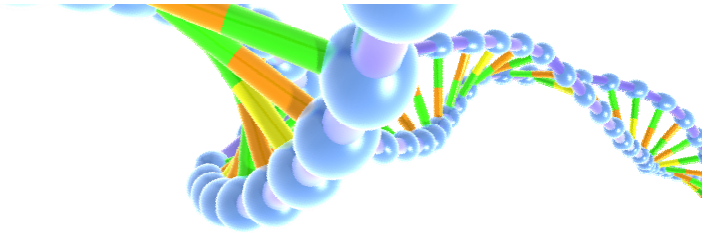
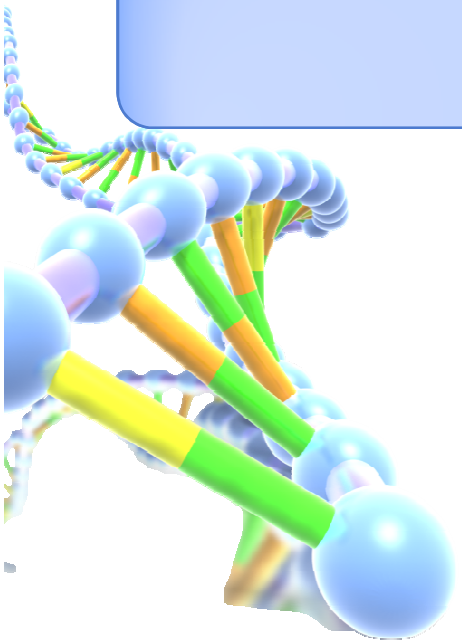


Les 6 – Basis toetsen in R (4)

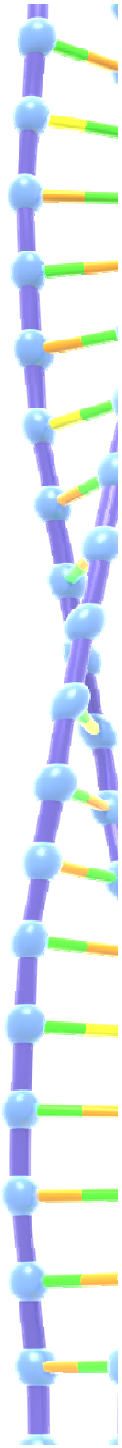
Emile Apol, 2013-2014



Hanze University Groningen
Applied Sciences
Institute for
Life Science & Technology

Les 4

- Stappenplan (verschil)toetsen
- Overzicht statistische toetsen
- 1-way ANOVA
- Formules in R



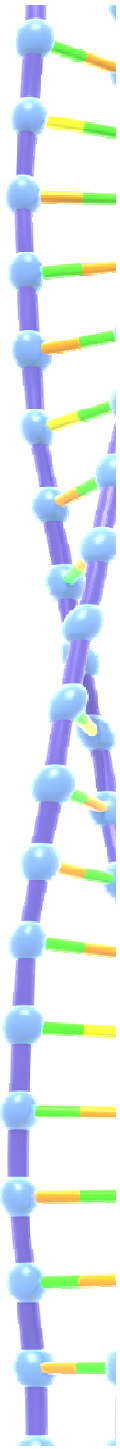
STATISTIEK 2: STAPPENPLAN (VERSCHIL)TOETSEN

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?

