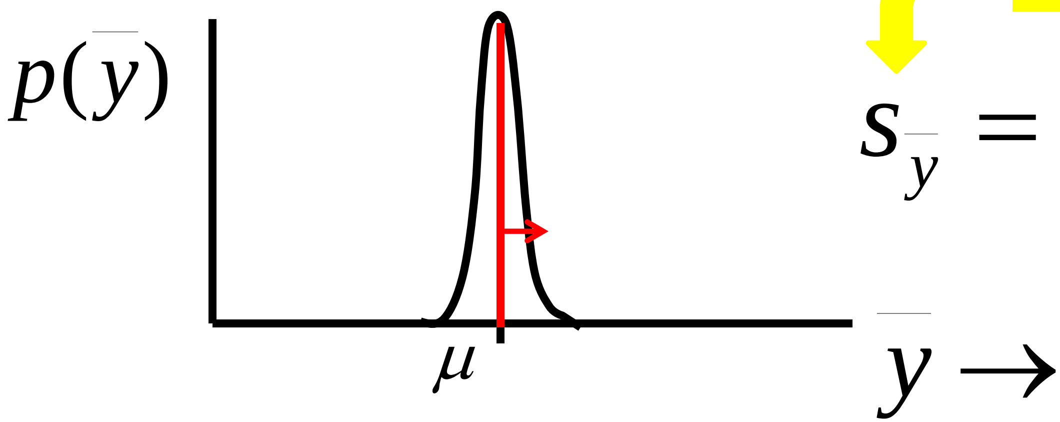
- Verdeling van meetwaarden



standaarddeviatie

$$s = s_y$$

- Verdeling van gemiddelde



fout in gemiddelde

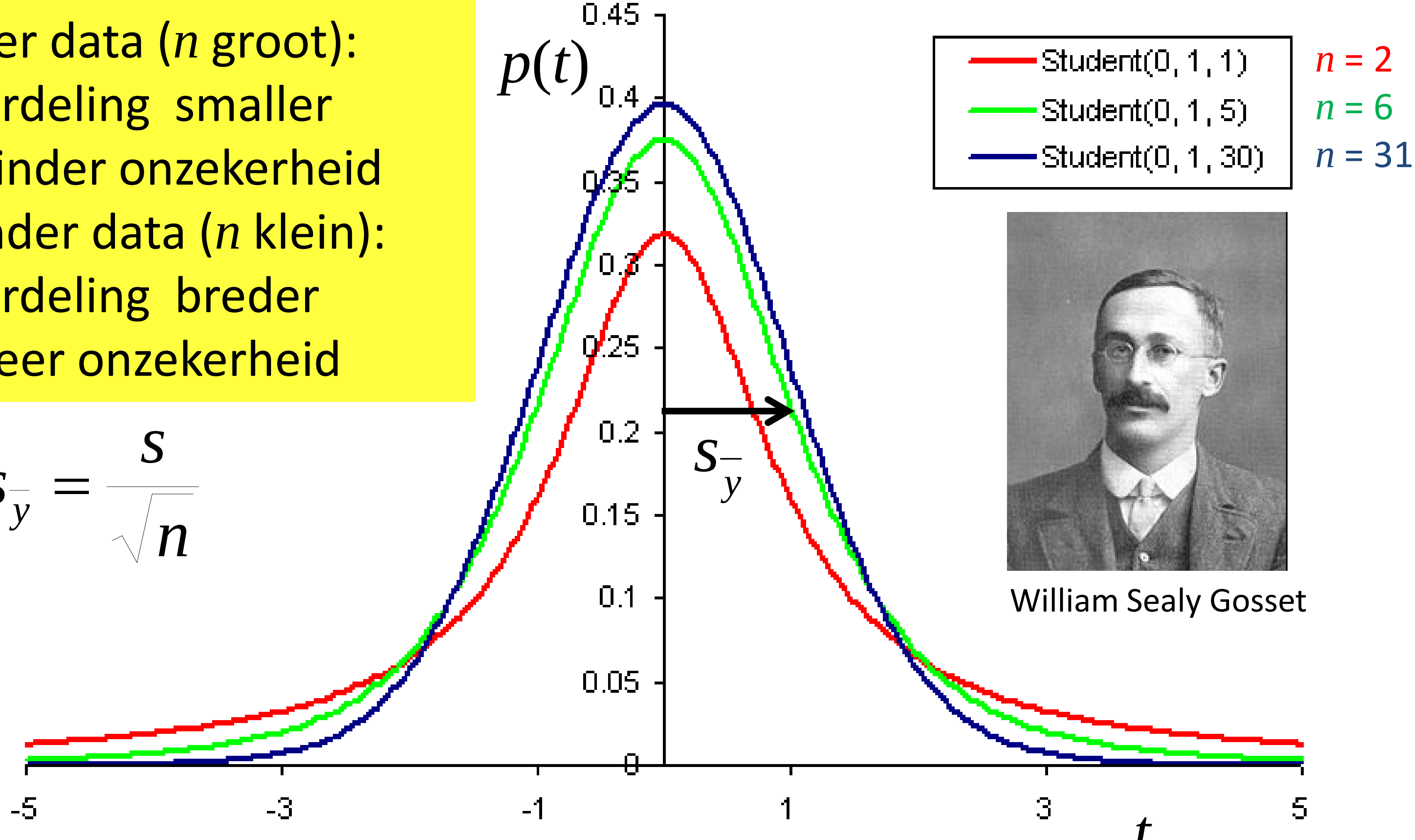
$$s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

