UNIVERSITEIT GENT

BIG N2N
Bioinformatics Institute Ghent
from nucleotides to networks

Niet-parametrische Statistiek

Lieven Clement

2^{de} bach. in de Biologie, Chemie, Biochemie en Biotechnologie en Biomedische Wetenschappen

Inleiding

Vorige hoofdstukken: parametrische methoden

- Inferentie enkel correct als voldaan aan param. veronderstellingen:
  - distributionele veronderstellingen: v.b. Normaal verdeelde gegevens.
  - gelijkheid van varianties (two-sample t-test en ANOVA)
- De p-waarde: $\mathrm{P}_0 [|T| \geq |t|]$.
  - Berekend o.b.v. nuldistributie van T die afgeleid is van verdeling van observaties
  - Fout als niet voldaan is aan veronderstellingen
- 95% BI steunt eveneens op veronderstellingen. Niet voldaan: geen garantie dat intervallen populatie parameterwaarde omvatten met 95% kans.

- Asymptotische theorie moeilijker te plaatsen: je kan stellen dat t-test asymptotisch niet-parametrisch is omdat bij erg grote steekproefgroottes de distributionele veronderstelling van normaliteit niet meer belangrijk is.
- Parametrische aanpak:
  - efficiënter: grotere power bij zelfde steekproefgrootte + smallere BI
  - meer flexibel: makkelijker inzetbaar voor experimenten met meer complexe designs

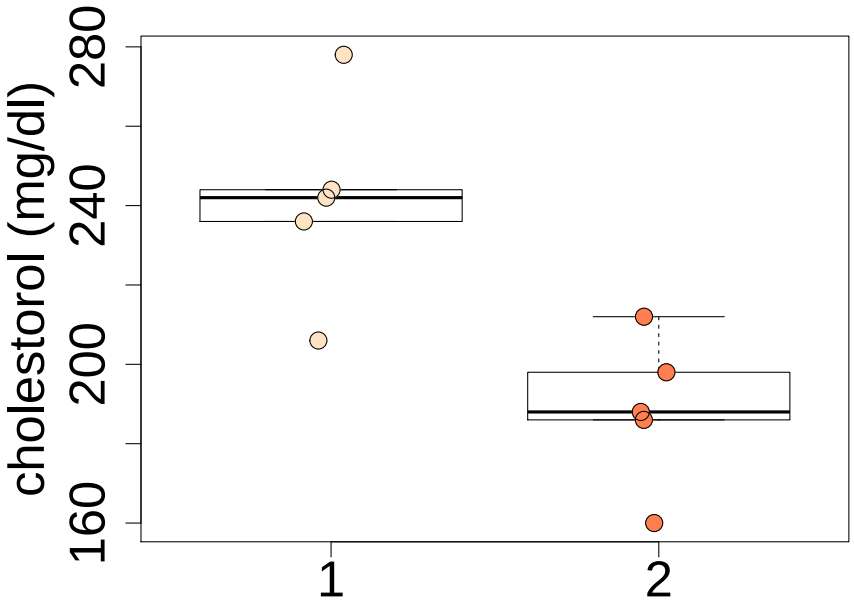
Vergelijken van twee groepen

Cholestorol voorbeeld

- Cholestorolconcentratie in bloed gemeten bij
- 5 patiënten (groep=1) die twee dagen geleden een hartaanval deden
- bij 5 gezonde personen (groep=2).
- Is cholestorolconcentratie verschillend bij hartpatiënten en gezonde personen?

```
chol <- read.table("dataset/chol.txt",header=TRUE)
chol$group <- as.factor(chol$group)
nGroups=table(chol$group)
n=sum(nGroups)
head(chol)
```

```
##   group cholest
## 1     1     244
## 2     1     206
## 3     1     242
```



- Mogelijks outliers
- Moeilijk om inzicht te krijgen in verdeling: maar 5 observaties per groep

Permutatie-testen

Vraagstelling Cholestorol voorbeeld vertaling naar nulhypothese. Klassiek:

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2.$$

- Groep 1: Hartpatiënten
- Groep 2: Gezonde individuen
- Testen van deze hypotheses d.m.v. two-sample t-test.

We gaan de voorwaarden na:

- Normaliteit?
- Gelijkheid van variantie?
- Te weinig observaties per groep.
- We kunnen veronderstellingen niet nagaan!
- Gevaarlijk!
- Oplossing: permutatietesten

Hypothesen

- Y_{1j} en Y_{2j} uitkomsten uit respectievelijk groep 1 en 2:

$$Y_{1j} \text{ iid } N(\mu_1, \sigma^2) \quad \text{en} \quad Y_{2j} \text{ iid } N(\mu_2, \sigma^2).$$

- Onder $H_0 : \mu_1 = \mu_2$, wordt dit (stel $\mu = \mu_1 = \mu_2$ onder H_0)

$$Y_{ij} \text{ iid } N(\mu, \sigma^2),$$

- Drukt uit dat alle $n = n_1 + n_2$ uitkomsten uit zelfde normale distributie komen en onafhankelijk zijn
- Laat toe om oorspronkelijke nulhypothese anders te schrijven:

$$H_0 : f_1(y) = f_2(y) \text{ voor alle } y$$

met

- f_1 en f_2 de verdeling van uitkomsten
- Bijkomende veronderstelling dat f_1 en f_2 normale verdelingen zijn.

Onder de alternatieve hypothese wordt een locatie-shift verondersteld:

$$H_1 : f_1(y) = f_2(y - \Delta) \quad \text{voor alle } y$$

met

- $\Delta = \mu_1 - \mu_2$
- en f_1 en f_2 normale verdelingen met dezelfde variantie.

We illustreren dit in R voor $f_1 \sim N(0, 1)$ en $f_2 \sim N(1, 1)$ en $\Delta = -1$.

```
mu1 <- 0; mu2 <- 1; sigma1 <- sigma2 <- 1
y <- -2:2
delta <- mu1-mu2; delta
```

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

```
rbind(dnorm(y,mu1,sigma1), dnorm(y-delta,mu2,sigma2))
```

```
##           [,1]      [,2]      [,3]      [,4]       [,5]
## [1,] 0.05399097 0.2419707 0.3989423 0.2419707 0.05399097
## [2,] 0.05399097 0.2419707 0.3989423 0.2419707 0.05399097
```

De permutatietesten die we ontwikkelen kunnen gebruikt worden voor het testen van

$$H_0 : F_1 = F_2 \quad \text{of} \quad H_0 : f_1 = f_2.$$

maar zonder de Normaliteitsveronderstellingen.

We weten dat onder H_0 geldt dat :

- Verdeling van de cholestorolconcentraties gelijk voor hartpatiënten en gezonde personen
- Groep-labels van de 10 personen niet informatief
- Groepering gebruikt om originele teststatistiek te bepalen is onder de nulhypothese een van de vele groeperingen die allemaal even zinvol/even weinig zinvol zijn.
- Elke groepering zou immers dezelfde uitkomsten hebben gegenereerd aangezien er geen effect is van de behandeling.
- Bereken nulldistributie door permuteren van de groepslabels!

Verdeling van de statistiek onder H_0

- $m = \binom{n_1+n_2}{n_1} = \binom{n}{n_1} = \binom{n}{n_2}$ mogelijke unique permutaties $\mathcal{G}$ van de groepslabels.
- In ons voorbeeld $m = 252$.
- Als m niet te groot is kunnen alle unieke permutaties van de groepslabels berekend worden.
- Vervolgens wordt voor iedere unique permutatie $g \in \mathcal{G}$ de teststatistiek t_g^* berekend
- Hier de t-test statistiek door de originele uitkomsten te gebruiken die nu gekoppeld worden aan de gepermuteerde groepslabels G_g^*.