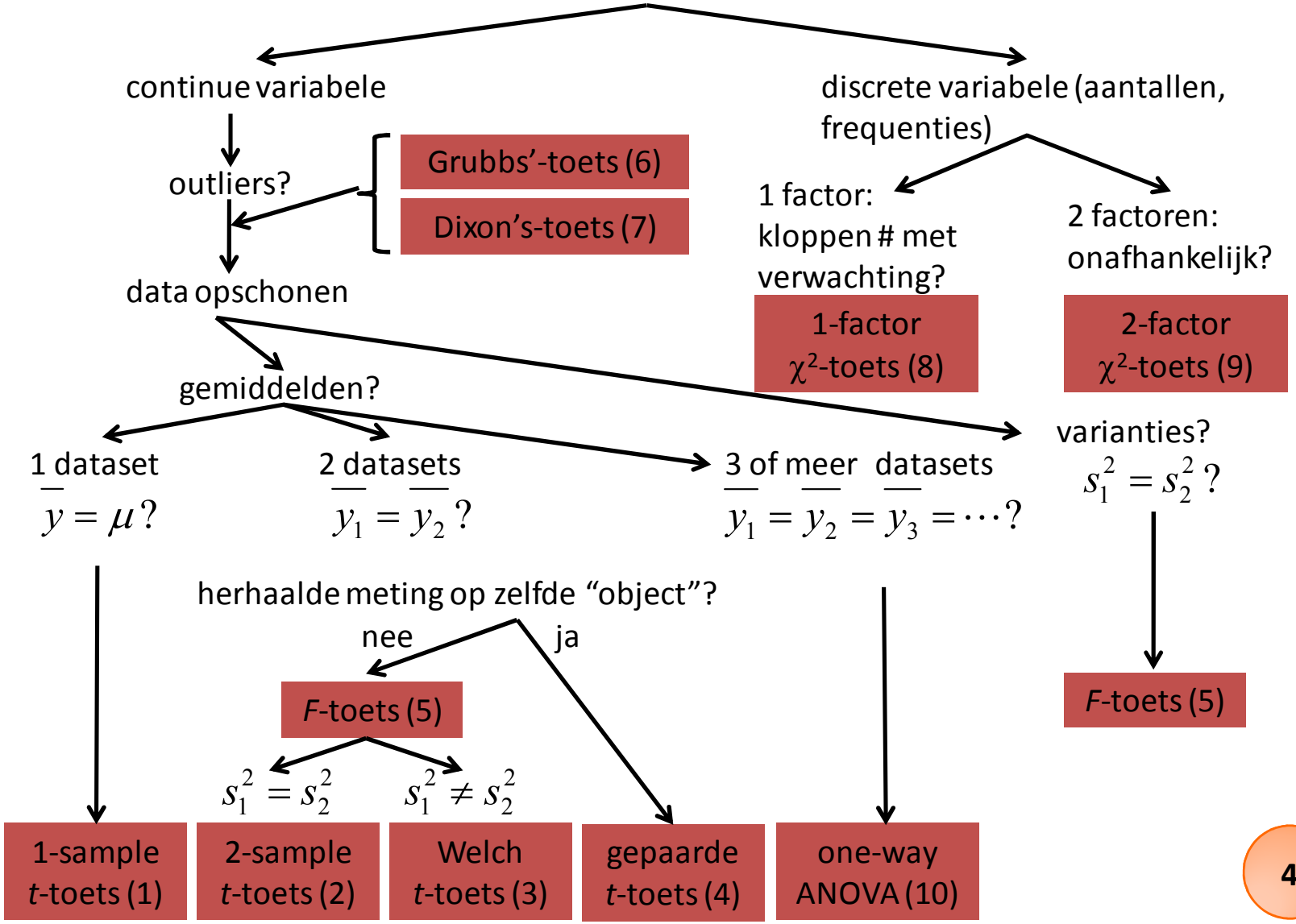
“kans dat H_0 waar is”

$$p < 0.05 = \alpha$$

6. Formuleer de conclusie in woorden



STATISTIEK 2: BESLISSCHEMA



2 LEVELS VS MEERDERE LEVELS

- Vergelijken van gemiddelden:

2 levels (groepen)

t-toets

`t.test()`

≥2 levels (groepen)

1-way ANOVA

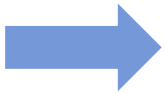
`aov()`

- Vergelijken van varianties:

2 levels (groepen)

F-toets

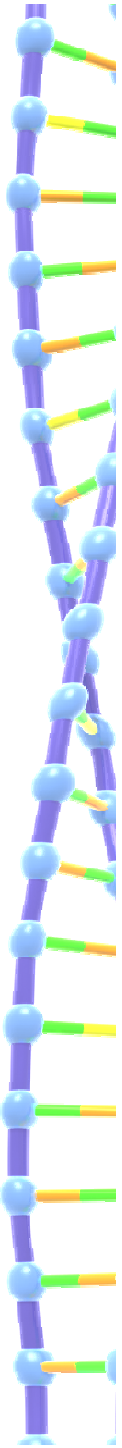
`var.test()`



≥ 2 levels (groepen)

