CANCER RESEARCH UK | Cambridge Institute

Together we are beating cancer

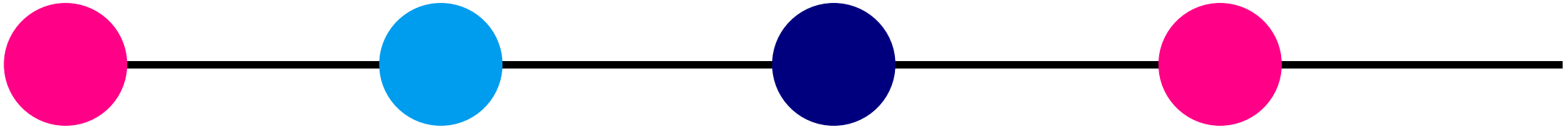
Luca Porcu & Chandra Chilamakuri (Bioinformatics core)

21st February 2025

Linear regression models

Fixed-effects models

Process flow



Stage 1

Effects and regression models

Stage 2

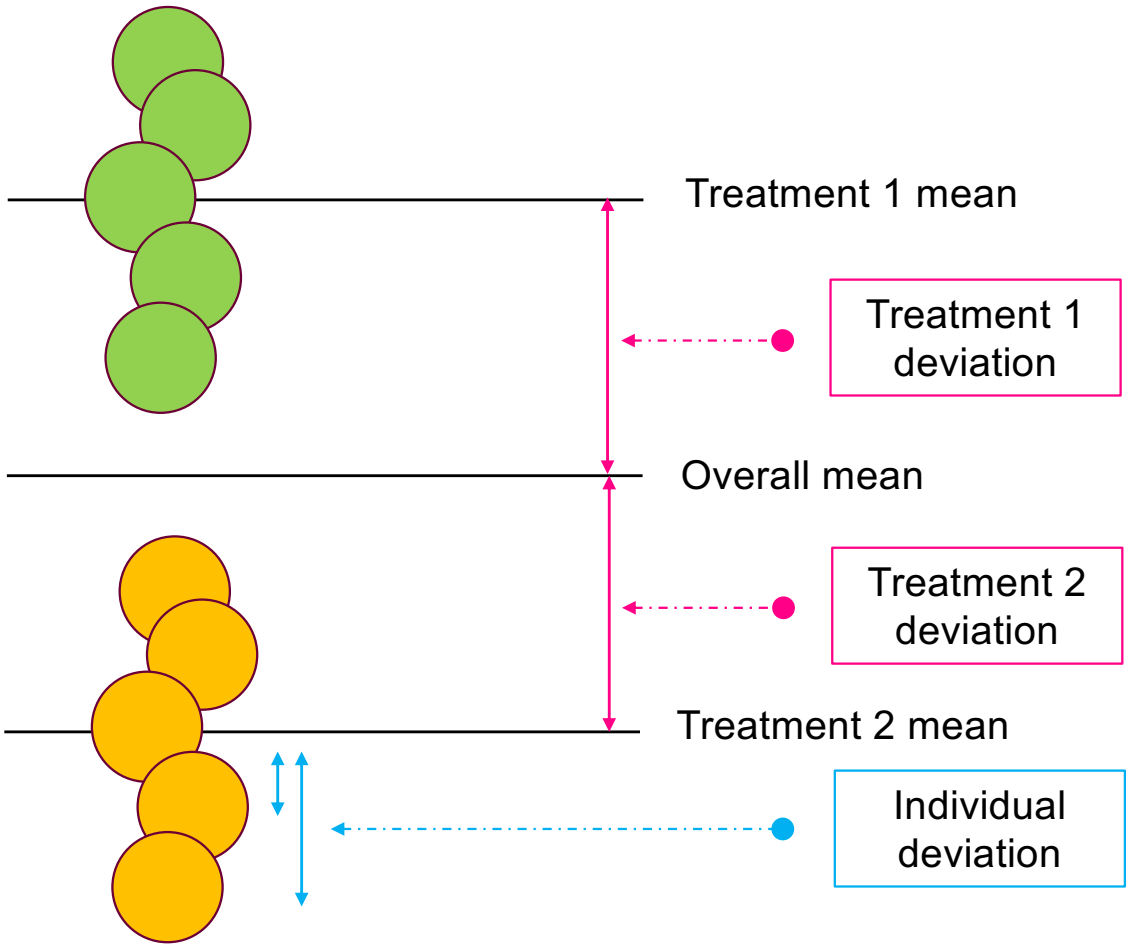
Analysis of variance (ANOVA)

Stage 3

Simple regression model

Stage 4

Multivariable regression model



CANCER RESEARCH UK | Cambridge Institute

Analysis of variance (ANOVA)

