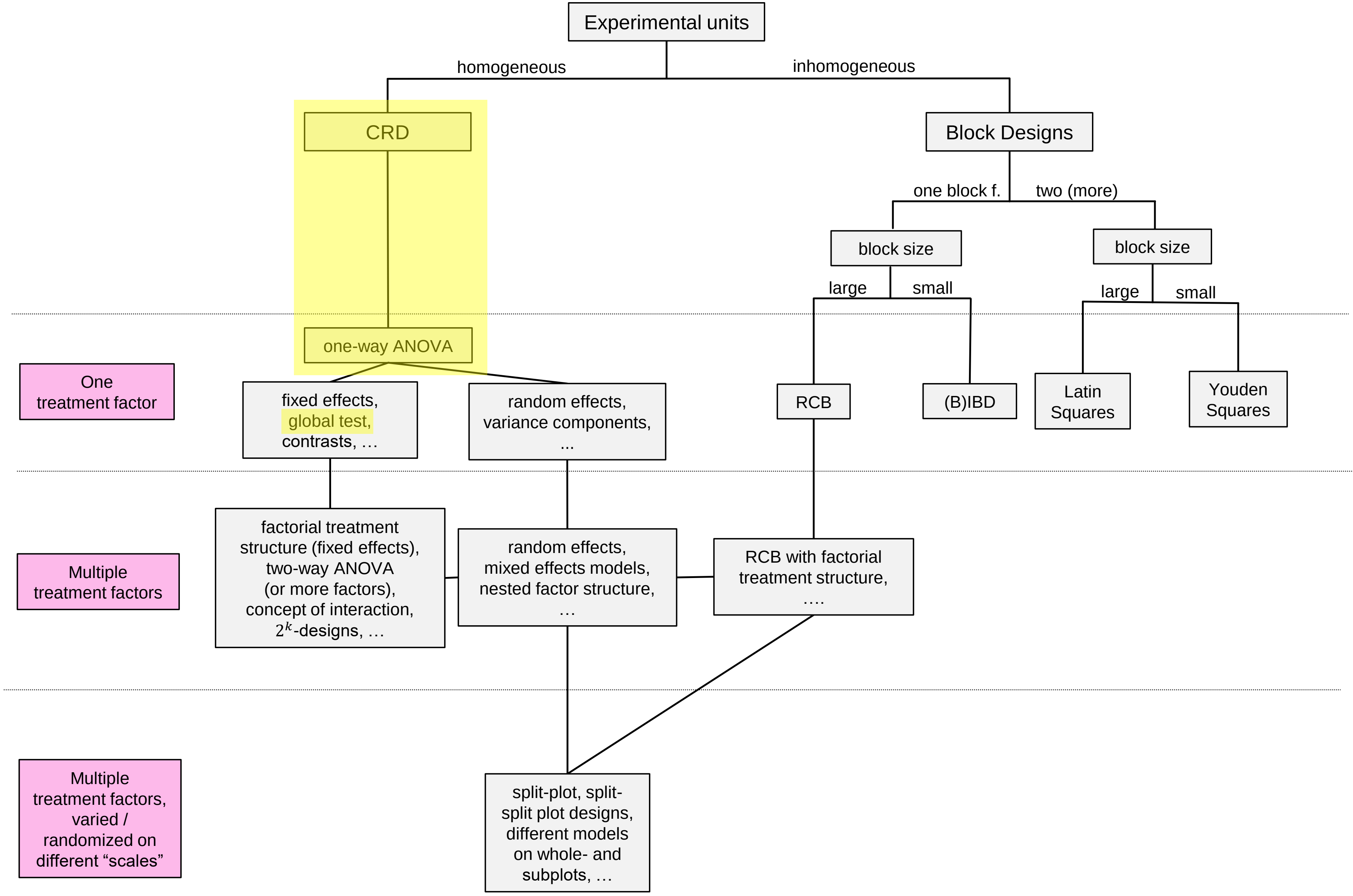
Completely Randomized Designs (CRD)
One-Way ANOVA



Experimental units

homogeneous

inhomogeneous

CRD

one-way ANOVA

Block Designs

one block f.

two (more)

block size

large

small

block size

large

small

One treatment factor

fixed effects, global test, contrasts, …

random effects, variance components, ...

RCB

(B)IBD

Latin Squares

Youden Squares

Multiple treatment factors

factorial treatment structure (fixed effects), two-way ANOVA (or more factors), concept of interaction, 2^k-designs, …

random effects, mixed effects models, nested factor structure, …

RCB with factorial treatment structure, ….