Verdeling gemiddelde: *t*-verdeling

Meer data (*n* groot):
- verdeling smaller
- minder onzekerheid

Minder data (*n* klein):
- verdeling breder
- meer onzekerheid

$$s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

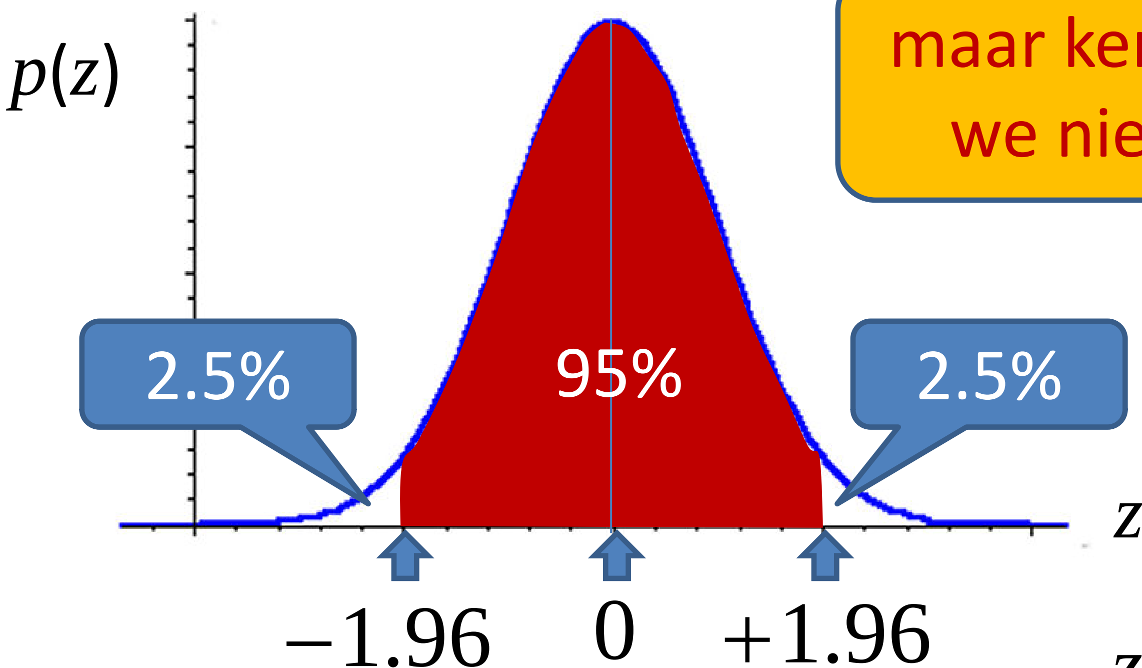


William Sealy Gosset

[vosesoftware.com]

Betrouwbaarheidsinterval: *z*-verdeling

Vraag: *als* de kansverdeling van gemiddelden een *normaal*verdeling *zou zijn*, in welk interval van waarden heb ik 95% kans dat het *echte* gemiddelde μ ligt?



maar kennen we niet...

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

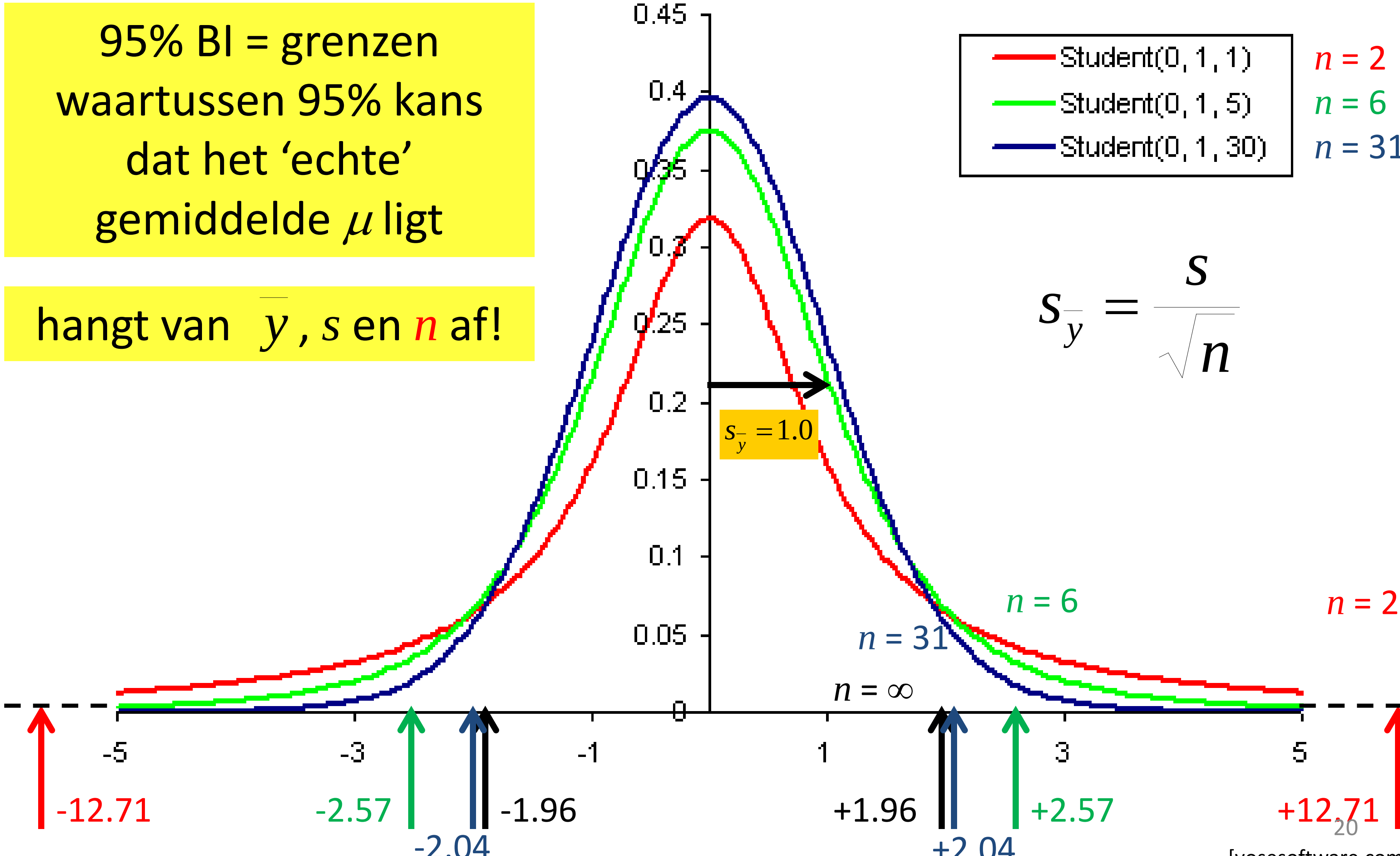
$$z = \frac{\bar{y} - \mu}{\sigma_{\bar{y}}}$$