Definition and classification

10.15 -10.40 am

Together we are beating cancer

Fisher's one-way ANOVA

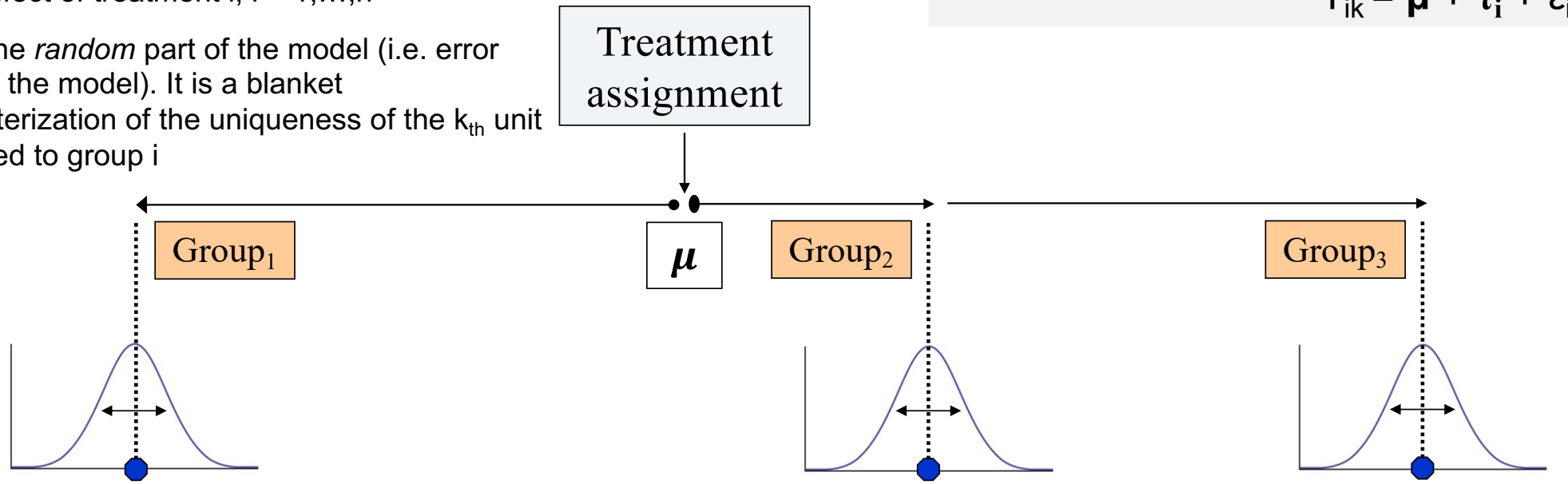
1. The unit k (e.g. mouse), k = 1,…,u_i; N = Σ_i u_i

2. τ_i: effect of treatment i, i = 1,…,n

3. ε_{ik}: the *random* part of the model (i.e. error term of the model). It is a blanket characterization of the uniqueness of the k_{th} unit assigned to group i

Equation of the statistical model:

$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$



Assumptions of ANOVA (ANalysis Of VAriance) models are the following:

- The effect of each factor is additive on μ (i.e. population mean) parameter
- ε_{ik} is assumed to be independent of one another and normally distributed with mean = 0 and common standard deviation = σ

Fisher's one-way ANOVA

Hypothesis to test: $\tau_1 = \ldots = \tau_n$

Test statistic:

Source of variation	**Sum of Squares**	**Degrees of freedom**	**Mean Squares**	**$F_{df1,df2}$**	**P-value**
Treatment	$SSB = \Sigma_i u_i (m_i - M)^2$	$df1 = n - 1$	$MSB = SSB / (n - 1)$	MSB / MSE	0.023
Residuals	$SSE = \Sigma_i \Sigma_k (x_{ik} - m_i)^2$	$df2 = N - n$	$MSE = SSE / (N - n)$		
Total	$SST = SSB + SSE$	$df_{TOT} = N - 1$			

Legend: m_i is the sample mean of group i

Note: the ANOVA divides the total variation in the response into parts.

R implementation		
Step	**Aim**	**R function**
1	We should fit our data to the ANOVA model	fitModel = lm(*Response* ~ *Treatment*, data=*dSet*)
2	We can get R to produce an ANOVA table	anova(fitModel)