We kunnen alle m=252 permutaties in R genereren a.d.h.v. de functie `combn(n,n_1)`. Dit wordt geïllustreerd in de onderstaande R code:

- G bevat volgnummers van de observaties uit groep 1 voor elke permutatie d
- We tonen enkel de eerste 10 permutaties.

```
G=combn(n,nGroups[1])
dim(G)
```

```
## [1]   5 252
```

```
G[,1:10]
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    1    1    1    1    1    1    1    1    1     1
## [2,]    2    2    2    2    2    2    2    2    2     2
## [3,]    3    3    3    3    3    3    3    3    3     3
## [4,]    4    4    4    4    4    4    5    5    5     5
## [5,]    5    6    7    8    9   10    6    7    8     9
```

We berekenen nu de teststatistiek voor elke permutatie

```
tOrig=t.test(cholest~group,chol)$statistic
tOrig
```

```
##         t
## 3.664425
```

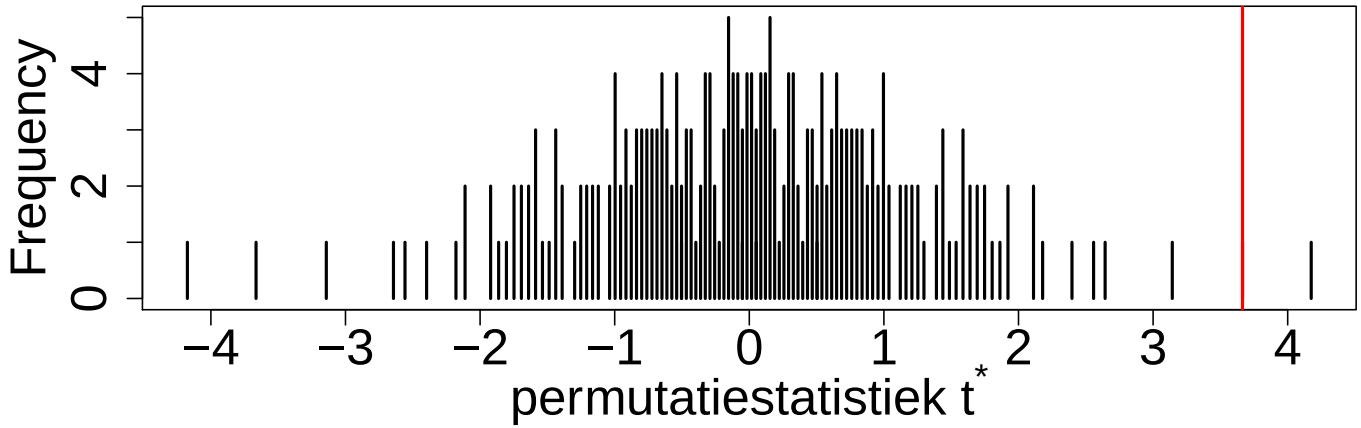
```
tStar=combn(n,nGroups[1],
    function(g,y=chol$cholest) t.test(y[g],y[-g])$statistic)
head(tStar)
```

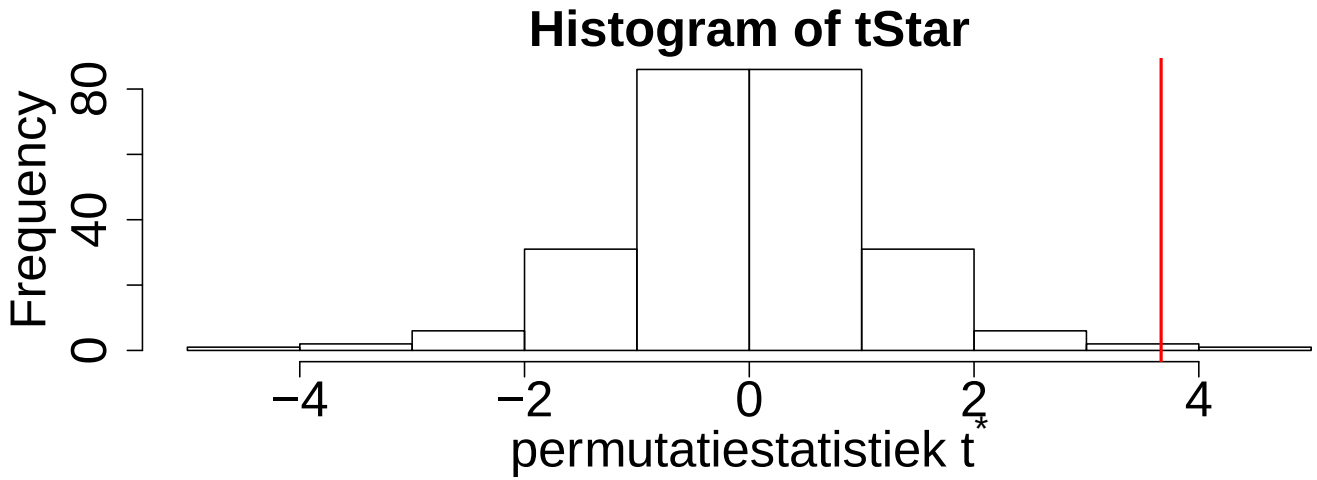
```
## [1] 3.6644253 1.6397911 2.3973923 1.5876250 1.9217173 0.99671
```

```
length(tStar)
```

```
## [1] 252
```

- We kunnen nu de verdeling van de teststatistiek onder H_0 bestuderen.
- Originele statistiek (rode verticale lijn) is extreem





p-waarde

- Nu we permutatienuldistributie hebben kunnen we hypothesetesten uitvoeren.
- tweezijdige permutatie p-waarde

$$p = \mathrm{P}_0 \left[|T| \geq |t| \mid \mathbf{y} \right].$$

- p-waarde geconditioneerd op geobserveerde cholestorolwaarden $\mathbf{y} = (y_{11}, \ldots, y_{51}, y_{12}, \ldots, y_{52})^T$.
- Gezien permutatienuldistrubutie van T bepaald wordt door t_g^*, $g \in \mathcal{G}$, berekenen we

$$p = \frac{\#\{g \in \mathcal{G} : |t_g^*| \geq |t|\}}{m}$$

```
pval=mean(abs(tStar)>=abs(tOrig))
pval
```

```
## [1] 0.01587302
```

- Op het 5% significantieniveau besluiten we dat de distributies van de cholestorol concentraties niet gelijk zijn bij hartpatiënten en bij gezonde personen. ($p = 0.0159$).
- De p-waarde op basis van alle permutaties wordt een **exacte** p-waarde genoemd.
- De permutatienuldistributie wordt een **exacte** nuldistributie genoemd.
- De term **exact** betekent dat de resultaten correct zijn voor iedere steekproefgrootte n.