Betrouwbaarheidsinterval: *t*-verdeling

95% BI = grenzen waartussen 95% kans dat het 'echte' gemiddelde μ ligt

hangt van $\bar{y}$, s en n af!

$$s_{\bar{y}} = \frac{s}{\sqrt{n}}$$

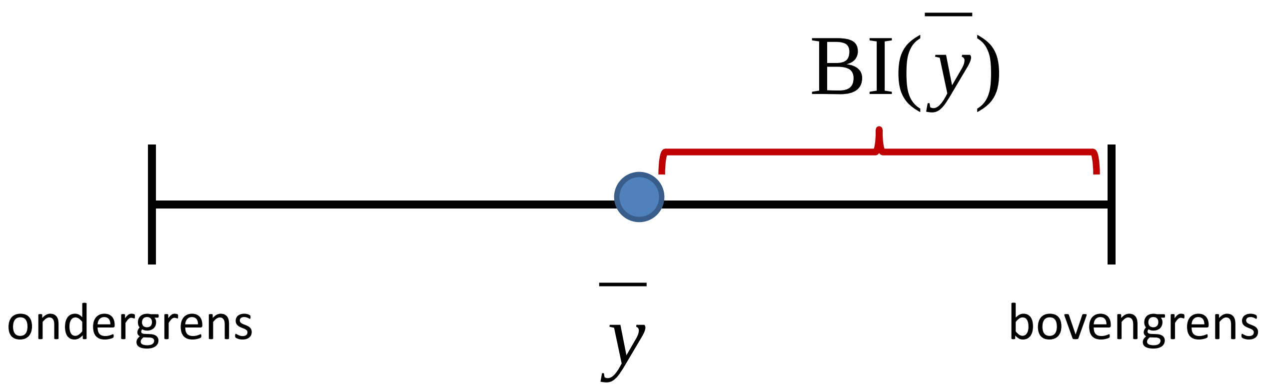


[vosesoftware.com]

95% Betrouwbaarheidsinterval (BI)

$$\bar{y} \pm \mathrm{BI}(\bar{y})$$

$$\mathrm{BI}(\bar{y}) = t_{n-1}^{0.05} \cdot s_{\bar{y}}$$

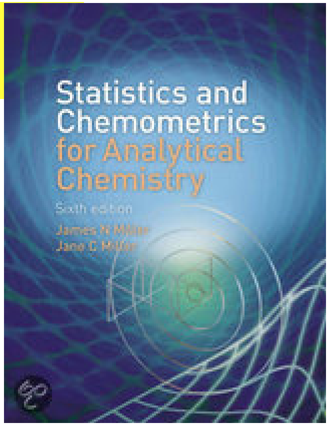


95% Betrouwbaarheidsinterval (BI)

95% BI (Engels: *Confidence Interval*, CI):

$$\bar{y} \pm t_{n-1}^{0.05} \times s_{\bar{y}} = \bar{y} \pm t_{n-1}^{0.05} \times \frac{s}{\sqrt{n}}$$

t-waarde, zie Miller & Miller, Tabel A.2



NB. De waarde n - 1 heet het aantal vrijheidsgraden (*degrees of freedom*, df)

STATISTIEK 2: STAPPENPLAN (VERSCHIL)TOETSEN

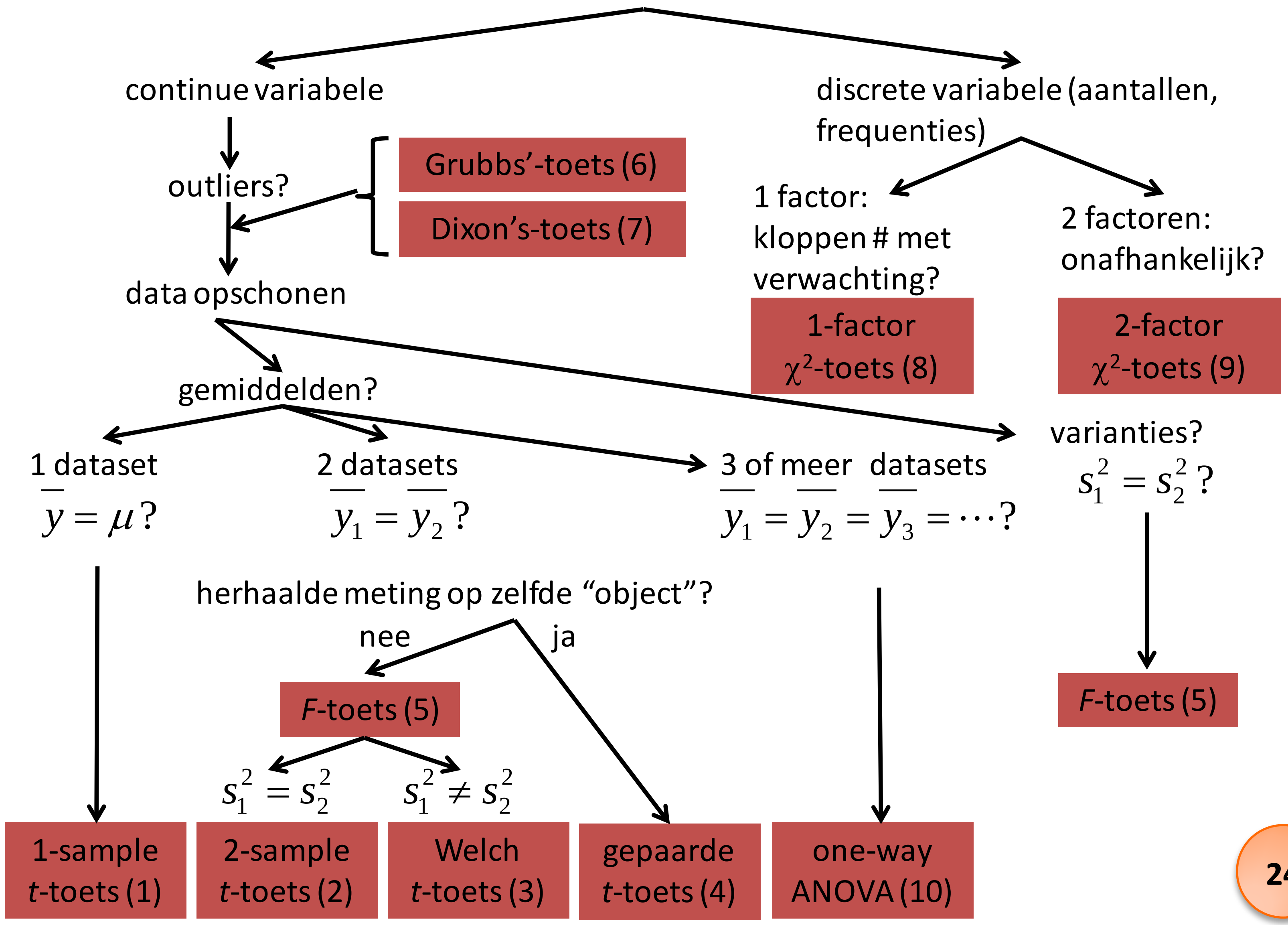
1. Formuleer de vraag helder
2. Kies op basis van de soort data de juiste toets
3. Formuleer nul-hypothese H_0 ("alles is gelijk")
4. Formuleer op basis van je vraag (achtergrond informatie) de alternatieve hypothese H_1 (of H_A)
   - 1-zijdig toetsen
   - 2-zijdig toetsen
5. Voer de toets uit: significant?