Fisher's two-way ANOVA

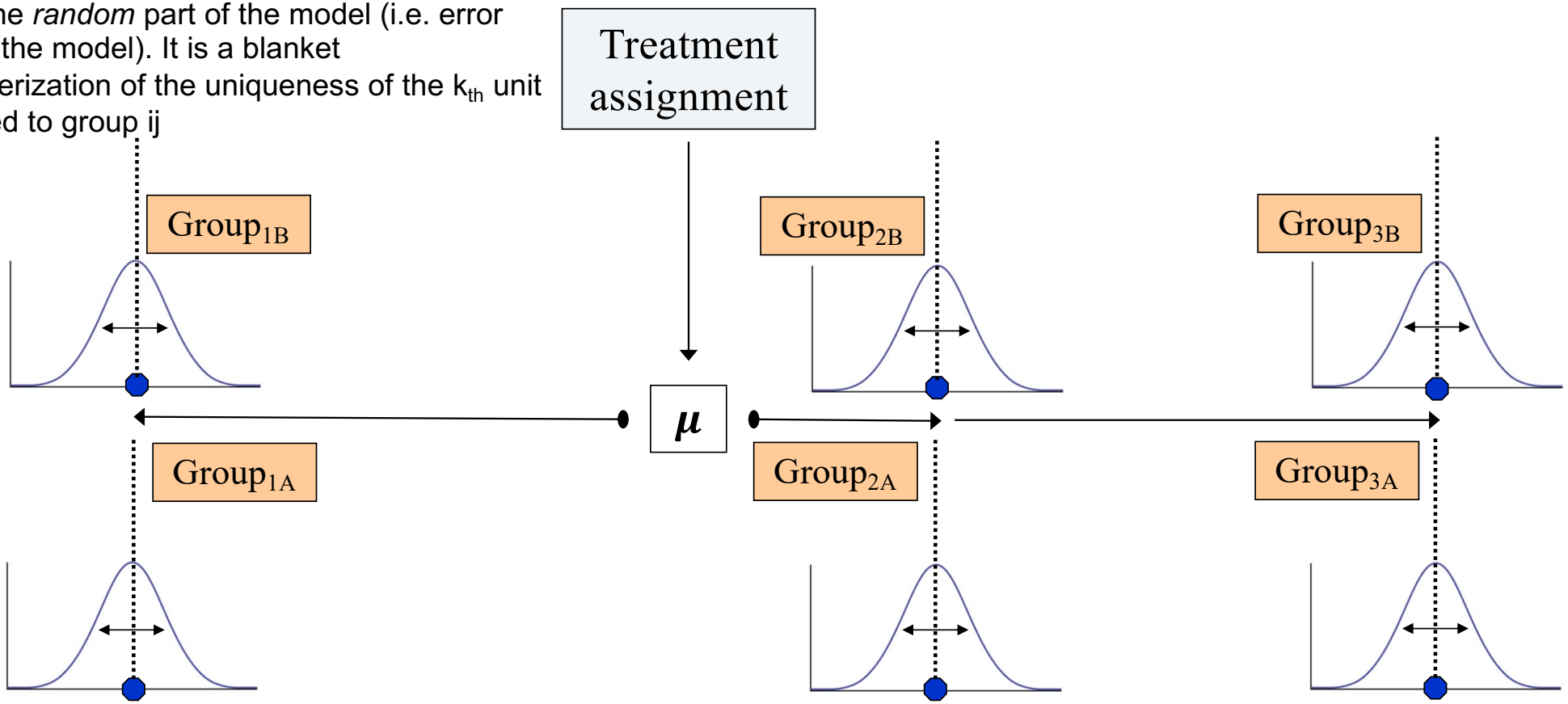
1. The unit k (e.g. mouse), k = 1,…,u_{ij}; N = Σ_{ij} u_{ij}

2. τ_i: effect of treatment i, i = 1,…,n; η_j: effect of treatment j, j = 1,…,r

3. ε_{ijk}: the *random* part of the model (i.e. error term of the model). It is a blanket characterization of the uniqueness of the k_{th} unit assigned to group ij

Equation of the statistical model:

$$Y_{ijk} = \mu + \tau_i + \eta_j + \varepsilon_{ijk}$$



Fisher's two-way ANOVA

Hypothesis to test n.1: $\tau_1 = \ldots = \tau_n$

Hypothesis to test n.2: $\eta_1 = \ldots = \eta_r$

Test statistic:

Source of variation	**Sum of Squares**	**Degrees of freedom**	**Mean Squares**	**$F_{df1,df2}$**	**P-value**
Treatment τ	$SSB_\tau = \Sigma_i u_i (m_i - M)^2$	$df1_\tau = n - 1$	$MSB_\tau = SSB_\tau / (n - 1)$	MSB_τ / MSE	0.023
Treatment η	$SSB_\eta = \Sigma_j u_i (m_j - M)^2$	$df1_\eta = r - 1$	$MSB_\eta = SSB_\eta / (r - 1)$	MSB_η / MSE	0.150
Residuals	$SSE = \Sigma_i \Sigma_k (x_{ijk} - m_{ij})^2$	$df2 = N - (n \cdot r)$	$MSE = SSE / [N - (n \cdot r)]$		
Total	$SST = SSB_\tau + SSB_\eta$ SSE	$df_{TOT} = N - 1$			

Note: the ANOVA divides the total variation in the response into parts.

R implementation		
Step	**Aim**	**R function**
1	We should fit our data to the ANOVA model	fitModel = lm(*Response* ~ $Treat_\tau$ + $Treat_\eta$, data=*dSet*)
2	We can get R to produce an ANOVA table	anova(fitModel)