Multiple treatment factors, varied / randomized on different “scales”

split-plot, split-split plot designs, different models on whole- and subplots, …

Similar to Lawson (2015)

Example: Meat Storage Study (Kuehl, 2000, Example 2.1)

- A researcher wants to investigate the **effect of packaging** on **bacterial growth** of stored meat.

- Some studies suggested controlled gas atmospheres as alternatives to existing packaging.

- Different **treatments** (= packaging types)
  - Commercial plastic wrap (ambient air)
  - Vacuum package

  } Current techniques (control groups)

  - 1% CO, 40% O_2, 59% N
  - 100% CO_2

  } New techniques

- **Experimental units**: 12 beef steaks (about 75g each).

- Measure effectiveness of packaging by measuring how successful they are in **suppressing bacterial growth**.

Example: Meat Storage Study

- Three beef steaks were **randomly assigned** to each of the packaging conditions.
- Each steak was packaged **separately** in its assigned condition.
- **Response**: (logarithm of the) number of bacteria per square centimeter.
- The number of bacteria was measured after nine days of storage at 4 degrees Celsius in a standard meat storage facility.

First Step (Always): Exploratory Data Analysis

- If very few observations: Plot **all** data points.
- With more observations: Use **boxplots** (side-by-side).
- Alternatively: Violin-plots, histogram side-by-side, …
- See examples in R: `02_meat_storage.R`

Such plots typically give you the same (or even more) information as a formal analysis (see later).

Side Remark: Factors

- Categorical variables are also called **factors**.

- The different values of a factor are called **levels**.

- Factors can be **nominal** or **ordinal** (= ordered).
  - Hair color: {black, blond, …} *nominal*
  - Gender: {male, female} *nominal*
  - Treatment: {commercial, vacuum, mixed, CO_2} *nominal*
  - Income: {<50k, 50-100k, >100k} *ordinal*

- Useful functions in R:
  - `factor`
  - `as.factor`
  - `levels`

Completely Randomized Design: Formal Setup

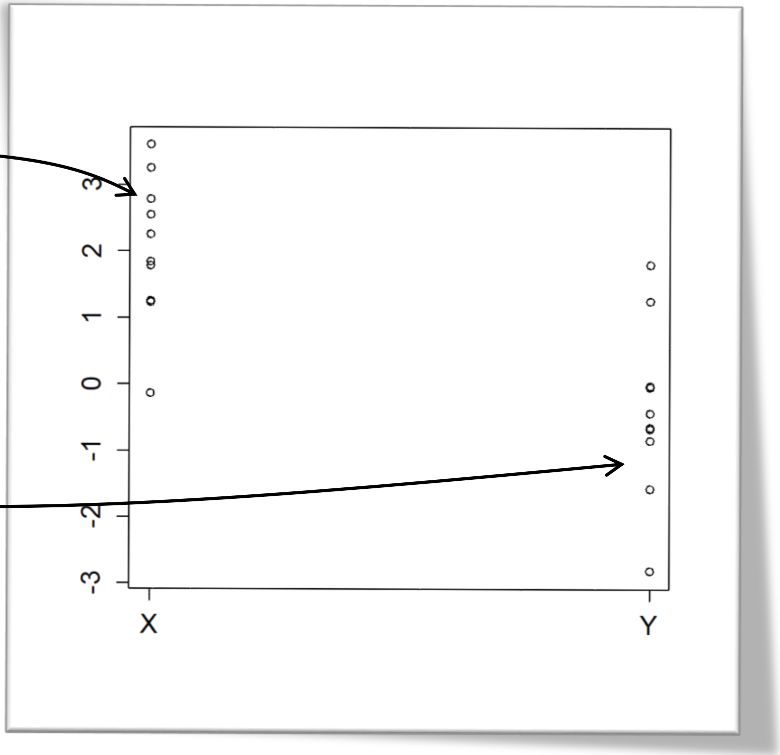
- Compare g treatments.
- Available resources: N experimental units
- Need to **assign** the N experimental units to g different **treatments** (**groups**) having n_i observations each, $i = 1, \dots, g$ (of course: $n_1 + n_2 + \dots + n_g = N$).
- Use randomization:
  - Choose n_1 units **at random** to get treatment 1,
  - n_2 units **at random** to get treatment 2,
  - ...
- The optimal choice of $n_1, \dots, n_g$ depends on the primary research question (if $n_1 = n_2 = \cdots = n_g$ the design is called **balanced**).
- This randomization produces a so called **completely randomized design (CRD)**.