"kans dat H_0 waar is"

$$p < 0.05 = \alpha$$

6. Formuleer de conclusie in woorden

STATISTIEK 2: BESLISSCHEMA



Basale toetsen in R

- *t*-toets (verschil tussen 2 gemiddelden?)
  - **`t.test( )`**
- *F*-toets (verschil tussen 2 varianties?)
  - **`var.test( )`**
- 1-way ANOVA (verschil tussen $\geq$ 2 gemiddelden?)
  - **`aov( )`**
- chi2-toets (aantalen: relatie tussen 2 nominale variabelen?)
  - **`chisq.test( )`**
- *z*-toets (standaard normaal verdeeld)
  - **`z.test( )`**   **`# in`** package:TeachingDemos

Basale toetsen in R

- Grubbs' en Dixon toets (waarde is uitbijter?):
  - `grubbs.test( ) # in` package:outliers
  - `dixon.test( )  # in` package:outliers

T-TOETSEN

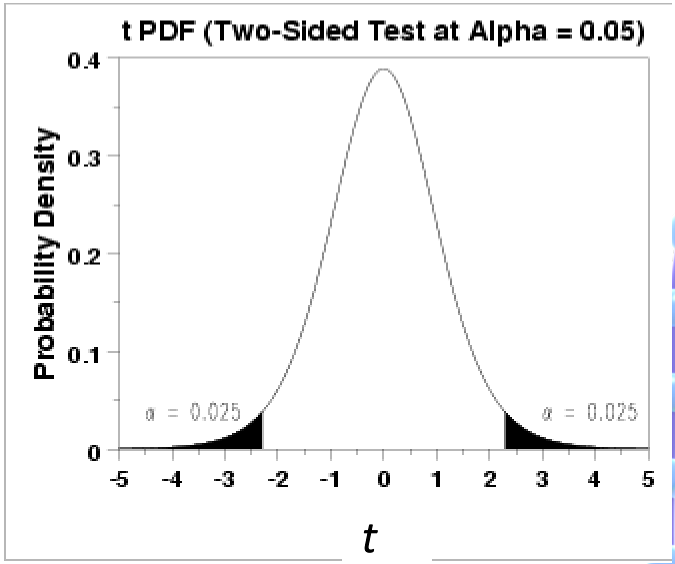
Als

- je verwacht dat twee gemiddelden gelijk zijn (H_0)
- de random fouten normaal verdeeld zijn

dan

is de grootheid (statistiek)