Fisher's two-way ANOVA with interaction

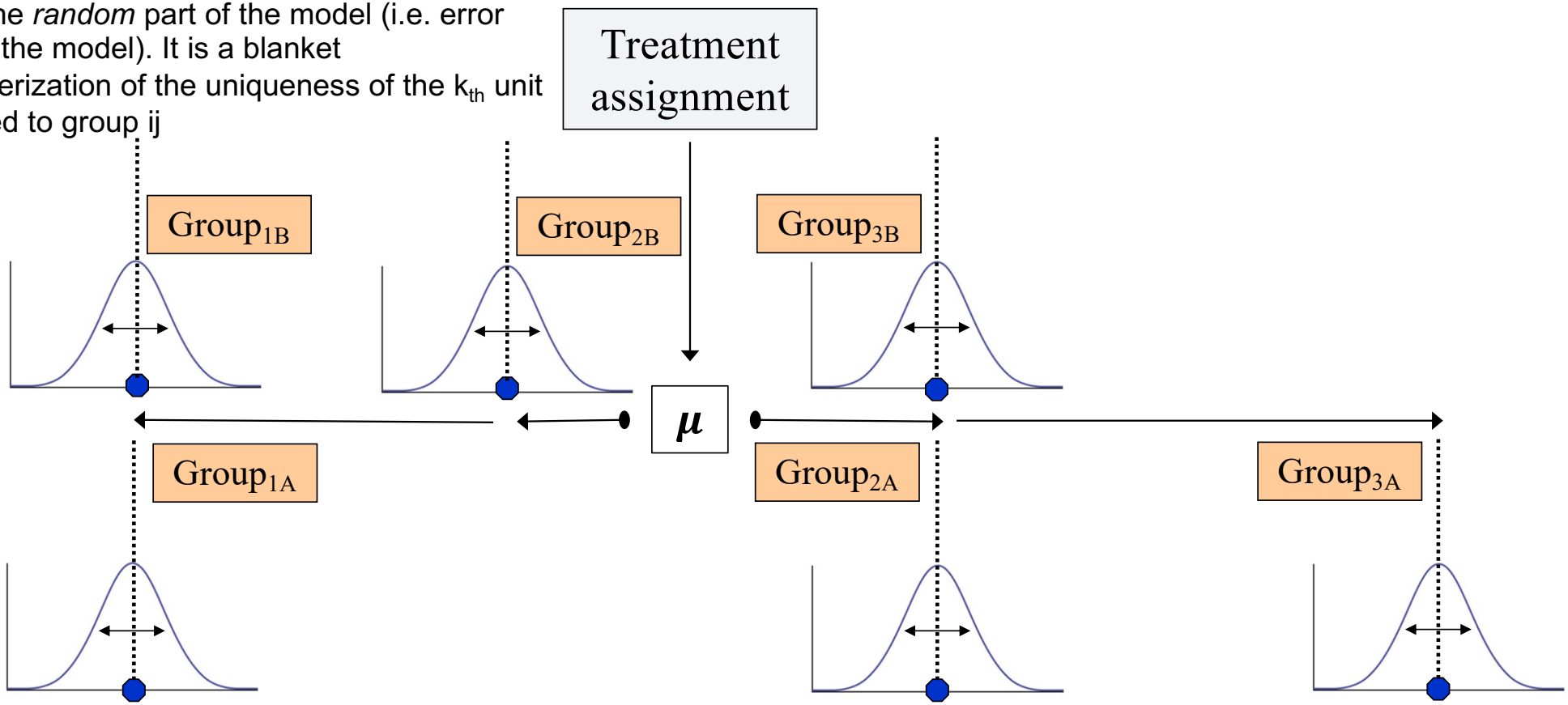
1. The unit k (e.g. mouse), k = 1,…,u_{ij}; N = Σ_{ij} u_{ij}

2. τ_i: effect of treatment i, i = 1,…,n; η_j: effect of treatment j, j = 1,…,r

3. ε_{ijk}: the *random* part of the model (i.e. error term of the model). It is a blanket characterization of the uniqueness of the k_{th} unit assigned to group ij

Equation of the statistical model:

$$Y_{ijk} = \mu + \tau_i + \eta_j + \tau_i{:}\eta_j + \varepsilon_{ijk}$$



Fisher's two-way ANOVA with interaction

Hypothesis to test n.1: $\tau_1 = \ldots = \tau_n$ Hypothesis to test n.2: $\eta_1 = \ldots = \eta_r$ Hypothesis to test n.3: $\tau{:}\eta = 0$

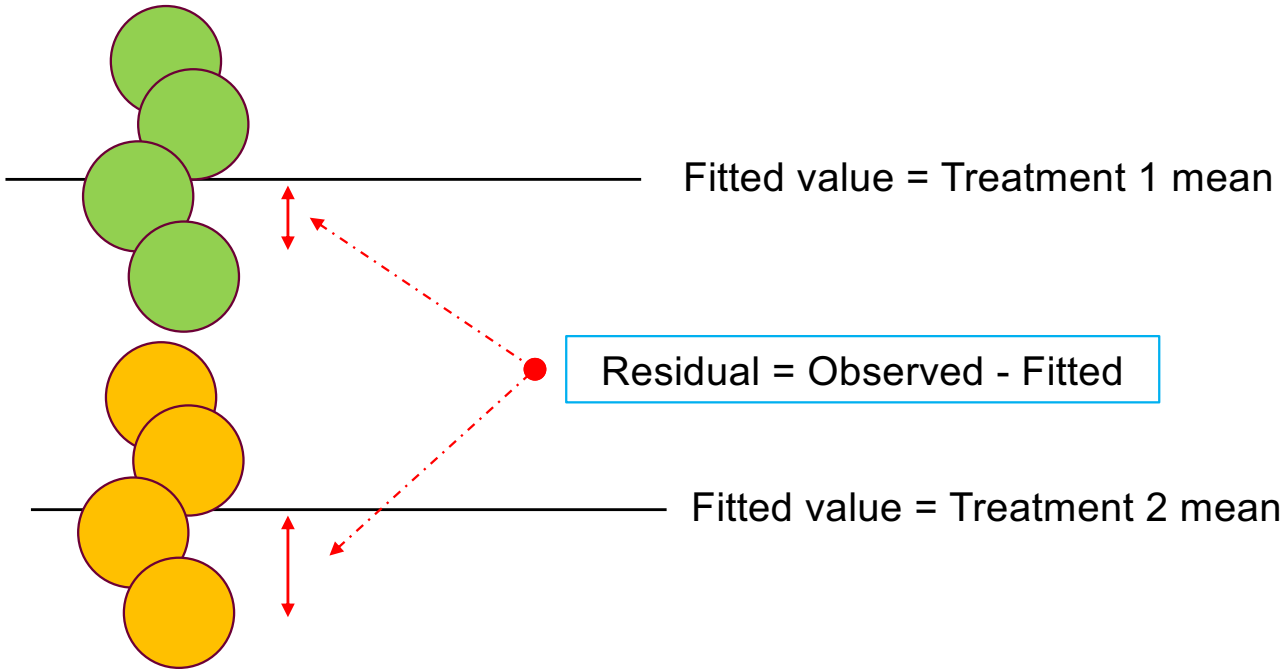
Test statistic:

Source of variation	**Sum of Squares**	**Degrees of freedom**	**Mean Squares**	**$F_{df1,df2}$**	**P-value**
Treatment τ	$SSB_\tau = \Sigma_i\, u_i\, (m_i - M)^2$	$df1_\tau = n - 1$	$MSB_\tau = SSB_\tau / df1_\tau$	MSB_τ / MSE	0.023
Treatment η	$SSB_\eta = \Sigma_j\, u_i\, (m_j - M)^2$	$df1_\eta = r - 1$	$MSB_\eta = SSB_\eta / df1_\eta$	MSB_η / MSE	0.150
Interaction $\tau{:}\eta$	$SSB_{\tau:\eta} = \Sigma_{jj}\, u_{ij}\, (m_{ij} - m_j - m_i + M)^2$	$df1_{\tau:\eta} = (n - 1) \cdot (r - 1)$	$MSB_{\tau:\eta}\, SSB_{\tau:\eta} / df1_{\tau:\eta}$	$MSB_{\tau:\eta}$ / MSE	0.401
Residuals	$SSE = \Sigma_i \Sigma_k\, (x_{ijk} - m_{ij})^2$	$df2 = N - (n \cdot r)$	$MSE = SSE / df2$		
Total	$SST = SSB_\tau + SSB_\eta + SSB_{\tau:\eta} + SSE$	$df_{TOT} = N - 1$			

Note: the ANOVA divides the total variation in the response into parts.

R implementation		
Step	**Aim**	**R function**
1	We should fit our data to the ANOVA model	fitModel = lm(Response ~ $Treat_\tau$ * $Treat_\eta$, data=*dSet*)
2	We can get R to produce an ANOVA table	anova(fitModel)

Diagnostics: residuals



The residuals are equal to the difference between the observations and the corresponding fitted values.

R implementation		
Step	**Aim**	**R function**
1	We should fit our data to the ANOVA model	fitModel = lm(*Response* ~ *Predictor*, data=*dSet*)
2	We want to obtain the *residuals* of the model	*dSet*$resid = resid(fitModel)