Setting up the Model

- Remember the research question: “Is there an **effect of packaging** on **bacterial growth** of stored meat?”
- Need to set up a **model** in order to do **statistical inference**.
- **Good message**: The problem looks rather easy.
- **Bad message**: Some complications ahead regarding parametrization.

Remember: Two Sample t-Test for Unpaired Data

- **Model**
  - X_i i.i.d. $\sim N(\mu_X, \sigma^2), i = 1, \dots, n$
  - Y_j i.i.d. $\sim N(\mu_Y, \sigma^2), j = 1, \dots, m$
  - X_i, Y_j independent



- ***t*-Test**
  - $H_0: \mu_X = \mu_Y$
  - $H_A: \mu_X \neq \mu_Y$ (or one-sided)
  - $T = \frac{(\bar{X}_n - \bar{Y}_m)}{S_{pool}\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$ under H_0

- Allows us to perform a **statistical test** or to construct **confidence intervals** for the true (unknown) difference $\mu_X - \mu_Y$.

- Note: Both groups have their “**individual” expected value** but they share a **common variance** (can be extended to more general situations).

From Two to More Groups

- In the meat storage example we had **4** groups.
- Hence, the t-test is **not** directly applicable anymore.
- Could try to construct something using only **pairs** of groups (e.g., doing **all pairwise comparisons**).
- Will do so later. Now we want to **extend** the model that we used for the two sample t-test to the more general situation of $g > 2$ groups.
- As we might run out of letters, we use a **common letter** (say Y) for all groups and put the grouping and replication information in the **index**.

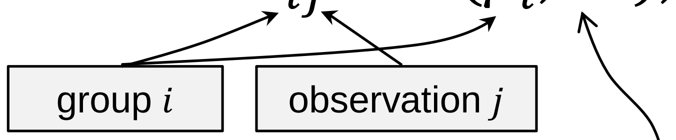
Cell Means Model

- We need **two indices** to distinguish between the different **treatments** (groups) and the different **observations**.

- Let Y_{ij} be the jth observation in the ith treatment group, $i = 1, \dots, g; j = 1, \dots, n_i$.

- **Cell means model**: Every **group** (treatment) has its **own expected** value, i.e.

$$Y_{ij} \sim N(\mu_i, \sigma^2), \text{independent}$$



- Also called **separate means model**.

- Note: Variance is **constant across groups** (as for standard two-sample t-test!)

Illustration of Cell Means Model

- See R-Code: `02_model_illustration.R`
- Or visit https://gallery.shinyapps.io/anova_shiny_rstudio/
- Why **cell means**? Have a look at the meat storage data:

Commercial	*Vacuum*	*Mixed*	*CO_2*
7.66 6.98 7.80	5.26 5.44 5.80	7.41 7.33 7.04	3.51 2.91 3.66

cell

Cell Means Model: Alternative Representation

- We can **“extract”** the **deterministic part** in $Y_{ij} \sim N(\mu_i, \sigma^2)$.

- Leads to

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

with $\epsilon_{ij}\ \mathrm{i.i.d.} \sim N(0, \sigma^2)$.

- The ϵ_{ij}’s are random **“errors”** that fluctuate around **zero**.

- In the regression context:
  - Y is the **response**.
  - Treatment is a categorical **predictor** (a **factor**).
  - Hence, this is nothing else than a **regression model** with a categorical predictor!

Yet Another Representation (!)

- We can also write $\mu_i = \mu + \alpha_i, i = 1, \dots, g$.
- E.g., think of μ as a **“global mean”** and α_i as the corresponding **deviation from the global mean**.
- α_i is also called the ith **treatment effect**.
- This looks like a needless complication now, but will be **very useful later** (with so called factorial treatment structure).
- Unfortunately this model is **not identifiable** anymore.
- Reason: $g + 1$ parameters $(\mu, \alpha_1, \dots, \alpha_g)$ for g different means $(\mu_1, \dots, \mu_g)$.

Ensuring Identifiability

- Need **side constraint**: Many options available.
- Sum of the treatment effects is **zero**, i.e.
  $\alpha_g = -(\alpha_1 + \cdots + \alpha_{g-1})$.
  (R: `contr.sum`)
- Sum of **weighted** treatment effects is zero: …
  (R: do manually)
- Set $\mu = \mu_1$, hence $\alpha_1 = 0, \alpha_2 = \mu_2 - \mu_1, \alpha_3 = \mu_3 - \mu_1, \ldots$
  i.e. a comparison with group 1 as **reference level**.
  (R: `contr.treatment`)
- Only $g - 1$ elements of the treatment effects are allowed to **vary freely**. We also say that the treatment effect has $g - 1$ **degrees of freedom (df)**.