Bartlett-toets

`bartlett.test()`

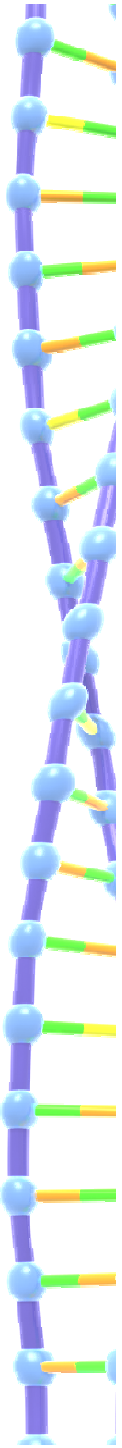


MEERDERE LEVELS: MEERDERE T-TOETSEN?

- Stel: k = 5 groepen (levels)
- Mogelijkheid: $R = \frac{1}{2}\, k\,(k - 1) = 10$ onderlinge *t*-toetsen:
  - `y <- c(y1, y2, y3, y4, y5)`
  - `sample <- factor(c(rep(1,5),…, rep(5,5)))`
  - `pairwise.t.test(y, sample, pool.sd=F, p.adjust.method="none")`
- Maar kans op "ergens" verkeerde beslissing:

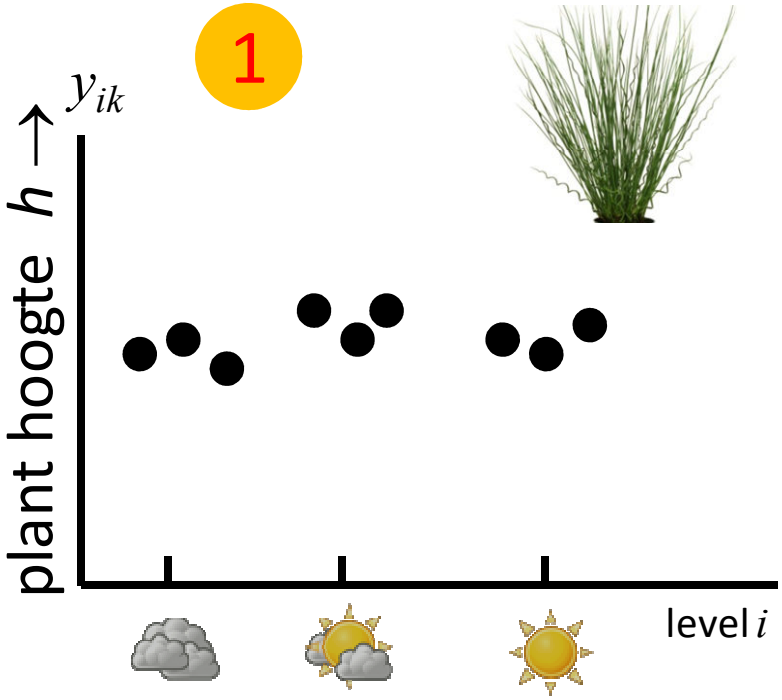
$$\text{kans} = 1 - (1 - \alpha)^R = 1 - (1 - 0.05)^{10} = 0.40$$