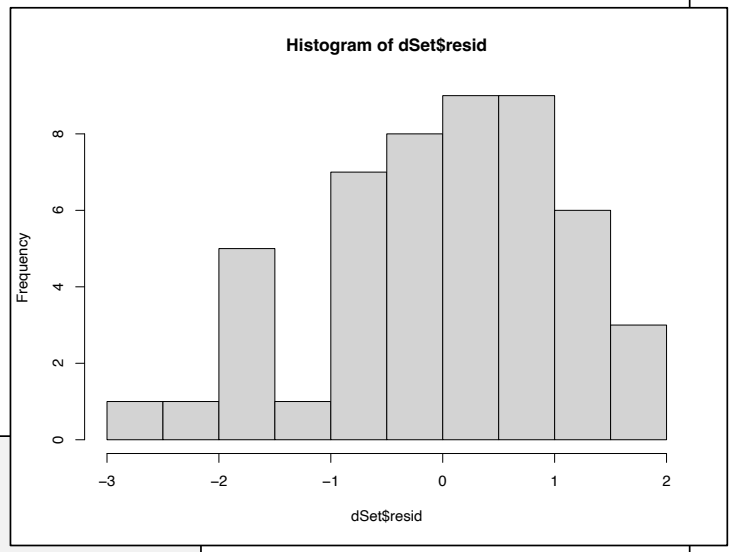
Diagnostics: residuals

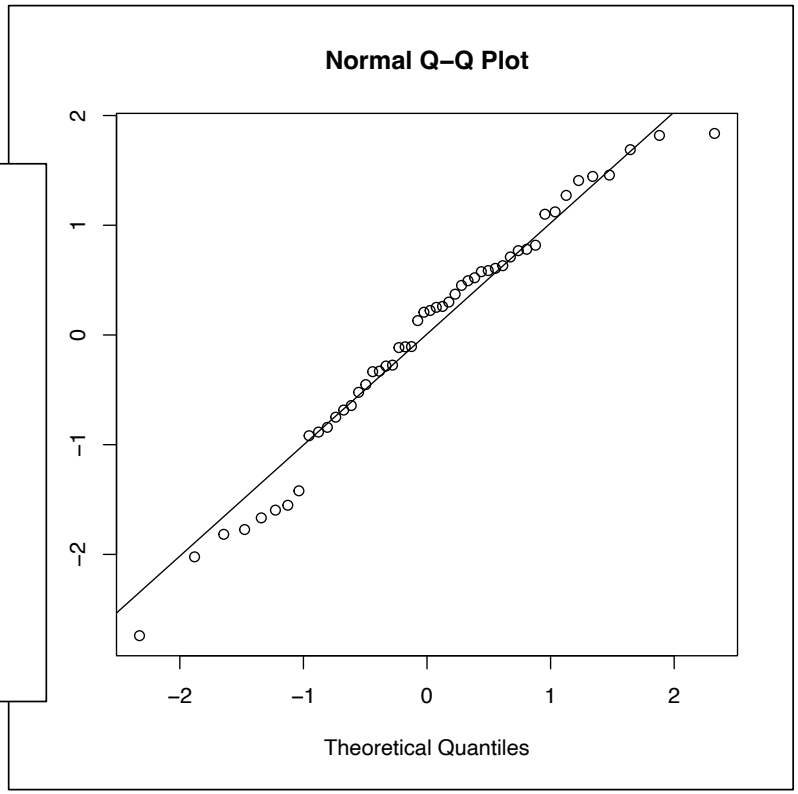
Shapiro-Wilk normality test

data: dSet$resid
W = 0.97324, p-value = 0.3119

Bartlett test of homogeneity of variances

data: resid by Predictor
Bartlett's K-squared = 1.5374, df = 1, p-value = 0.215





R implementation

Step	Aim	Tool	R function
1	We should plot the *residuals*	Histogram	hist(*dSet*$resid)
		Q-Q plot	qqnorm(*dSet*$resid); qqline(*dSet*$resid)
2	We could test the assumptions	Shapiro-Wilk *normality* test	shapiro.test(*dSet*$resid)
		Bartlett's *homoscedasticity* test	bartlett.test(resid ~ *Predictor*, data = *dSet*)

Diagnostics: residuals

Equation of the statistical model:

$$Y_{ijk} = \mu + \tau_i + \eta_j + \tau_i{:}\eta_j + \varepsilon_{ijk}$$

Assumptions of normality and homoscedasticity **must be satisfied** by residuals of single treatment group and **combined** treatment groups (e.g. **pooled** residuals of $Group_{1A}$, $Group_{3A}$ and $Group_{3B}$.

Source of problems and possible solutions

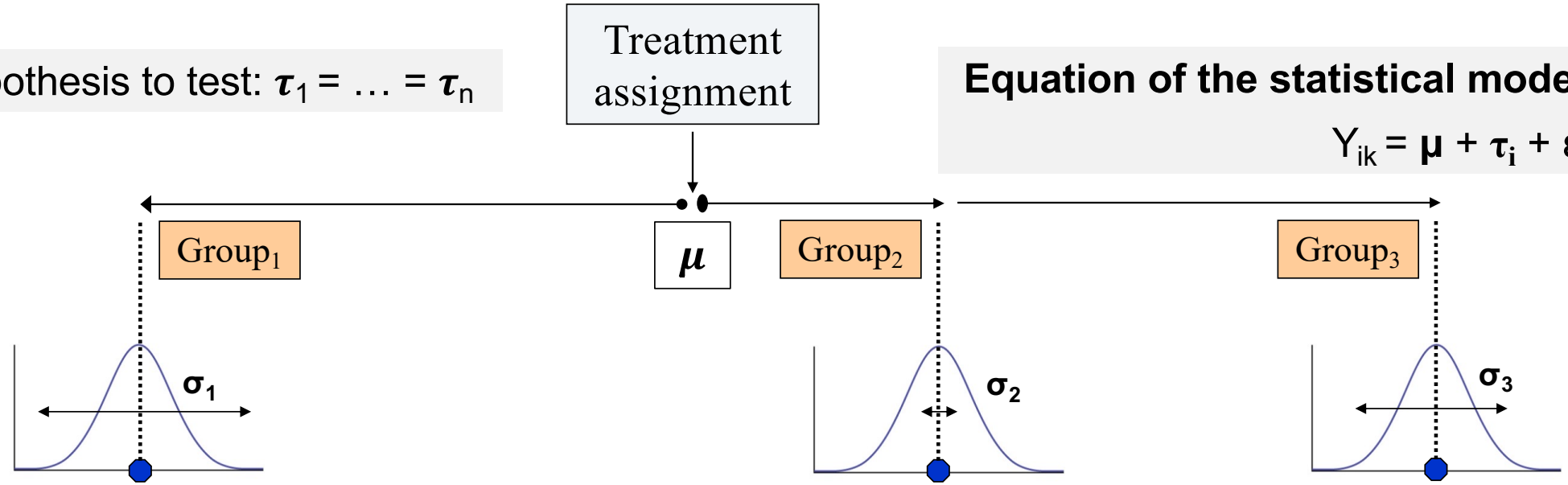
Solution	Normality	Unequal variance	Outliers
Welch's one-way ANOVA		✓	
Weighting		✓	
Distribution-free methods ⊙	✓	✓	✓
Data transformation	✓	✓	✓

⊙ e.g. Kruskal-Wallis test

Welch's one-way ANOVA

The Welch version of one-way ANOVA do not assume that all the groups are sampled from populations with equal variances.