Kritieke waarde

- Ook de kritieke waarde c kan eenvoudig bekomen worden.

$$\mathrm{P}_0\left[|T| > c \mid \mathbf{y}\right] = \alpha.$$

- Door discrete natuur van de permutatienuldistributie onwaarschijnlijk om een kritieke waarde c te vinden zodat deze gelijkheid exact opgaat.
- Daarom zoeken we de kleinste waarde c zodat

$$\mathrm{P}_0\left[|T| > c \mid \mathbf{y}\right] \leq \alpha.$$

- Permutatietest daarom mogelijks te conservatief
- Meestal is m voldoende groot zodat de kans op een type I fout ($\mathrm{P}_0\left[|T| > c \mid \mathbf{y}\right]$) erg dicht bij het nominale significantieniveau ligt.

```r
alpha<-0.05
m <- length(tStar)
t.crit<-sort(abs(tStar))[ceiling((1-alpha)*m)]
t.crit
```

```
## [1] 2.179236
```

```r
mean(abs(tStar)>t.crit)
```

```
## [1] 0.04761905
```

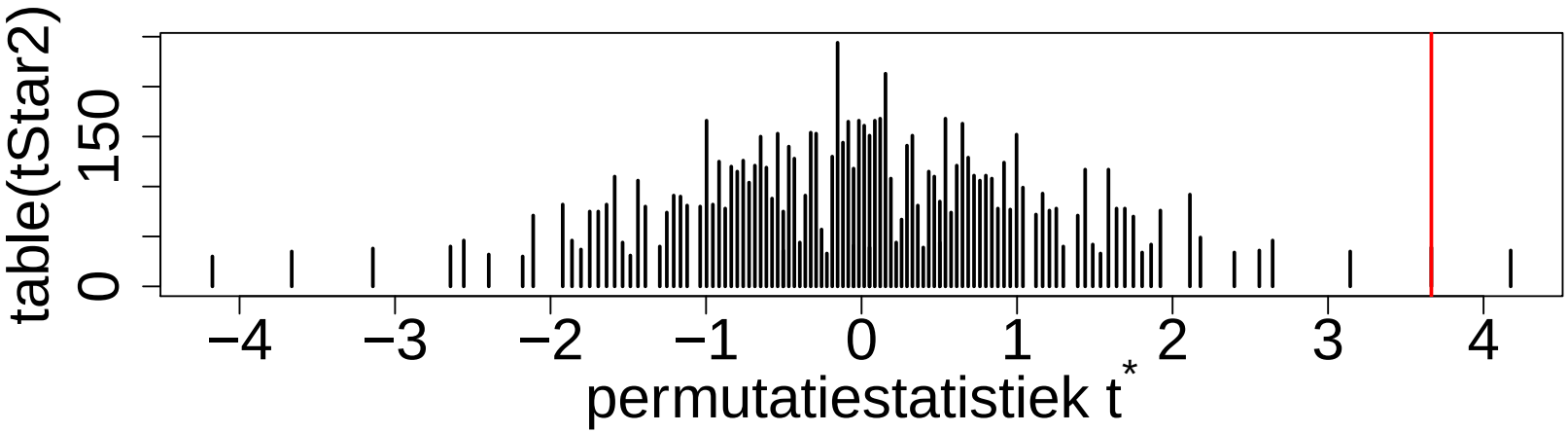
- De kans op een type I fout wordt dus gecontroleerd door een permutatietest, maar wel conditioneel op de geobserveerde uitkomsten data **y**.
- **We kunnen ons nu afvragen of we de conclusies kunnen veralgemenen naar de populatie toe? Het antwoord is ja, als de subjecten at random getrokken zijn uit de populatie.**
- Het bewijs hiervan valt buiten het bestek van de cursus.

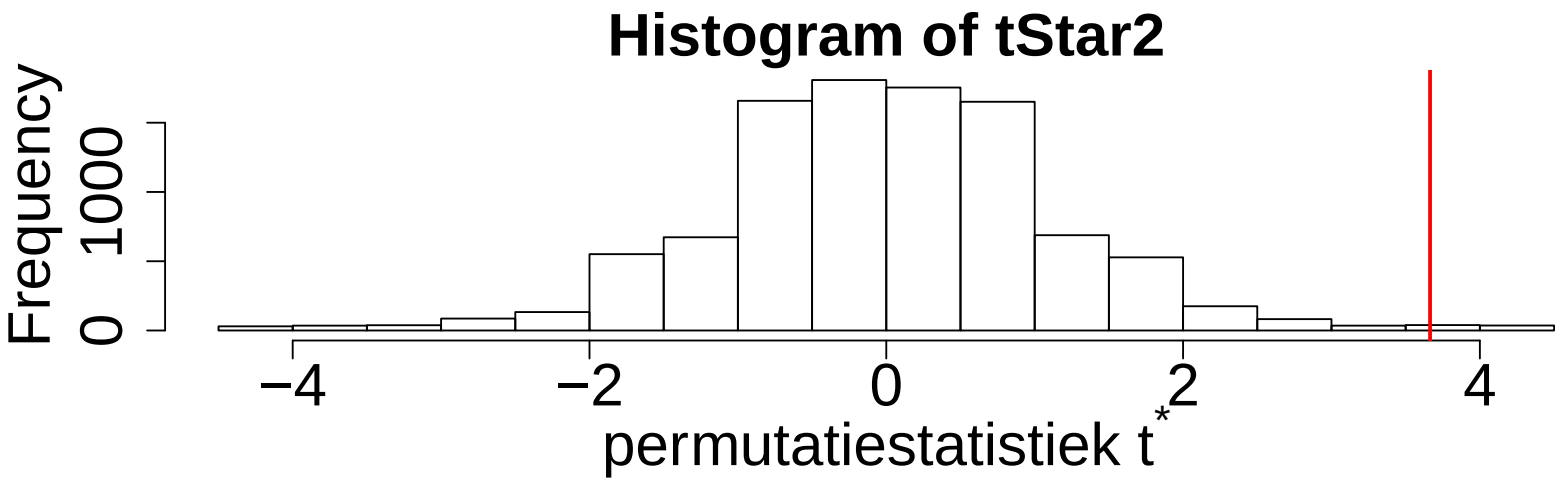
- Soms probleem omdat het aantal permutaties $m = \#\mathcal{G}$ erg groot is
- $\binom{20}{10} = 184756$
- $\binom{30}{15} = 1.55\text{e+}08$
- $\binom{40}{20} = 1.38\text{e+}11$.
- Daarom niet alle $g \in \mathcal{G}$ te beschouwen, maar groot aantal random permutaties uitvoeren (b.v. 10000)
- Als men niet alle permutaties uitvoert, kan het dat de geobserveerde statistiek niet berekend wordt in de random geselecteerde permutaties
- Als de statistiek erg extreem is, kan het dat de de statistiek groter is dan alle statistieken die in de permutaties werden berekend. $\rightarrow p = 0$
- Permutatie p-waarde $p = 0$ theoretisch niet mogelijk! Elke mogelijke t-waarde minimum 1 keer behaald.
- Andere manier om p-waarde te berekenen o.b.v. B willekeurige permutaties:

$$p = \frac{\#\{|t_g^*| \geq |t|\} + 1}{B + 1},$$

```r
set.seed(304)
B=10000
tStar2=sapply(X=1:B, FUN=function(b,y,groep)
    {t.test(y~sample(groep))$statistic}
          ,y=chol$cholest,groep=chol$group)
pval2=(sum(abs(tStar2)>=mean(tOrig))+1)/(B+1)
pval2
```

```
## [1] 0.01409859
```





Approximatieve $p = 0.0141$ niet ver van de exacte $p = 0.0159$

Rank Testen

- Belangrijkste groep van niet-parametrische testen
- Populariteit:
- Niet-parametrisch,
- Exacte p-waarden d.m.v.permutatienuldistributie.
- Geen nood aan aparte permutatienuldistributie voor iedere nieuwe dataset.
- Permutatienuldistributie van rank testen hangt alleen af van steekproefgroottes.
- Erg robust zijn tegen uitschieters (Engels: *outliers*)
- Nuttig als locatie-shift model niet opgaat.

Ranks

Rank testen starten vanuit rank-getransformeerde uitkomsten.