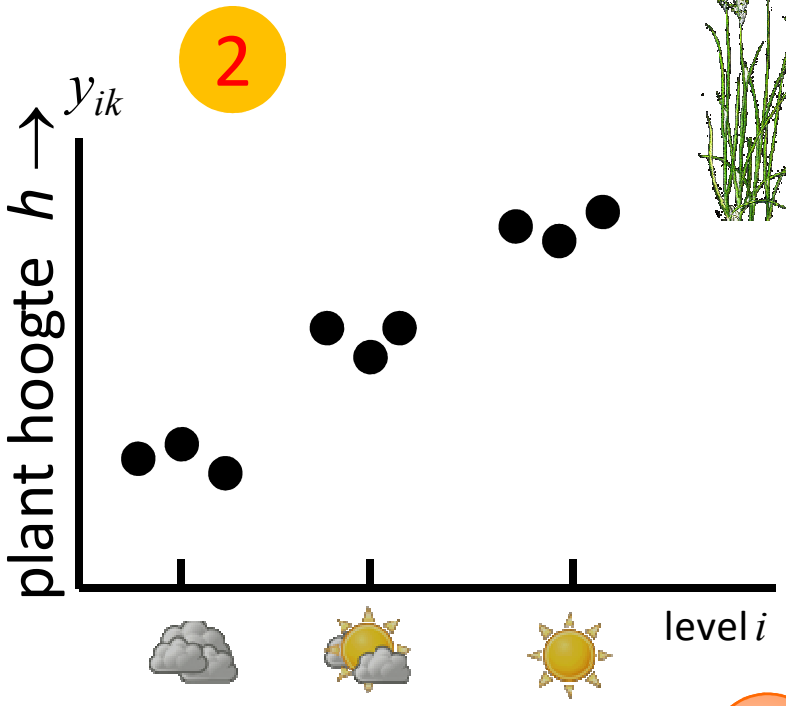
Dit is de "family-wise" error rate



1-WAY ANOVA

Waarom vind je in situatie 1 dat er geen effect van licht is op de plant hoogte, terwijl je dat in situatie 2 wel vindt?

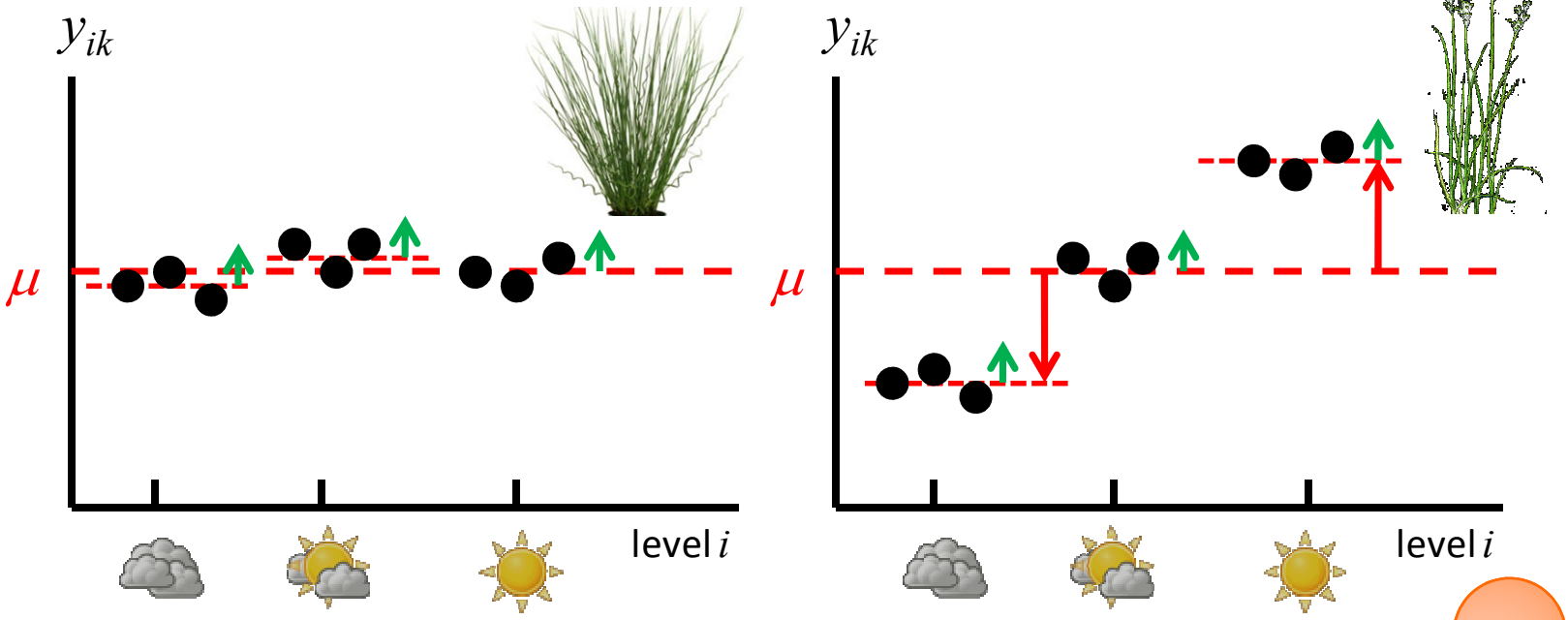




1-WAY ANOVA

Verhouding tussen spreiding (= variantie) *tussen* levels (groepen) en *binnen* levels:

$$F = \frac{s^2_{\text{tussen}}}{s^2_{\text{binnen}}} = \frac{MS_{\text{tussen}}}{MS_{\text{binnen}}} = \frac{MS_{\text{A}}}{MS_{\text{error}}} = \frac{\text{var}(\uparrow)}{\text{var}(\uparrow)}$$