$$t = \frac{\text{verschil tussen gemiddelden}}{\text{fout in verschil}}$$



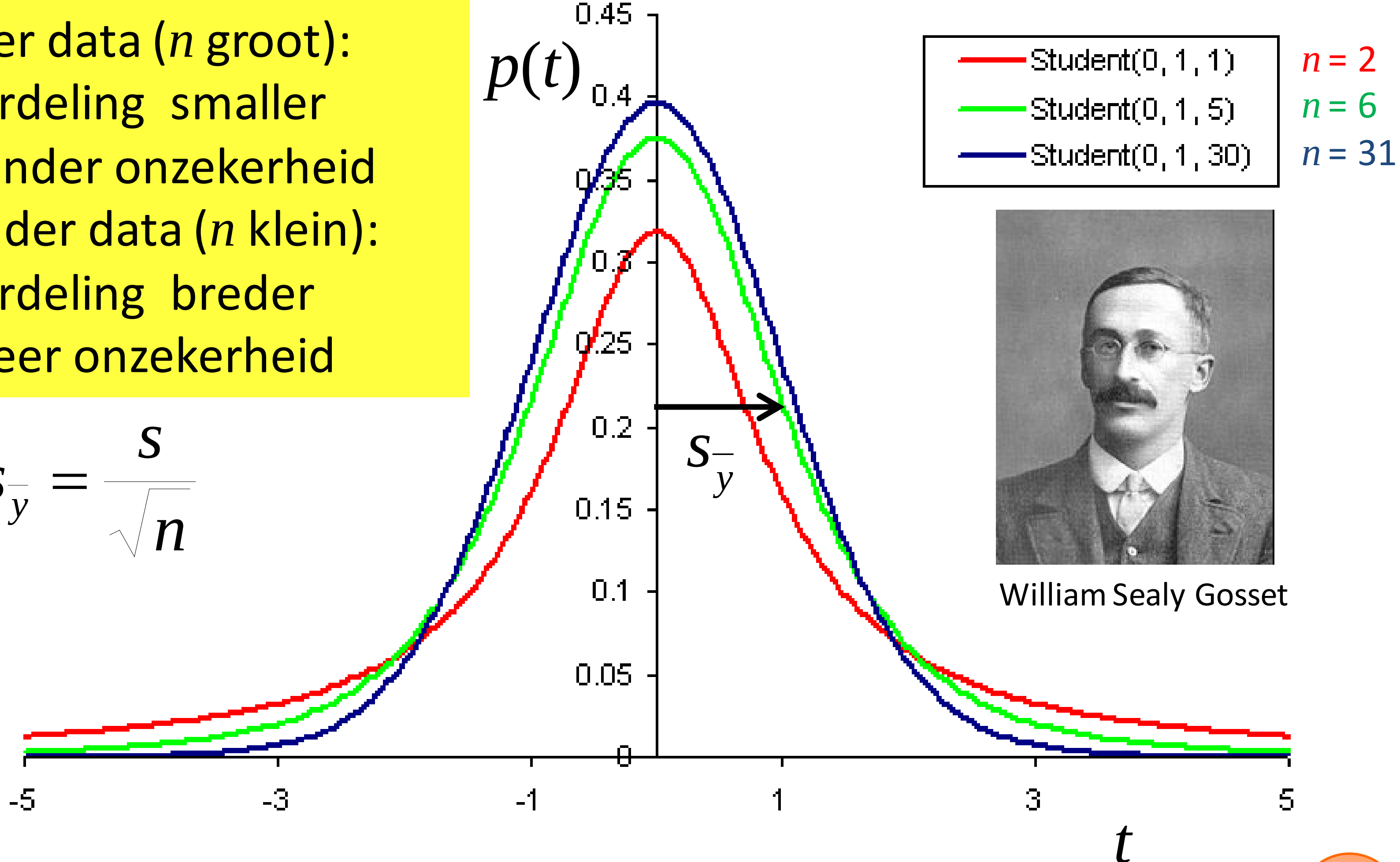
t-verdeeld, met d.f. vrijheidsgraden

(STUDENT'S) T-VERDELING

Meer data (n groot):
- verdeling smaller
- minder onzekerheid

Minder data (n klein):
- verdeling breder
- meer onzekerheid

$$s_{\bar{y}} = \frac{s}{\sqrt{n}}$$



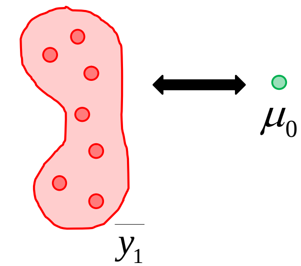
William Sealy Gosset

T-TOETSEN: 1-SAMPLE, 2-SAMPLE, WELCH T-TOETS

- One-sample *t*-toets:

$$t = \frac{\overline{y_1} - \mu}{s\sqrt{\frac{1}{n}}}$$

$$\mathrm{df} = n - 1$$

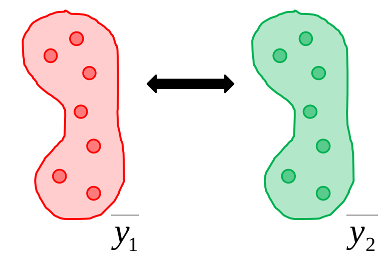


- Two-sample *t*-toets:

$$t = \frac{\overline{y_1} - \overline{y_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\mathrm{df} = n_1 + n_2 - 2$$

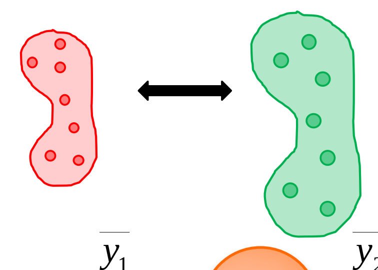
$$s_p^2 = \frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 + n_2 - 2}$$



- Welch *t*-toets:

$$t = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$



T-TOETSEN IN R: 1-SAMPLE, 2-SAMPLE, WELCH T-TOETS

- Twee vectoren **`y1`** en **`y2`** met continue waarden (interval/ratio):
  - `y1 <- c(1, 2, 5, 4, 2)`
  - `y2 <- c(4, 5, 8, 6, 7)`
- Uitvoeren van *t*-toetsen:
- One-sample *t*-toets:
  - `t.test(y1, mu = 5.0)`
- Two-sample *t*-toets:
  - `t.test(y1, y2 , var.equal=T)`
- Welch *t*-toets:
  - `t.test(y1, y2)`

T-TOETSEN: 1- OF 2-ZIJDIG

Hangt van vraagstelling (en/of voorkennis) af:

- Zijn twee waarden significant verschillend?

  Je weet niet welke kant het verschil gaat

  2-zijdig

- Is de ene waarde hoger/lager/beter/slechter/… dan de andere waarde?