- Beschouw $Y_1, \ldots, Y_n$.
- Afwezigheid van twee gelijke observaties (i.e. geen *ties*).

$$R_i = R(Y_i) = \#\{Y_j : Y_j \leq Y_i; j = 1, \ldots, n\}$$

- Kleinste observatie krijgt dus rank 1, de tweede kleinste rank 2, enzovoort, en de grootste observatie, tenslotte, krijgt rank n.

```
chol$cholest
```

```
##  [1] 244 206 242 278 236 188 212 186 198 160
```

```
rank(chol$cholest)
```

```
##  [1]  9  5  8 10  7  3  6  2  4  1
```

Soms komen *ties* voor in de data, i.e. minstens twee observaties hebben dezelfde numerieke waarde. Een klein voorbeeld:

```
metTies=c(403,507,507,610,651,651,651,830,900)
rank(metTies)
```

```
## [1] 1.0 2.5 2.5 4.0 6.0 6.0 6.0 8.0 9.0
```

- Ties: 507 komt tweemaal voor, 651 komt driemaal voor. Dit zijn voorbeelden van *ties*.
- Wanneer *ties* voorkomen, worden *midranks* gebruikt.
- **midrank** van observatie Y_i wordt

$$R_i = \frac{\#\{Y_j : Y_j \leq Y_i\} + (\#\{Y_j : Y_j < Y_i\} + 1)}{2}.$$

- Dikwijls de ranks van de uitkomsten nodig in de gepoolde steekproef.
- Bijvoorbeeld: beschouw de uitkomsten Y_{ij}, $i = 1, \ldots, n_j$ en $j = 1, 2$.
- Deze uitkomsten kunnen ook worden voorgesteld door $Z_1, \ldots, Z_n$ ($n = n_1 + n_2$), de uitkomsten uit de gepoolde steekproef.

```
t(chol)
```

```
##         [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
## group   "1"   "1"   "1"   "1"   "1"   "2"   "2"   "2"   "2"
## cholest "244" "206" "242" "278" "236" "188" "212" "186" "198"
```

```
z=chol$cholest
z
```

```
##  [1] 244 206 242 278 236 188 212 186 198 160
```

```
rank(z)
```

```
##  [1]  9  5  8 10  7  3  6  2  4  1
```

Wilcoxon-Mann-Whitney Test

- Gelijktijdig ontwikkeld door Wilcoxon en door Mann en Whitney
- **Wilcoxon-Mann-Whitney**, **Wilcoxon rank sum test** of **Mann-Whitney U test**
- $H_0 : f_1 = f_2$ vs $H_1 : \mu_1 \neq \mu_2$ (of de eenzijdige versies).
- Eerst **locatie-shift** model veronderstellen later relaxeren we aanname.
- Klassieke t-test: verschil in steekproefgemiddelden $\bar{Y}_1 - \bar{Y}_2$.
- Hier: verschil in steekgroepgemiddelde op basis van rank-getransformeerde uitkomsten.
- Ranks op basis van gepoolde sample (na samenvoegen van uitkomsten uit groep 1 en groep 2)

- $R_{ij} = R(Y_{ij})$ is de rank van uitkomst Y_{ij} in de gepoolde steekproef.

$$T = \frac{1}{n_1}\sum_{i=1}^{n_1} R(Y_{i1}) - \frac{1}{n_2}\sum_{i=1}^{n_2} R(Y_{i2}).$$

- Onder H_0 verwachten we dat gemiddelde rank in de eerste groep ongeveer gelijk is aan de gemiddelde rank in de tweede groep en T dicht bij nul.
- Als H_1 waar is dan verwachten we dat gemiddelde ranks verschillen en dus dat T niet dicht bij nul zal liggen.
- Eigenlijk voldoende om enkel

$$S_1 = \sum_{i=1}^{n_1} R(Y_{i1})$$

.

- S_1 is som van de ranks van observaties uit groep 1 vandaar de naam *rank sum test*.
- Want

$$S_1 + S_2 = \text{som van alle ranks} = 1 + 2 + \cdots + n = \frac{1}{2}n(n+1).$$

- S_1 (of S_2) een goede teststatistiek
- Permutatietestmethode toegepassen om exacte permutatienuldistributie te bekomen
- Voor een gegeven steekproefgrootte n, en veronderstellend dat er geen *ties* zijn, zijn rank-getransformeerde uitkomsten altijd

$$1, 2, \ldots, n$$

- Voor gegeven groepsgroottes n_1 en n_2, zal de permutatienuldistributie dan ook steeds dezelfde zijn!
- Tot 1980 werd dit als een groot voordeel beschouwd omdat de nuldistributies voor gegeven n_1 en n_2 getabuleerd konden worden
- Door rekencapaciteit speelt dit argument niet meer, wel andere belangrijke redenen.

- Vaak wordt gestandaardiseerde teststatistiek gebruikt

$$T = \frac{S_1 - \mathrm{E}_0[S_1]}{\sqrt{\mathrm{Var}_0[S_1]}},$$

- met $\mathrm{E}_0[S_1]$ en $\mathrm{Var}_0[S_1]$ de verwachtingswaarde en variantie van S_1 onder H_0.
- Dit zijn dus het gemiddelde en variantie van de permutatienuldistributie van S_1.
- Onder H_0 geldt

$$\mathrm{E}_0[S_1] = \frac{1}{2} n_1 (n+1) \quad \text{en} \quad \mathrm{Var}_0[S_1] = \frac{1}{12} n_1 n_2 (n+1).$$

- Verder kan men onder H_0 en als $\min(n_1, n_2) \to \infty$ opgaat aantonen dat,

$$T = \frac{S_1 - \mathrm{E}_0[S_1]}{\sqrt{\mathrm{Var}_0[S_1]}} \to N(0,1).$$

- Asymptotisch volgt gestandaardiseerde teststatistiek een standaardnormaal verdeling!

We illustreren de WMW test aan de hand van de R functie `wilcox.test`.

```
wilcox.test(cholest~group,data=chol)
```

```
##
##  Wilcoxon rank sum test
##
## data:  cholest by group
## W = 24, p-value = 0.01587
## alternative hypothesis: true location shift is not equal to 0
```

- We verwerpen H_0 ($p = 0.016 < 0.05$)
- De output geeft de teststatistiek $W = 24$?
- In volgende lijnen berekenen we S_1 en S_2 manueel voor de dataset.

```
S1=sum(rank(chol$cholest)[chol$group==1])
S2=sum(rank(chol$cholest)[chol$group==2]); c(S1,S2)
```

```
## [1] 39 16
```

Waar komt $W = 24$ vandaan?

- Mann en Whitney test in afwezigheid van ties:

$$U_1 = \sum_{i=1}^{n_1} \sum_{k=1}^{n_2} \mathrm{I}\{Y_{i1} \geq Y_{k2}\}.$$

- waarbij $\mathrm{I}\{.\}$ een indicator is die 1 is als de uitdrukking waar is en 0 als dit niet het geval is.
- U telt dus hoeveel keer een observatie uit de eerste groep groter of gelijk is aan een observatie uit de tweede groep.

```
y1=subset(chol,group==1)$cholest
y2=subset(chol,group==2)$cholest
u1Hlp=sapply(y1,function(y1i,y2) {y1i>=y2},y2=y2)
colnames(u1Hlp)=y1;rownames(u1Hlp)=y2
```

```
u1Hlp
```

```
##       244   206  242  278  236
## 188 TRUE  TRUE TRUE TRUE TRUE
## 212 TRUE FALSE TRUE TRUE TRUE
## 186 TRUE  TRUE TRUE TRUE TRUE
## 198 TRUE  TRUE TRUE TRUE TRUE
## 160 TRUE  TRUE TRUE TRUE TRUE
```

```
U1=sum(u1Hlp); U1
```

```
## [1] 24
```

Er kan worden aangetoond dat $U_1 = S_1 - \frac{1}{2} n_1(n_1 + 1)$.

```
S1-nGroups[1]*(nGroups[1]+1)/2
```

```
##  1
## 24
```

1. U_1 en S_1 dezelfde informatie bevatten,
2. U_1 is ook een rankstatistiek is en
3. Exacte testen gebaseerd op U_1 en S_1 equivalent zijn.