1-WAY ANOVA

- Model: $y_{ik} = \mu + \alpha_i + \varepsilon_{ik}$

effect van factor 1

error (ruis)

$$y_{ik} = \bar{y}_{\bullet\bullet} + \left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right) + \left(y_{ik} - \bar{y}_{i\bullet}\right)$$

- *Sum of Squares* ("variatie") opsplitsing:

$$\underbrace{\sum_{i=1}^{q}\sum_{k=1}^{n_c}\left(y_{ik} - \bar{y}_{\bullet\bullet}\right)^2}_{SS_{\text{tot}}} = \underbrace{n_c\sum_{i=1}^{q}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2}_{SS_{\text{tussen}}} + \underbrace{\sum_{i=1}^{q}\sum_{k=1}^{n_c}\left(y_{ik} - \bar{y}_{i\bullet}\right)^2}_{SS_{\text{binnen}}}$$

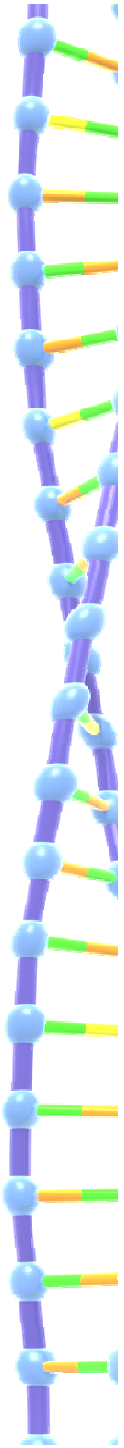
1-WAY ANOVA

H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_q$

of wel $\alpha_1 = \alpha_2 = \alpha_3 = \ldots = \alpha_q = 0$

(alle levels hetzelfde gemiddelde)

H_1: niet zo…



1-WAY ANOVA

- Op basis van SS_{tussen} en SS_{binnen} bereken MS's:

var(↑) $MS_{\text{tussen}} = \dfrac{SS_{\text{tussen}}}{q-1}$ $\quad$ $MS_{\text{binnen}} = \dfrac{SS_{\text{binnen}}}{q(n_c-1)}$ var(↑)

- Verhouding MS_{tussen} / MS_{binnen} volgt F- verdeling:

$$F = \frac{MS_{\text{tussen}}}{MS_{\text{binnen}}}; \quad \text{verwachting } F_{\text{krit}} = F_{q-1,q(n_c-1)}$$

als $F > F_{\text{krit}}$: sign. verschil tussen levels factor 1

òf als $p < 0.05$

(eenzijdig toetsen)

1-WAY ANOVA: ANOVA TABEL

- Standaard manier van weergave ANOVA:

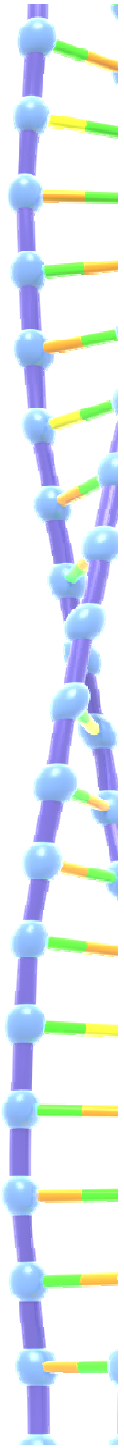
factor

error / residuals

source	SS	df	MS	F	*p*-value
between groups	0.3089	2	0.1544	19.86	0.00226
within groups	0.0467	6	0.0078		
total	0.3556	8			