Hypothesis to test: $\tau_1 = \ldots = \tau_n$



Equation of the statistical model:

$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

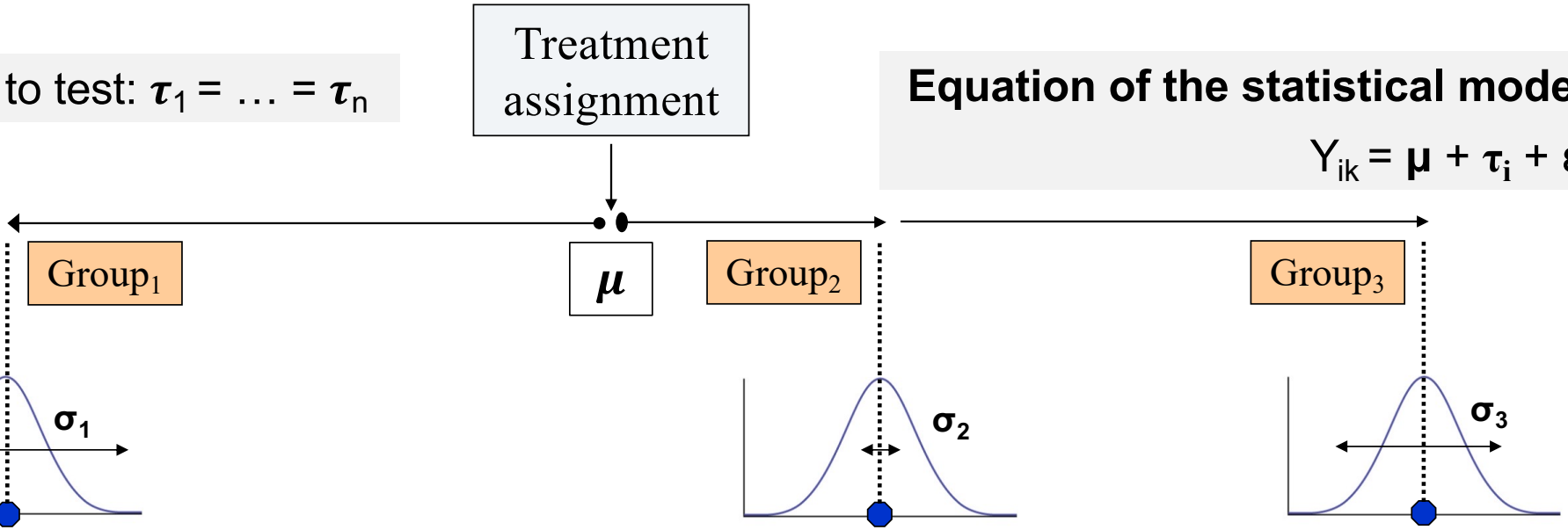
Assumptions of ANOVA (ANalysis Of VAriance) models are the following:

- The effect of each factor is additive on μ (i.e. population mean) parameter
- ε_{ik} is assumed to be independent of one another and normally distributed with mean = 0. **Standard deviation could be different between groups: $\sigma_i \neq \sigma_j$, $i \neq j$**

Weighted least square

The *gls* function of the R package *nlme* using generalized least squares. The errors are allowed to be correlated and/or have unequal variances.

Hypothesis to test: $\tau_1 = \ldots = \tau_n$



Equation of the statistical model:

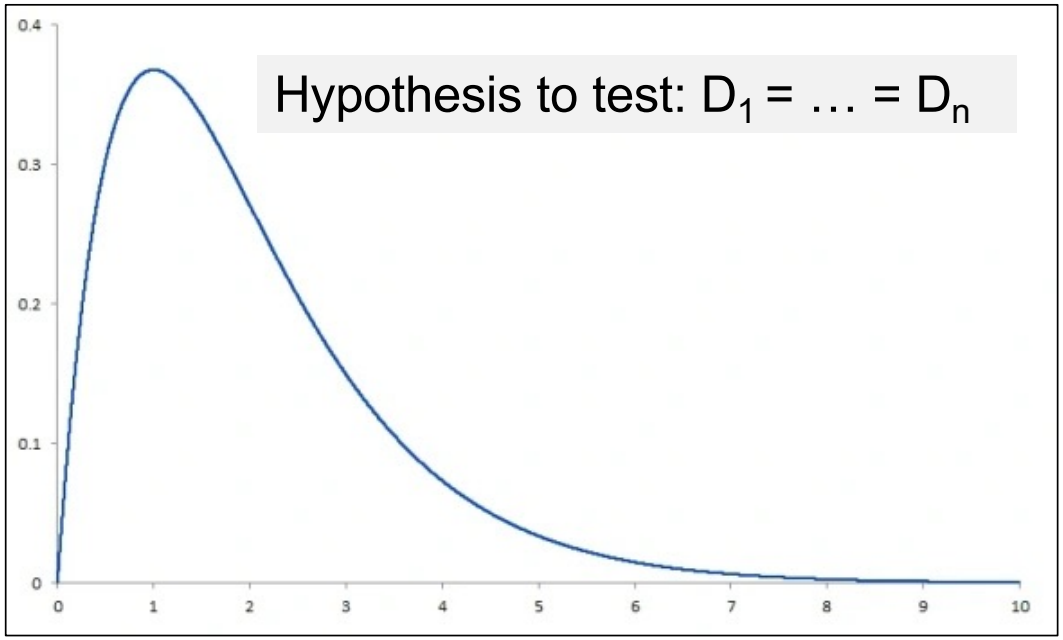
$$Y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

Assumptions of linear models fitted with the gls function are the following:

- The effect of each factor is additive on μ (i.e. population mean) parameter
- ε_{ik} could be **correlated**. They are normally distributed with mean = 0. **Standard deviation could be different between groups: $\sigma_i \neq \sigma_j$, $i \neq j$**

Kruskal-Wallis test

The Kruskal-Wallis test (i.e. one-way ANOVA on ranks) works on ranks. It tests whether samples originate from the same distribution.



10.2	24.7	33.2	..	99.7	99.9

Replacement of data by their ranks

1	2	3	..	N-1	N

Assumptions of Kruskal-Wallis test are the following:

- We only assume that the observations in the data set are independent of each other.