- U_1 heeft interpretatievoordeel
- Stel Y_j een willekeurige uitkomst uit behandelingsgroep j $(j = 1, 2)$. Dan geldt

$$\frac{1}{n_1 n_2} \mathrm{E}\left[U_1\right] \quad = \quad \mathrm{P}\left[Y_1 \geq Y_2\right].$$

- Intuïtief voelen we dit aan:
- Op basis van de steekproef kunnen we die kans schatten door het gemiddelde te berekenen van alle indicator waarden $\mathrm{I}\left\{Y_{i1} \geq Y_{k2}\right\}$.
- We voerden inderdaad $n_1 \times n_2$ vergelijkingen uit.

```r
mean(u1Hlp)
```

```
## [1] 0.96
```

```r
U1/(nGroups[1]*nGroups[2])
```

```
##    1
## 0.96
```

- De kans $\mathrm{P}[Y_1 \geq Y_2]$ wordt een **probabilistische index** (Engels: *probabilistic index*) genoemd.
- Het is de kans dat een uitkomst uit de eerste groep groter is dan een uitkomst uit de tweede groep.
- Als H_0 waar is, dan is $\mathrm{P}[Y_1 \geq Y_2] = \frac{1}{2}$.

- De R functie `wilcox.test` geeft niet de Wilcoxon rank sum statistiek, maar wel de Mann-Whitney statistiek U_1.
- We bekijken nogmaals de output

```
wTest=wilcox.test(cholest~group,data=chol); wTest ; U1
```

```
##
##  Wilcoxon rank sum test
##
## data:  cholest by group
## W = 24, p-value = 0.01587
## alternative hypothesis: true location shift is not equal to 0
```

```
## [1] 24
```

```
probInd=wTest$statistic/prod(nGroups); probInd
```

```
##    W
## 0.96
```

- Aangezien $p = 0.0159 < 0.05$ besluiten we op het 5% significantieniveau dat de gemiddelde cholestorolconcentratie groter is bij hartpatiënten kort na een hartaanval dan bij gezonde personen. (We nemen aan dat locatie-shift opgaat)
- We weten ook dat een cholestorolwaarde van hartpatiënten met een kans van $U1/(n_1 \times n_2) = 96\%$ groter is die van gezonde personen.
- We zouden de veronderstelling van de locatie-shift moeten nagaan, maar met slechts 5 observaties in elke behandelingsgroep is dit zinloos.

- Zonder locatie-shift veronderstelling blijft de conclusie in termen van de probabilistische index correct!
- Dus wanneer we geen locatie-shift veronderstellen en een tweezijdige test uitvoeren testen we eigenlijk

$$H_0 : F_1 = F_2 \text{ vs } \mathrm{P}(Y_1 \geq Y_2) \neq 0.5.$$

Conclusie: Er is een significant verschil in de distributie van de cholestorolconcentraties bij hartpatiënten 2 dagen na hun hartaanval en gezonde individuen ($p = 0.0159$). Het is meer waarschijnlijk om een hogere cholestorolconcentraties te observeren hartpatiënten dan bij gezonde individuen. De puntschatting voor deze kans bedraagt 96%.

Vergelijken van g Behandelingen

- Veralgemenen naar niet-parametrische tegenhangers van de F-test uit een one-way ANOVA.

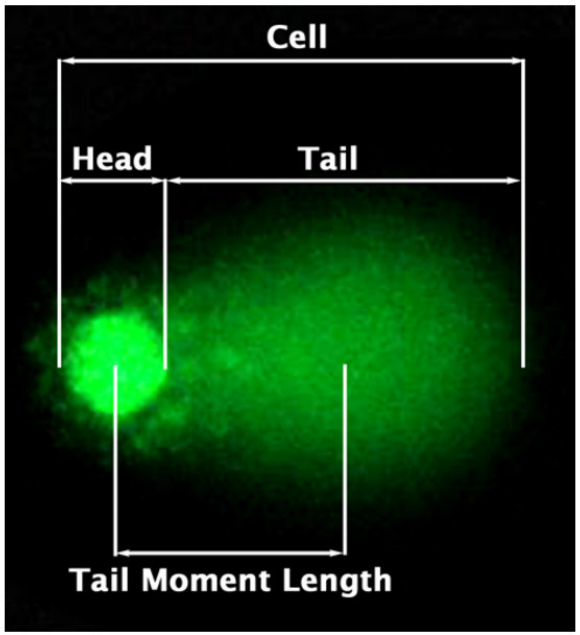
DMH Voorbeeld

1,2-dimethylhydrazine dihydrochloride (DMH) testen op genotoxiciteit (EU directive) - 24 ratten - Vier groepen volgens dagelijkse DMH dosis - controle - laag - medium - hoog

- Genotoxiciteit in de lever a.d.h.v. een comet assay op 150 levercellen per rat.
- De onderzoekers wensen na te gaan of verschillen zijn in de DNA schade tengevolge van de DMH dosis.

Comet Assay:

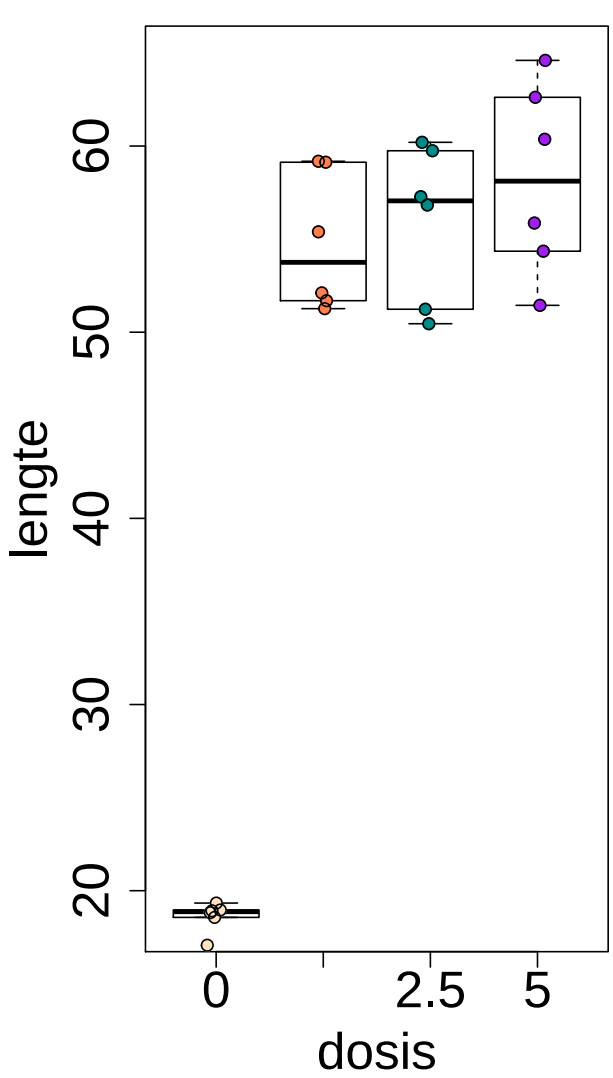
- DNA strengbreuken visualiseren
- Lengte comet staart is proxy voor strengbreuken.