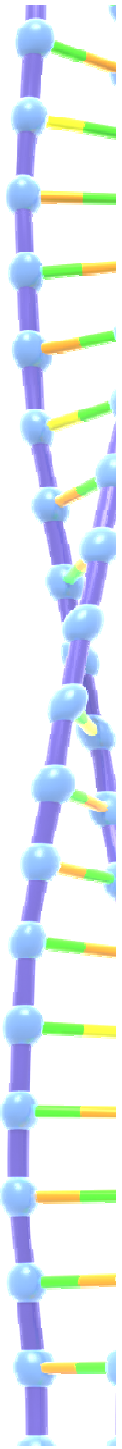
- Uitvoer in R:

```
            Df  Sum Sq Mean Sq F value  Pr(>F)
sample       2 0.30889 0.15444   19.86 0.00226 **
Residuals    6 0.04667 0.00778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1-WAY ANOVA IN R

- Datasets:
  - `y1 <- c(1.1, 1.3, 1.2)`
  - `y2 <- c(1.5, 1.6, 1.5)`
  - `y3 <- c(1.0, 1.2, 1.1)`
  - `y <- c(y1, y2, y3)`
  - `sample <- factor(c(rep(1,3), rep(2,3), rep(3, 3)))`
  - `myData <- data.frame(y, sample)`
- 1-way ANOVA analyse:
  - `fit <- aov(y ~ sample, data=myData)`
  - `summary(fit)`

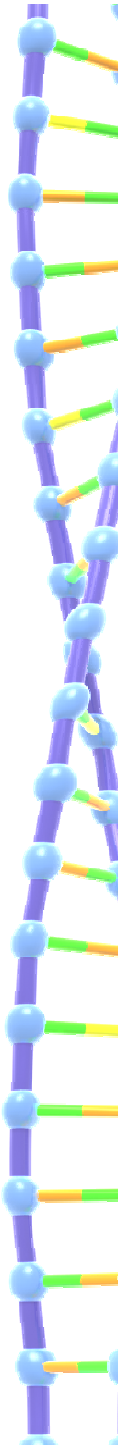


1-WAY ANOVA IN R

- Uitvoer:

```
> fit <- aov(y ~ sample, data=myData)
> summary(fit)
            Df  Sum Sq Mean Sq F value  Pr(>F)
sample       2 0.30889 0.15444   19.86 0.00226 **
Residuals    6 0.04667 0.00778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Zijn alle gemiddelden gelijk?