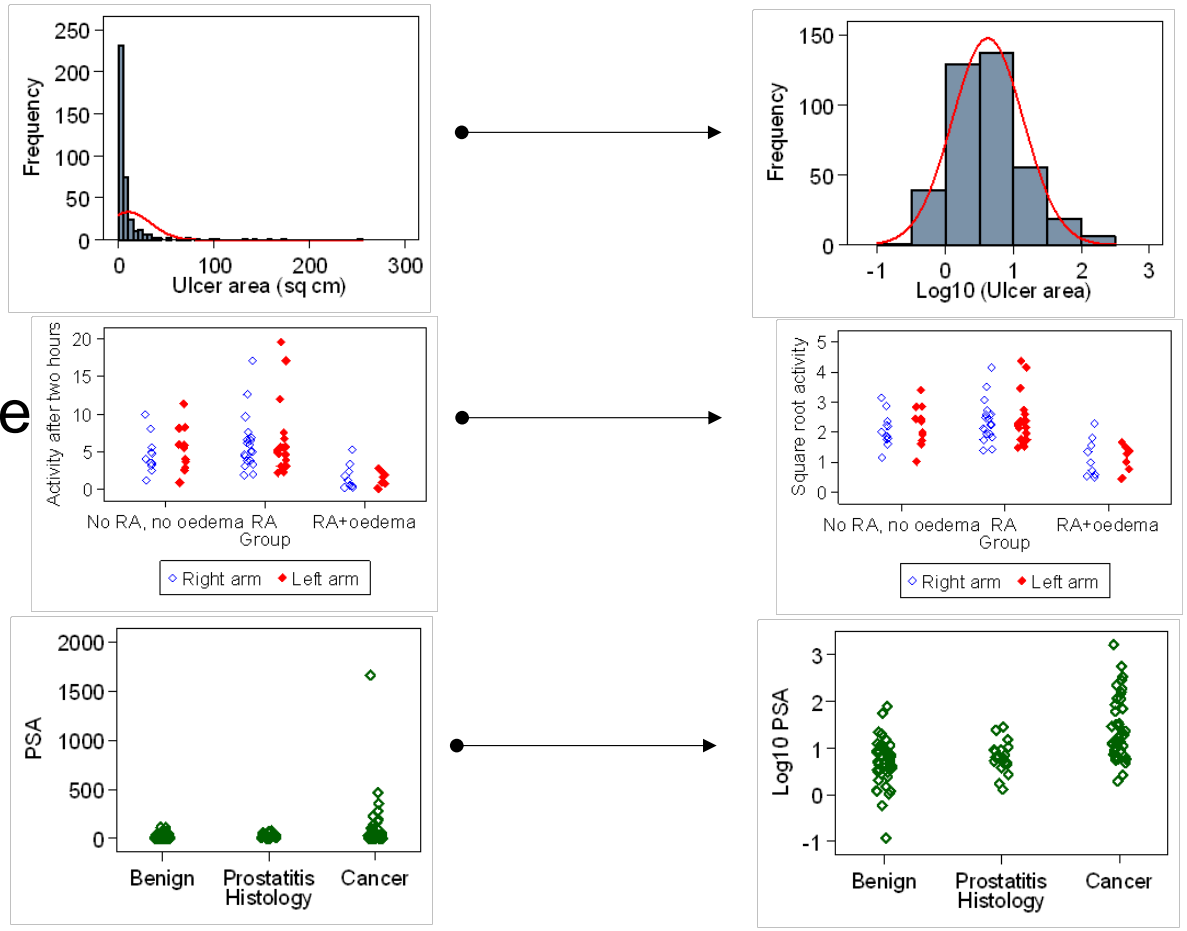
R functions

R implementation		
Test	**R**	
Welch's one-way ANOVA	Function	• oneway.test(*Response ~ Predictor*, data = *dSet*, var.equal = FALSE)
	Output	One-way analysis of means (not assuming equal variances) data: Response and Predictor F = 118.34, num df = 1.000, denom df = 45.143, p-value = 3.342e-14
Weighted least square	Function	• fitModel <- gls(*Response ~ Predictor*, weights = varIdent(form= ~ 1 | *Predictor*), data = *dSet*) • summary(fitModel)
	Output	Variance function: Structure: Different standard deviations per stratum Formula: ~1 | Predictor Parameter estimates: 1 2 1.000000 1.293192 Coefficients: Value Std.Error t-value p-value (Intercept) -0.001177 0.1890228 -0.006228 0.9951 Predictor 3.361487 0.3090014 10.878548 0.0000
Kruskal-Wallis	Function	• kruskal.test(*Response ~ Predictor*, data = *dSet)*
	Output	Kruskal-Wallis rank sum test data: Response by Predictor Kruskal-Wallis chi-squared = 34.222, df = 1, p-value = 4.917e-09

Data transformation

We can transform the data mathematically…

- to make them fit the normality more closely
- to obtain more similar variance
- to handle outliers



Data transformation

Common and useful transformations of the response variable:

1. the logarithm ($x_i > 0$, i=1,…n)
2. the square root ($x_i \geq 0$, i=1,…n)
3. the square power ($x_i \geq 0$, i=1,…n)
4. the ranks (e.g. Welch's one-way ANOVA on ranks)

http://bioinformatics-core-shared-training.github.io/IntroductionToStats/practical.html