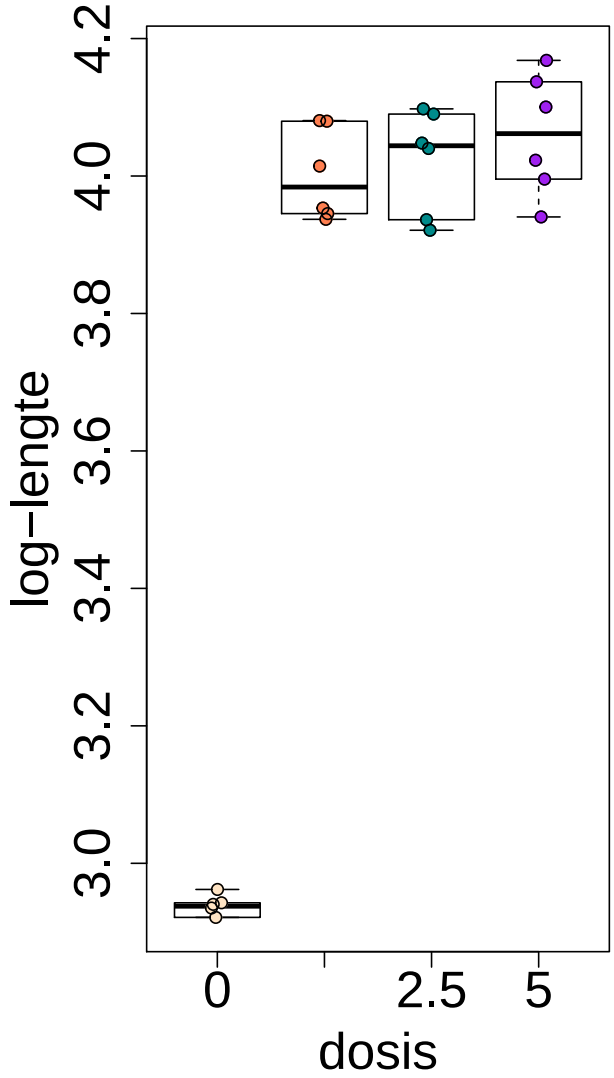


Figuur 1: Comet assay

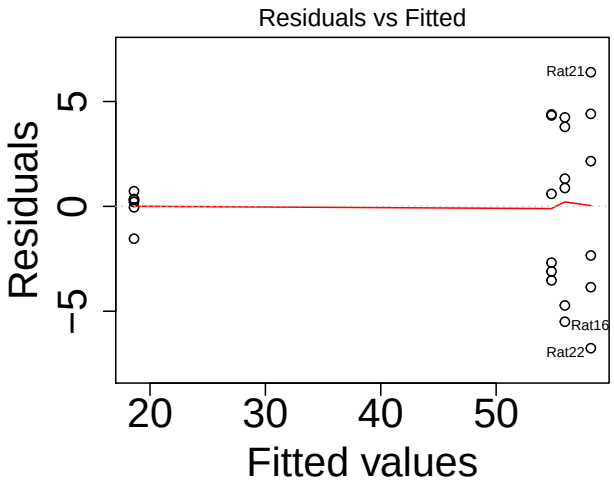
```r
dna <- read.table("dataset/dna.txt",header=TRUE)
dna$dose <- as.factor(dna$dose)
head(dna)
```

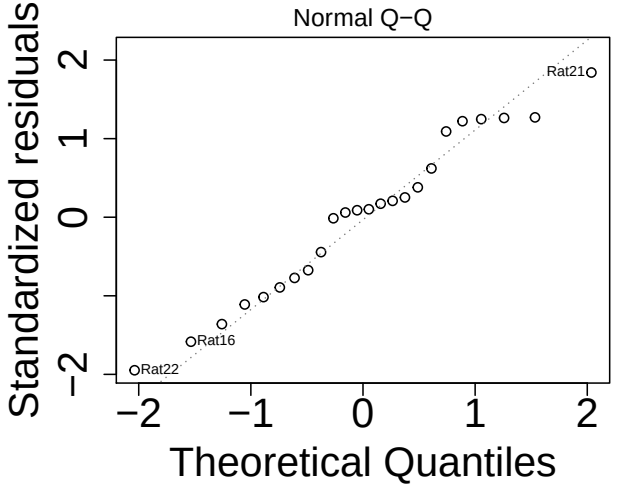
```
##        length dose
## Rat1 19.33632    0
## Rat2 18.92102    0
## Rat3 18.56595    0
## Rat4 18.96406    0
## Rat5 17.07120    0
## Rat6 18.82054    0
```

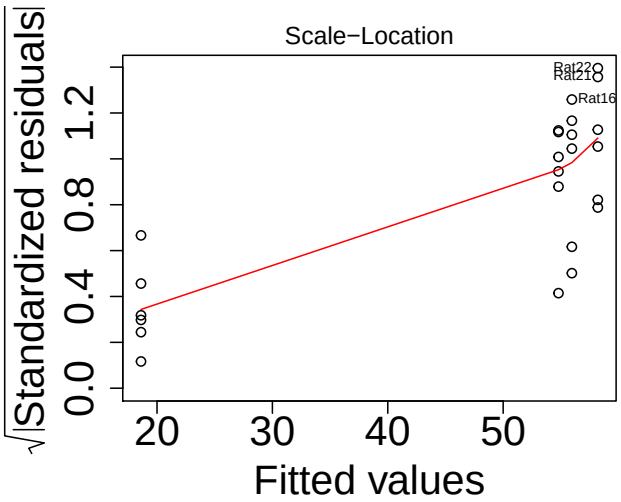


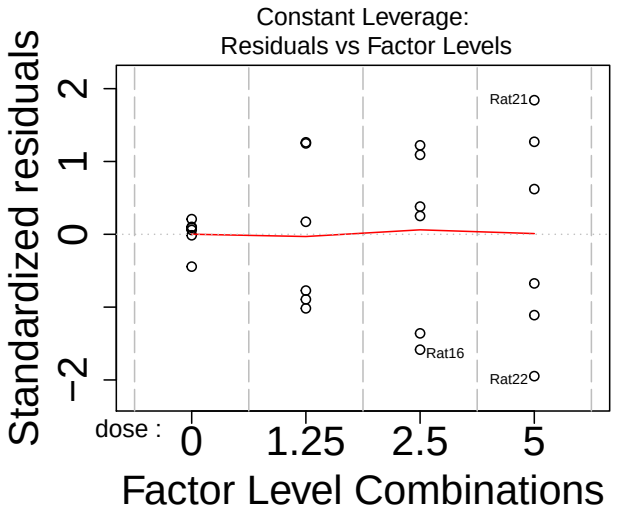


- Indicatie dat de controle groep andere variabiliteit heeft.
- 6 observaties per groep te weinig om aannames na te gaan.