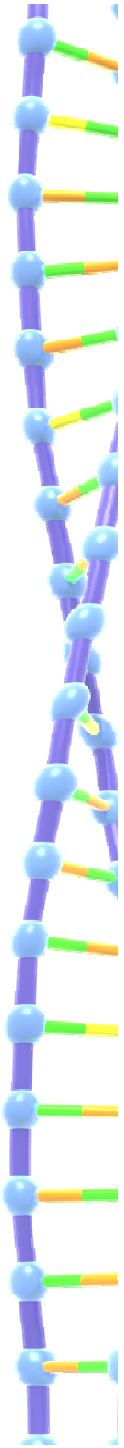


1-WAY ANOVA IN R

- Uitvoer:

```
> fit <- aov(y ~ sample, data=myData)
> summary(fit)
            Df  Sum Sq Mean Sq F value  Pr(>F)
sample       2 0.30889 0.15444   19.86 0.00226 **
Residuals    6 0.04667 0.00778
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

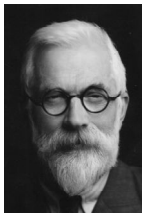
Nee, er is een significant verschil tussen de 3 samples, want $p < 0.05$



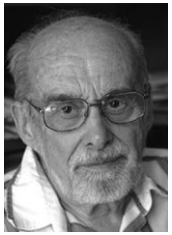
1-WAY ANOVA: *POST-HOC*

Als er een significant effect is van factor 1:
welk(e) level(s) verschillen dan van elkaar?

- Gelijke aantallen n_c per level i:
  - LSD, Tukey HSD, Scheffé, …
- Ongelijke aantallen per level i:
  - Bonferroni, Tukey HSD, …
- Vergelijken met referentielevel (bijv. placebo):
  - Dunnett





Allemaal *t*-toetsen, aangepast zodat de family-wise error rate < α

1-WAY ANOVA: *POST-HOC*: FISHER'S LSD

- Idee: voer tussen alle groepen (levels) onderling 2-sample *t*-toetsen uit; haal de gepoolde s_p uit de MS_{err} van de 1-way ANOVA:

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

$$t_{ij} = \frac{\overline{y_i} - \overline{y_j}}{\sqrt{MS_{\text{err}}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \rightarrow p_{ij}$$

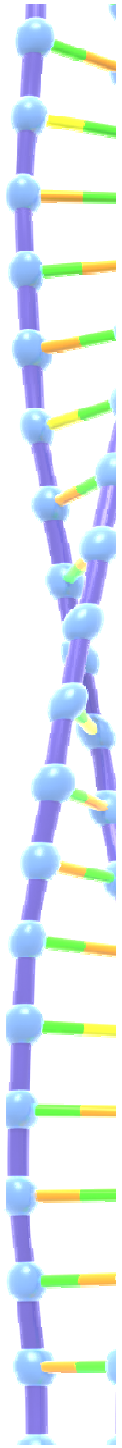
- Significant verschil tussen i en j als $p_{ij} < 0.05$

1-WAY ANOVA IN R: *POST-HOC*

- Fisher's LSD-toets voor significante ANOVA:
  - **`pairwise.t.test(y, sample, pool.sd=T, p.adjust.method="none")`**
- Uitvoer:

```
> pairwise.t.test(y, sample, pool.sd=T, p.adjust.method="none")

 Pairwise comparisons using t tests with pooled SD

data:  y and sample

  1       2
2 0.00358 -
3 0.21427 0.00095

P value adjustme    method: none
```

Welke samples verschillen significant van elkaar?



1-WAY ANOVA IN R: *POST-HOC*

- Fisher's LSD-toets voor significante ANOVA:
  - **`pairwise.t.test(y, sample, pool.sd=T, p.adjust.method="none")`**
- Uitvoer:

```
> pairwise.t.test(y, sample, pool.sd=T, p.adjust.method="none")

 Pairwise comparisons using t tests with pooled SD

data:  y and sample

  1       2
2 0.00358 -
3 0.21427 0.00095

P value adjustment method: none
```

p < 0.05

1 en 2 verschillen
2 en 3 verschillen
1 en 3 verschillen niet

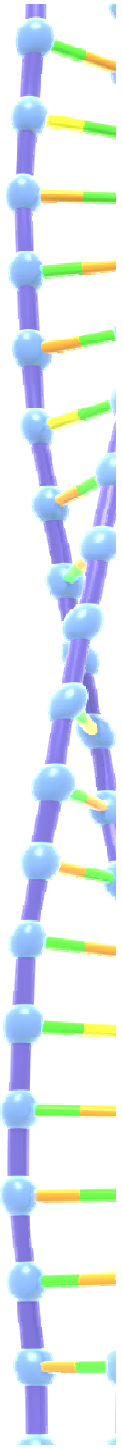
Er zijn 2 "subsets": samples (1,3) en (2)

1-WAY ANOVA IN R: *POST-HOC*

- Tukey-toets voor significante ANOVA:
  - **`TukeyHSD(fit)`**
- Uitvoer:

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ sample, data = myData)

$sample
           diff        lwr        upr     p adj
2-1   0.3333333  0.1123923  0.5542743 0.0085322
3-1  -0.1000000 -0.3209410  0.1209410 0.4037051
3-2  -0.4333333 -0.6542743 -0.2123923 0.0023012
```

Welke samples verschillen significant van elkaar?



1-WAY ANOVA IN R: *POST-HOC*

- Tukey-toets voor significante ANOVA:
  - **`TukeyHSD(fit)`**
- Uitvoer:

```
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ sample, data = myData)

$sample
           diff        lwr       upr     p adj
2-1   0.3333333  0.1123923 0.5542743 0.0085322
3-1  -0.1000000 -0.3209410 0.1209410 0.4037051
3-2  -0.4333333 -0.6542743 -0.2123923 0.0023012
```

$p < 0.05$

Er zijn 2 “subsets”: samples (1,3) en (2)

MODEL FORMULE IN R

- Statistisch 1-way ANOVA model:

$$y_{ik} = \mu + \alpha_i + \varepsilon_{ik} \qquad i = 1 \ldots q, \quad k = 1 \ldots n_c$$