  Je hebt een sterk vermoeden welke kant het verschil gaat

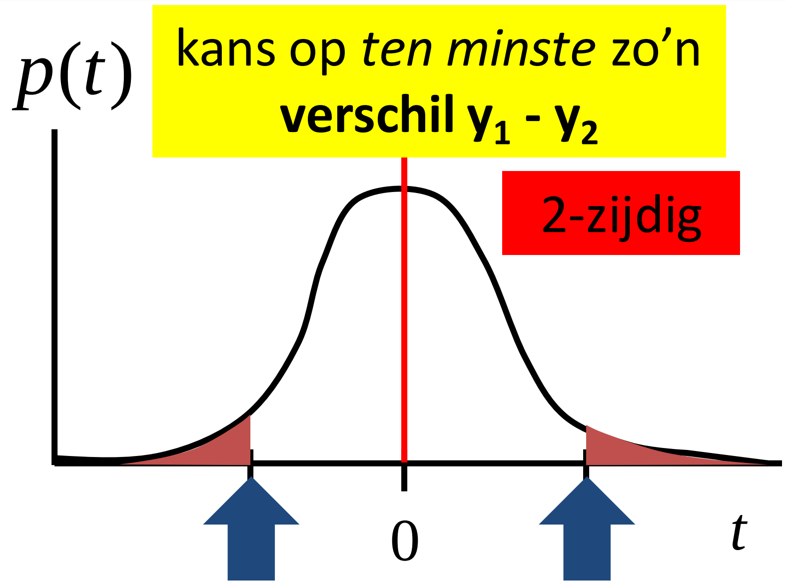
  1-zijdig

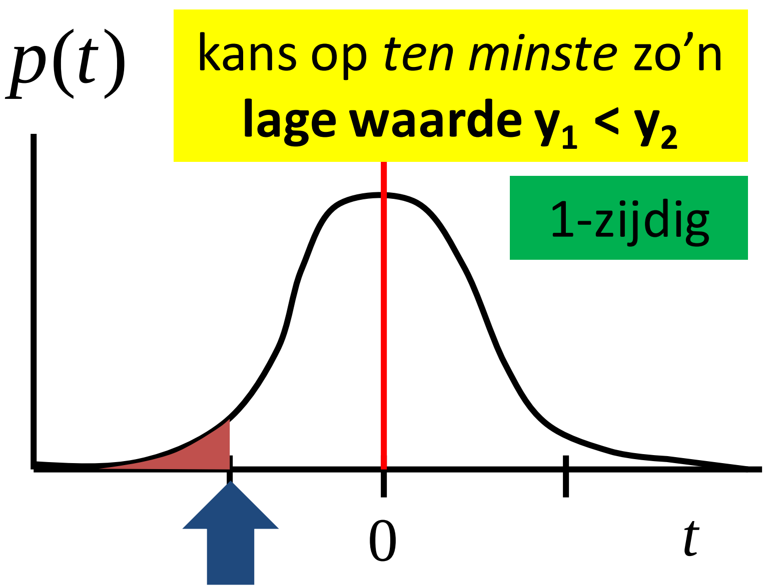
Dit wordt dus bepaald door H_1 !

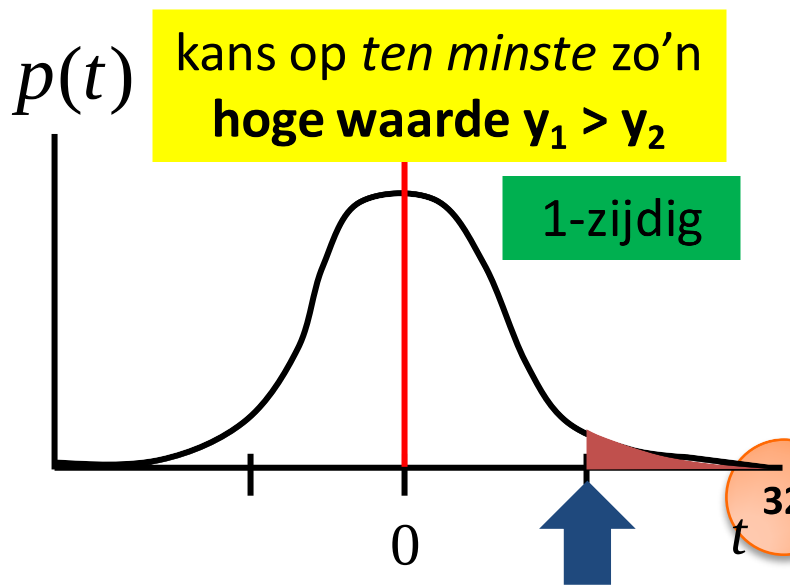
T-TOETSEN: 1- OF 2-ZIJDIG

1- of 2-zijdige kansen?

$$t = \frac{\text{verschil tussen gemiddelden}}{\text{fout in verschil}}$$







T-TOETSEN IN R: 1- OF 2-ZIJDIG

- Vraag: is gemiddelde van y1 *anders* dan van y2 (μ)?
  2-zijdig toetsen
  - `t.test(y1, y2, …, alternative = "two.sided") #default`
  - `t.test(y1, y2, …)`

  $y_1 \neq y_2$

- Vraag: is gemiddelde van y1 *lager* dan van y2 (μ)?
  1-zijdig toetsen
  - `t.test(y1, y2, …, alternative = "less")`

  $y_1 < y_2$

- Vraag: is gemiddelde van y1 *hoger* dan van y2 (μ)?
  1-zijdig toetsen
  - `t.test(y1, y2, …, alternative = "greater")`

  $y_1 > y_2$

T-TOETSEN IN R: 2 VECTOREN OF VECTOR + FACTOR

- Data: twee vectoren **`y1`** en **`y2`** met continue waarden (interval/ratio):
  - `y1 <- c(1, 2, 5, 4, 2)`
  - `y2 <- c(4, 5, 8, 6, 7)`
- Uitvoeren van *t*-toets:
  - `t.test(y1, y2, …)`
- Data: vector **`y`** met continue waarden en vector **`sample`** met labels (= factor):
  - `y <- c(1,2,5,4,2,4,5,8,6,7)`
  - `sample <- factor(c(rep(1,5),rep(2,5)))`
- Uitvoeren van *t*-toets:
  - `t.test(y ~ sample, …)` "model formule"
  - `t.test(y[sample==1], y[sample==2], …)`

T-TOETSEN IN R: VECTOR + FACTOR

- Bij gebruik van een model formule als
  - **`y ~ sample`**

  in 1-zijdige toets moet je bedenken dat R de levels in **`sample`** in *logische volgorde* neemt, dus
  - **`sample <- c(1, 2, 1, 1, 2, 2)`**

  geeft groepen "**1**" en "**2**"
  - **`sample <- c(2, 2, 1, 2, 1, 1)`**

  geeft groepen "1" en "2"
  - **`sample <- c("low,"low","high","high")`**

  geeft groepen: "**h**igh" en "low"

H_1: 1^e groep "less"/"greater" 2^e groep

T-TOETSEN IN R: VECTOR + FACTOR

- Als factor uit meer dan 2 levels bestaat: gebruik **subset**
  - `y <-c(0.15,0.21,0.13,0.15,0.16,0.17)`
  - `sample <- c("a","a","b","b","c","c")`
- 2-sample *t*-toets tussen groepen “a” en “b”:
  - `t.test(y[sample=="a"], y[sample=="b"], var.equal=T)`
  - `t.test(y ~ sample, var.equal=T, subset=(sample=="a" | sample=="b") )`
  - `t.test(y ~ sample, var.equal=T, subset=(sample != "c") )`

Veel functies in R, zoals t.test, var.test, aov, lm, chisq.test etc., kunnen werken met **subset**!