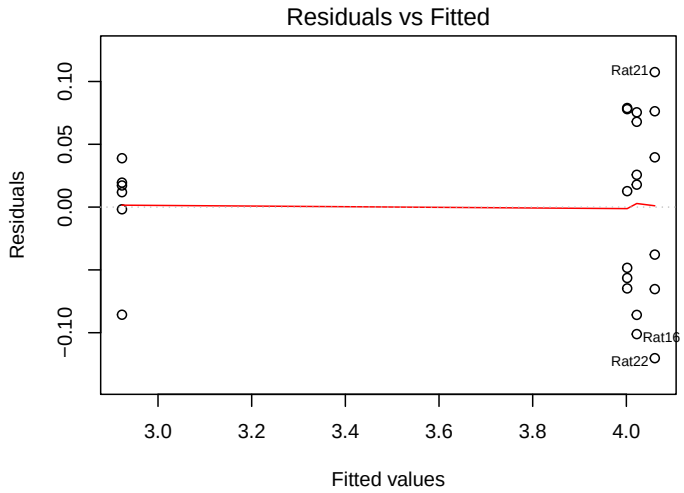


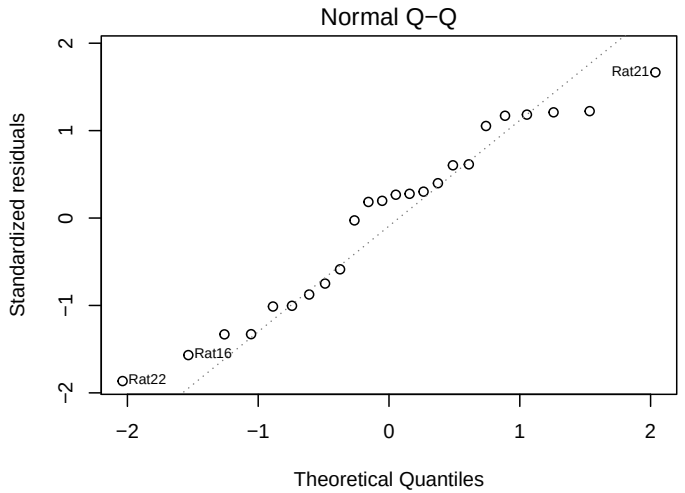


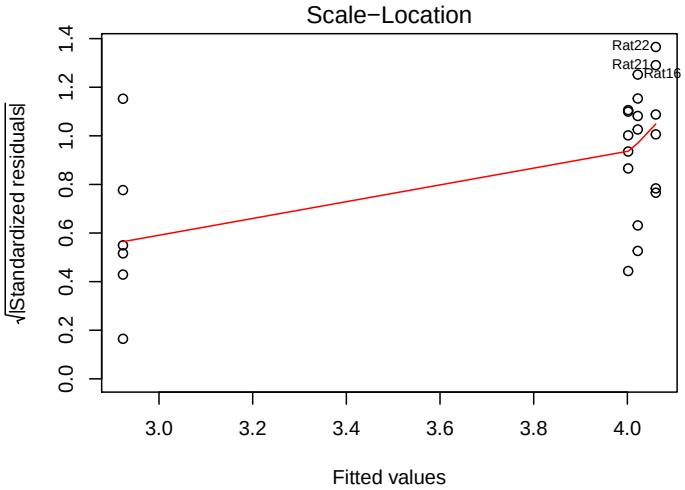


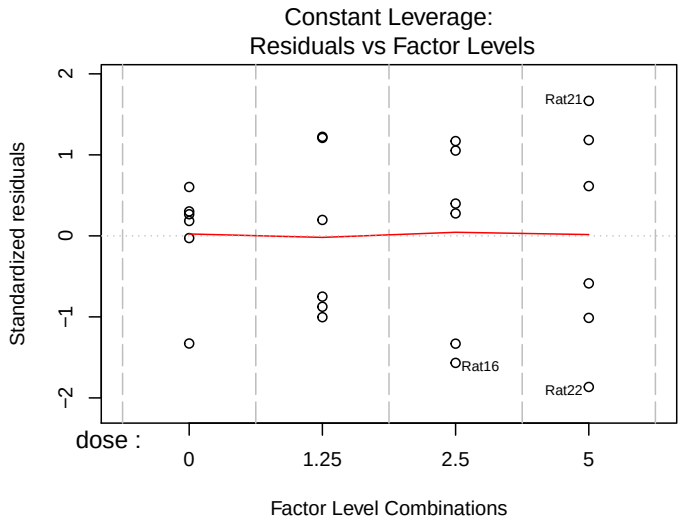



```
## null device
##           1
```










```
## null device
##           1
```

Permutatietest

- One-way ANOVA model impliceert locatie-shift.
- Onder veronderstellingen van one-way ANOVA model volgt meer algemene nulhypothese

$$H_0 : f_1(y) = f_2(y) = \ldots = f_t(y) \text{ voor alle } y.$$

- We veronderstellen niet langer normale distributies.
- Als we locatie-shift model kunnen veronderstellen is de alternatieve hypothese analoog als bij de ANOVA test nl.

$$H_1 : \exists\, j, k \in \{1, \ldots, g\} : \mu_j \neq \mu_k.$$

- We kunnen opnieuw groepslabels permuteren om de nuldistributie van de test-statistiek te bekomen.
- Men kan aantonen dat er

$$m = \frac{n!}{n_1! \ldots n_g!}$$

  unieke permutaties $\mathcal{G}$ bestaan.
- Voor ons voorbeeld zijn dat er $m = (24!)/(6!)^4 = 2.31\text{e}{+12}$.
- De permutatienuldistributie voor F-teststatistiek bekomen door f te berekenen voor iedere $g \in \mathcal{G}$ of voor een willekeurige steekproef van permutaties uit $\mathcal{G}$.

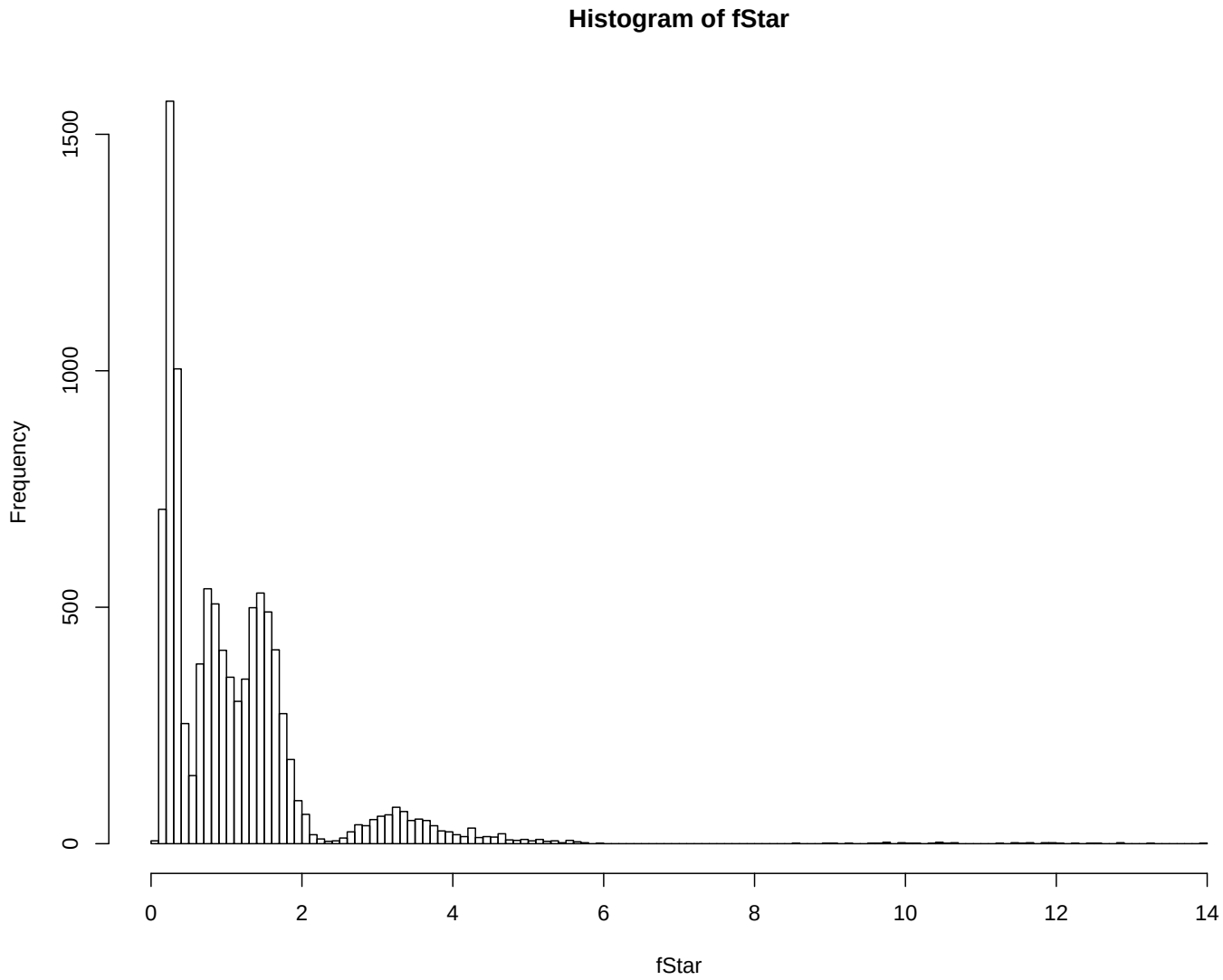
```r
set.seed(165)
B=10000
fOrig=anova(lm(log(length)~dose,data=dna))$F[1]
fStar=sapply(X=1:B, FUN=function(b,y,groep)
      {anova(lm(y~sample(groep)))$F[1]},y=log(dna$length)
                ,groep=dna$dose)
fOrig
```

```
## [1] 367.7574
```

```r
pval2=(sum(fStar>=fOrig)+1)/(B+1)
pval2
```

```
## [1] 9.999e-05
```

```
hist(fStar,breaks=100)
```