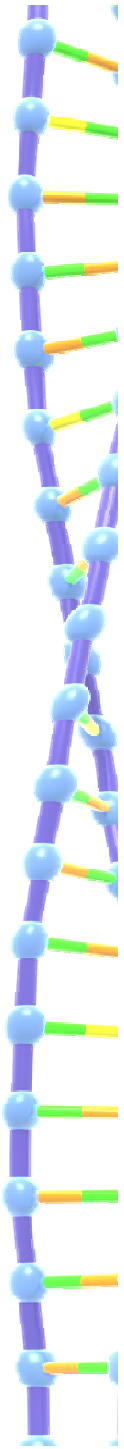
error

factor x

- R-model:
  - `y ~ x`
- Statistisch 2-sample *t*-toets model:

$$y_{ik} = \mu + \alpha_i + \varepsilon_{ik} \qquad i = 1 \ldots 2, \quad k = 1 \ldots n_c$$

- R-model:
  - `y ~ x`



MODEL FORMULE IN R

- Statistisch regressie model (1):

$$y = a_0 + a_1 \cdot x$$

- R model:
  - `y ~ x`
- Statistisch regressie model (2):

$$y = a_1 \cdot x$$

- R model:
- `y ~ x - 1`

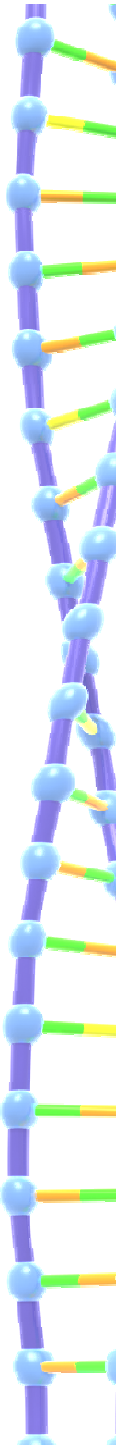
constante weg

- Statistisch regressiemodel (3):

$$y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_1^2$$

- R model:
  - `y ~ x1 + x2 + I(x1^2)`

I( ) : “letterlijk”



MODEL FORMULE IN R

- Statistisch 1-way ANOVA model:

$$y_{ik} = \mu + \alpha_i + \varepsilon_{ik} \qquad i = 1 \ldots q, \quad k = 1 \ldots n_c$$

- R model:
  - `y ~ x`
- Statistisch 2-way ANOVA model (zonder interactie):

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \qquad i = 1 \ldots q, \quad j = 1 \ldots r, \quad k = 1 \ldots n_c$$

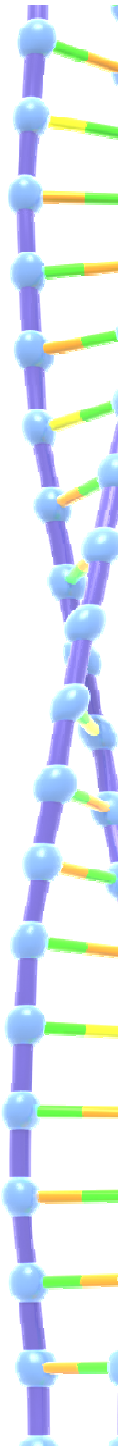
- R model:
  - `y ~ x1 + x2`
- Statistisch 2-way ANOVA model (met interactie):

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \qquad i = 1 \ldots q, \quad j = 1 \ldots r, \quad k = 1 \ldots n_c$$

- R model:
  - `y ~ x1 + x2 + x1:x2`
  - `y ~ x1*x2`

FACTOR VS NUMERIEKE VECTOR

- (Tijdelijk) switchen van numerieke vector naar factor:
  - `as.factor( )`
- (Tijdelijk) switchen van factor naar numerieke vector:
  - `as.numeric( )`
- Voorbeeld:
  - `sample <-factor(c(rep(1,5),rep(2,5)))`
  - `y[sample < 2]  # geeft error`
  - `y[as.numeric(sample) < 2]`



Jullie kunnen nu de (rest v/d) opdrachten van les 6 maken

Hanze University Groningen
Applied Sciences
Institute for Life Science & Technology