T-TOETSEN IN R: DATAFRAMES (1)

- Twee soorten dataframes:
  - `data1 <- data.frame(y1, y2)`
  - `data2 <- data.frame(y, sample)`
  - `rm(y1, y2, y, sample)  # uit Workspace`
- Zonder attachen van data1 of data2:
  - `t.test(y ~ sample, data=data2)`
- Met attachen van data1 of data2:
  - `attach(data1)`
  - `attach(data2)`
  - `t.test(y1, y2)`
  - `t.test(y ~ sample)`
  - `detach(data1)`
  - `detach(data2)`

Werkt alleen met de “formule” manier ~

T-TOETSEN IN R: DATAFRAMES (2)

- Zonder attachen, met gebruik van de functie `with( )`:
  - `with(data=data.1, t.test(y1,y2,…))`
  - `with(data=data.2, t.test(y~sample,…))`

"Scope" van variabelen in R

- R heeft een search path voor objecten/variabelen:
  - **`search()`**

```
> search()
 [1] ".GlobalEnv"         "package:VGAM"       "package:stats4"     "package:splines"
 [5] "package:mvtnorm"    "genExpr.2"          "genExpr.1"          "tools:rstudio"
 [9] "package:stats"      "package:graphics"   "package:grDevices"  "package:utils"
[13] "package:datasets"   "package:methods"    "Autoloads"          "package:base"
```

- R zoekt *vanaf begin van deze lijst* totdat object gevonden is.
- Volgorde:

zelf gedefinieerde variabelen

  - 1e positie: altijd .GlobalEnv
  - laatste positie: altijd package:base
  - 2e positie: altijd laden nieuw package / attachen dataframe

TYPE I EN TYPE II ERRORS

- Onderscheid tussen Type I en Type II errors

		Waarheid	
		H_0 is waar	H_0 is niet waar
Uitkomst toets	H_0 niet verworpen	true negatives (goede beslissing) $1 - \alpha$	false negatives (Type II error) β
	H_0 verworpen	false positives (Type I error) α	true positives (goede beslissing) $1 - \beta$

Deze kans stellen we van te voren in, bijv. $\alpha = 0.05$

Dit is de "power" van een statistische toets

T-TOETSEN IN R

```
> t.test(y1, y2, var.equal=T)

 Two Sample t-test

data:  y1 and y2
t = -3.1379, df = 8, p-value = 0.01385
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.551672 -0.848328
sample estimates:
mean of x mean of y
      2.8       6.0
```

Zijn beide gemiddelden gelijk?

T-TOETSEN IN R

```
> t.test(y1, y2, var.equal=T)

 Two Sample t-test

data:  y1 and y2
t = -3.1379, df = 8, p-value = 0.01385
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.551672 -0.848328
sample estimates:
mean of x mean of y 
      2.8       6.0 
```

Nee, beide gemiddelden verschillen significant, want $p < 0.05$

T-TOETSEN IN R: OUTPUT

- De uitvoer van `t.test( )` is een list met allerlei nuttige gegevens:
  - `testOut <- t.test(y1, y2)`
  - `str(testOut)`
- Elementen van uitvoer:

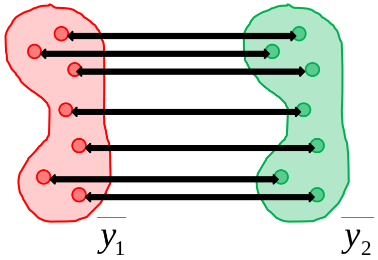
Element	Betekenis	Element	Betekenis
$statistic	waarde *t*	$null.value	verwacht verschil
$parameter	df	$alternative	1- of 2-zijdig
$p.value	*p*-waarde	$method	soort *t*-toets
$conf.int	BI van verschil	$data.name	data
$estimate	gemiddelden		

T-TOETSEN: GEPAARDE T-TOETS

- Gepaarde *t*-toets:

$$t = \frac{\bar{d} - 0}{s_d \sqrt{\frac{1}{n_d}}}$$

$$\mathrm{df} = n_d - 1$$



met

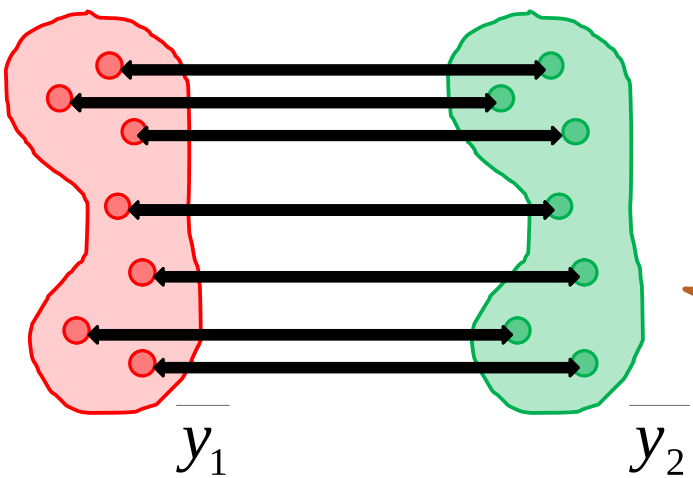
$$d_i = y_{1,i} - y_{2,i}$$

en n_d = aantal paren, s_d = standaard deviatie van verschillen d_i

- Gebruik:
  - weg filteren van extra ruisfactor(bijv. natuurlijke verschillen tussen proefpersonen/samples/etc.
  - = block ANOVA voor 2 groepen

T-TOETSEN IN R: GEPAARDE T-TOETS

- Voor gepaarde data (bijv. herhaalde meting bij verschillende patiënten onder 2 omstandigheden):
  - `t.test(y1, y2, paired=T)`
  - `t.test(y ~ sample, paired=T)`
  - `t.test(y[sample==1], y[sample==2], paired=T)`



Beide groepen moeten evenveel data bevatten!