- De benaderde p-waarde is $p << 0.001$, dus het effect van de dosis van DMH op DNA beschadiging in levercellen van ratten extreem significant is.
- Via een posthoc analyse zouden we de groepen paarsgewijs met elkaar kunnen vergelijken.
- Merk op, dat als het locatie-shift model niet opgaat, het moeilijk is om inzicht te krijgen in de precieze alternatieve hypothese van de toets.

Kruskal-Wallis Rank Test

- De Kruskal-Wallis Rank Test (KW-test) is een niet-parameterisch alternatief voor de ANOVA F-test.
- De klassieke F-teststatistiek kan geschreven worden als

$$F = \frac{\text{SST}/(g-1)}{\text{SSE}/(n-g)} = \frac{\text{SST}/(g-1)}{(\text{SSTot} - \text{SST})/(n-g)},$$

- met g het aantal groepen.
- SSTot hangt enkel af van uitkomsten $\mathbf{y}$ en zal niet variëren bij permutaties.
- Voldoende om SST als teststatistiek te gebruiken.

$$\text{SST} = \sum_{j=1}^{t} n_j (\bar{Y}_j - \bar{Y})^2$$

- De KW teststatistiek maakt gebruik van SST maar dan gebaseerd op de rank-getransformeerde uitkomsten[1],

$$\text{SST} = \sum_{j=1}^{g} n_j \left(\bar{R}_j - \bar{R}\right)^2 = \sum_{j=1}^{t} n_j \left(\bar{R}_j - \frac{n+1}{2}\right)^2,$$

- met $\bar{R}_j$ het gemiddelde van de ranks in behandelingsgroep j, en $\bar{R}$ het gemiddelde van alle ranks,

$$\bar{R} = \frac{1}{n}(1 + 2 + \cdots + n) = \frac{1}{n}\frac{1}{2}n(n+1) = \frac{n+1}{2}.$$

- De KW teststatistiek wordt gegeven door

$$KW = \frac{12}{n(n+1)} \sum_{j=1}^{g} n_j \left(\bar{R}_j - \frac{n+1}{2}\right)^2.$$

- De factor $\frac{12}{n(n+1)}$ zorgt ervoor dat KW een eenvoudige asymptotische nuldistributie heeft. In het bijzonder, onder H_0, als $\min(n_1, \ldots, n_g) \to \infty$,

$$KW \to \chi^2_{t-1}.$$

[1] we veronderstellen afwezigheid van *ties*

- De exacte KW-test kan uitgevoerd worden via de berekening van de permutatienuldistributie (die enkel afhangt van $n_1, \ldots, n_g$) voor het testen van

$$H_0 : f_1 = \ldots = f_g \text{ vs } H_1 : \text{ minstens twee gemiddelden verschillend.}$$

- Om toe te laten dat H_1 geformuleerd is in termen van gemiddelden, moet locatie-shift verondersteld worden.
- Indien locatie-shift niet opgaat, zou H_1 eigenlijk geformuleerd moeten worden in termen van probabilistische indexen:

$$H_0 : f_1 = \ldots = f_g \text{ vs } H_1 : \exists\, j, k \in \{1, \ldots, g\} : \mathrm{P}\left[Y_j \geq Y_k\right] \neq 0.5$$

DNA Schade Voorbeeld

```
kruskal.test(length~dose,data=dna)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  length by dose
## Kruskal-Wallis chi-squared = 14, df = 3, p-value = 0.002905
```

- Op het 5% significantieniveau kan de nulhypothese worden verworpen.
- R-functie `kruskal.test` heeft enkel de asymptotische benadering voor berekening van *p*-waarden.
- Met slechts 6 observaties per groep, is dit geen optimale benadering van de exacte *p*-waarde!

- Met de coin R package kan de exacte p-waarde wel berekenen

```
library(coin)
kwPerm=kruskal_test(length~dose,data=dna,
                    distribution=approximate(B=100000))
kwPerm
```

```
##
##  Approximative Kruskal-Wallis Test
##
## data:  length by dose (0, 1.25, 2.5, 5)
## chi-squared = 14, p-value = 0.00036
```

- We besluiten dat er een extreem significant verschil is in distributie van de DNA schade ten gevolge van de dosis.

- Posthoc analyse a.d.h.v WMW testen.

```
pairwise.wilcox.test(dna$length,dna$dose)
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  dna$length and dna$dose
##
##      0     1.25  2.5
## 1.25 0.013 -     -
## 2.5  0.013 0.818 -
## 5    0.013 0.721 0.788
##
## P value adjustment method: holm
```

- Alle DMH behandelingen verschillen significant van de controle.
- U1 niet in `pairwise.wilcox.test` output. Puntschatter op de kans op hogere DNA-schade?

```
nGroep <- table(dna$dose)
probInd <- combn(levels(dna$dose),2,function(x)
          {
          test=wilcox.test(length~dose,subset(dna,dose%in%x))
          return(test$statistic/prod(nGroep[x]))
          }
          )
names(probInd) <- combn(levels(dna$dose),2,paste,collapse="vs")
probInd
```

```
##    0vs1.25    0vs2.5      0vs5 1.25vs2.5   1.25vs5    2.5vs5
## 0.0000000 0.0000000 0.0000000 0.4444444 0.2777778 0.3333333
```

Omdat er twijfels zijn of het locatie-shift model geldig is, doen we enkel uitspraken in termen van de probabilistische index.

Conclusie

- Er extreem significant verschil is in de distributie van de DNA-schade metingen tengevolge van de DMH behandeling ($p < 0.001$ KW-test).
- DNA-schade is meer waarschijnlijk na behandeling met DMH dan in de controle behandeling (alle p=0.013, WMW-testen).
- De kansen op hogere DNA-schade na blootstelling aan DMH bedraagt 100% (Berekenen van BI op kans buiten bestek van cursus).
- Er zijn geen significante verschillen in de distributies van de comit-lengtes tussen de DMH behandelingen onderling ($p =$ 0.72-0.82).
- DMH vertoont dus al bij de lage dosis genotoxische effecten.
- (Alle paarsgewijze testen werden gecorrigeerd voor multiple testing d.m.v. Holm's methode).