Met een gepaarde *t*-toets filter je extra ruis uit je signaal!

F-TOETS

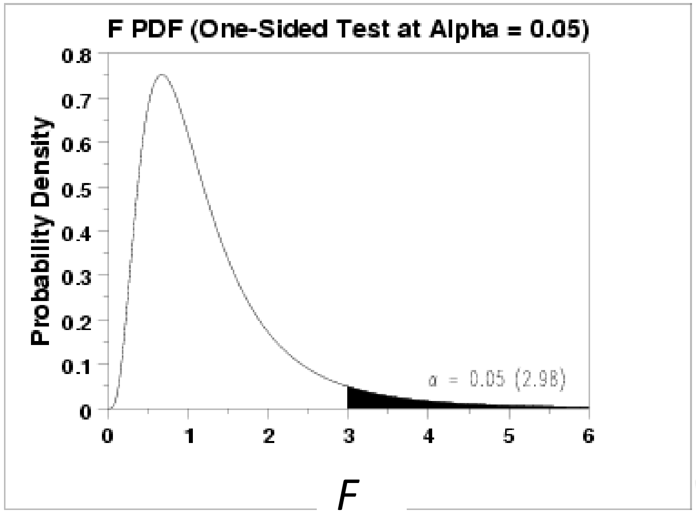
Als

- twee variabelen normaal verdeeld zijn: $N(\mu_1, \sigma_1)$ en $N(\mu_2, \sigma_2)$
- je verwacht dat beide varianties σ_1^2 en σ_2^2 gelijk zijn (H_0)

dan

is de grootheid (statistiek)

$$F = \frac{s_A^2}{s_B^2} \qquad s_A^2 > s_B^2$$



F-verdeeld, met df1 en df2 vrijheidsgraden

F-TOETS IN R

- Vraag: is variantie van **`y1`** anders dan variantie van **`y2`**?
  2-zijdig
  - **`var.test(y1, y2)`**
  - **`var.test(y ~ sample)`**

Gebruik je o.a. om te kiezen tussen 2-sample en Welch *t*-toets!

- Vraag: is variantie van **`y1`** groter dan van **`y2`**?
  1-zijdig
  - **`var.test(y1, y2, alternative="greater")`**
- Vraag: is variantie van **`y1`** kleiner dan van **`y2`**?
  1-zijdig
  - **`var.test(y1, y2, alternative="less")`**

F-TOETS IN R

```
> var.test(y1,y2)

 F test to compare two variances

data:  y1 and y2
F = 1.08, num df = 4, denom df = 4, p-value = 0.9423
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.1124469 10.3728923
sample estimates:
ratio of variances
              1.08
```

Zijn beide varianties gelijk?

F-TOETS IN R

```
> var.test(y1,y2)

 F test to compare two variances

data:  y1 and y2
F = 1.08, num df = 4, denom df = 4, p-value = 0.9423
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.1124469 10.3728923
sample estimates:
ratio of variances
              1.08
```

Ja, beide varianties zijn gelijk, want $p > 0.05$

F-TOETS IN R

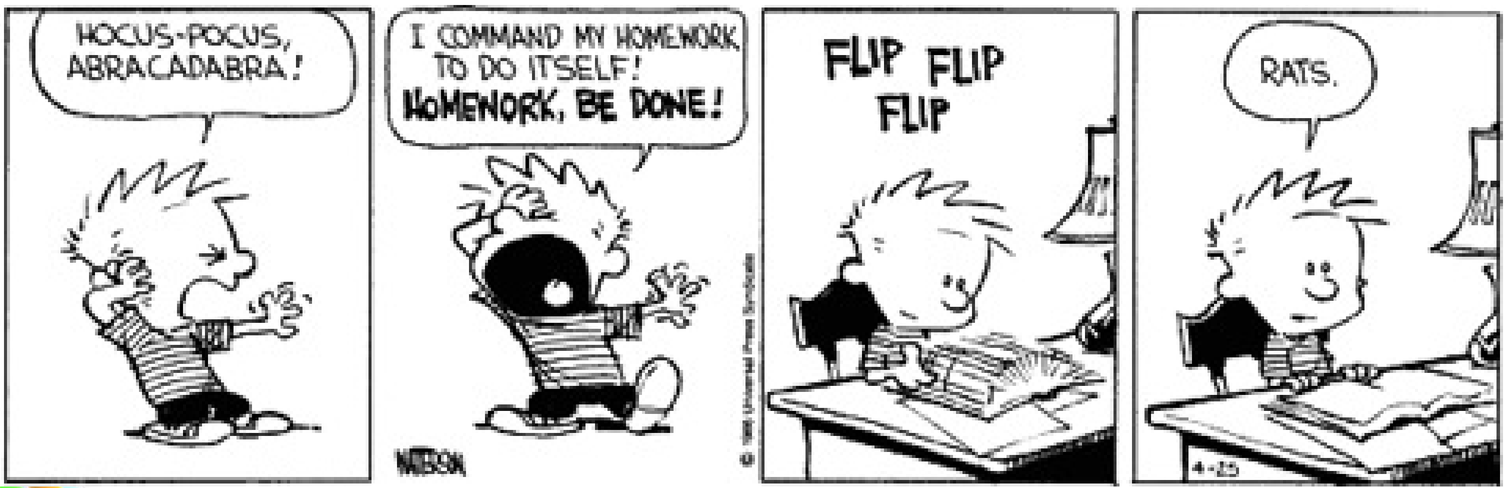
```
> var.test(y1,y2)

 F test to compare two variances

data:  y1 and y2
F = 1.08, num df = 4, denom df = 4, p-value = 0.9423
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.1124469 10.3728923
sample estimates:
ratio of variances
              1.08
```

Maar: absence of proof is NOT proof of absence!

Ja, beide varianties zijn gelijk, want $p > 0.05$

"Beide varianties zijn niet significant ongelijk"

Jullie kunnen nu (een deel van) de opdrachten van les 3 maken



Hanze University Groningen
Applied Sciences
Institute for Life Science & Technology