Encoding Scheme of Factors

- The **encoding scheme** (i.e., the side constraint being used) of a factor is called **contrast** in R.

- To summarize: We have a total of g parameters $\mu, \alpha_1, \dots, \alpha_{g-1}$ to parametrize the g group means $\mu_1, \dots, \mu_g$.

- The interpretation of the parameters $\mu, \alpha_1, \dots, \alpha_{g-1}$ **strongly depends** on the parametrization that is being used.

- We will re-discover the word “contrast” in a different way later…

Parameter Estimation

- Choose **parameter estimates** $\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_{g-1}$ such that the model fits the data “well”.

- Criterion: Choose parameter estimates such that

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} \Big(\overset{\text{observed value}}{y_{ij}} - \overbrace{\hat{\mu} - \hat{\alpha}_i}^{\text{fitted value}} \Big)^2$$

is **minimal** (so called **least squares criterion**, exactly as in regression).

- The **predicted values per treatment group** (or: **estimated cell means**) are simply

$$\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$$

Illustration of Goodness of Fit

- See blackboard (incl. definition of **residual**).

Some Notation

Symbol	*Meaning*	*Formula*
$y_{i\cdot}$	**Sum** of all values in **group *i***	$y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$
$\bar{y}_{i\cdot}$	**Mean** of **group *i***	$\bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n_i} y_{i\cdot}$
$y_{\cdot\cdot}$	**Sum** of **all** observations	$y_{\cdot\cdot} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij}$
$\bar{y}_{\cdot\cdot}$	**Overall** (or grand) **mean**	$\bar{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{N}$

Rule: If we replace an index with a **dot** ("·") it means that we are **summing up** the values over that index.

Parameter Estimates, the Other Way Round

- "Obviously", the $\hat{\mu}_i$'s that minimize the least squares criterion are $\hat{\mu}_i = \bar{y}_{i\cdot}$
- Means: **Expectation** of group i is estimated by **sample mean** of group i.
- The $\alpha_i' s$ are then simply estimated by applying the corresponding parametrization, i.e.

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$$

for the sum of weighted treatment effects constraint.

The **fitted** values $\hat{\mu}_i$ (and the **residuals**) are **independent** of the parametrization, but the $\hat{\alpha}_i$'s **(heavily) depend** on it!

Parameter Estimation

- We denote the **residual** (or **error**) **sum of squares** by SS_E, that is

$$SS_E = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

- Estimator for σ^2 is MS_E, **mean squared error**, i.e.

empirical variance in group i

$$\hat{\sigma}^2 = MS_E = \frac{1}{N-g} SS_E = \frac{1}{N-g} \sum_{i=1}^{g} (n_i - 1) s_i^2$$

- This is an **unbiased estimator** for σ^2 (reason for $N - g$ instead of N in the denominator).

- We also say that the error estimate has $N - g$ **degrees of freedom** (N observations, g parameters) or

$$N - g = \sum_{i=1}^{g} (n_i - 1).$$

Estimation Accuracy

- **Standard errors** for the parameters (using the sum of weighted treatment effects constraint).

Parameter	Estimator	Standard Error
μ	$\bar{y}_{..}$	$\sigma/\sqrt{N}$
μ_i	$\bar{y}_{i.}$	$\sigma/\sqrt{n_i}$
α_i	$\bar{y}_{i.} - \bar{y}_{..}$	$\sigma\sqrt{\frac{1}{n_i} - \frac{1}{N}}$
$\mu_i - \mu_j = \alpha_i - \alpha_j$	$\bar{y}_{i.} - \bar{y}_{j.}$	$\sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$

- Therefore, a 95% confidence interval for α_i is given by

$$\hat{\alpha}_i \pm t_{N-g}^{0.975} \cdot \hat{\sigma}\sqrt{\frac{1}{n_i} - \frac{1}{N}}$$

97.5% quantile of t_{N-g} distribution

$N - g$ degrees of freedom because of the degrees of freedom of MS_E

Single Mean Model

- Extending the null hypothesis of the t-test to the situation where $g > 2$, we can (for example) use the (very strong) null hypothesis that there is **no treatment effect at all** on the response.

- In such a setting, all values (also across **different** treatments) fluctuate around the **same “global” mean** μ.

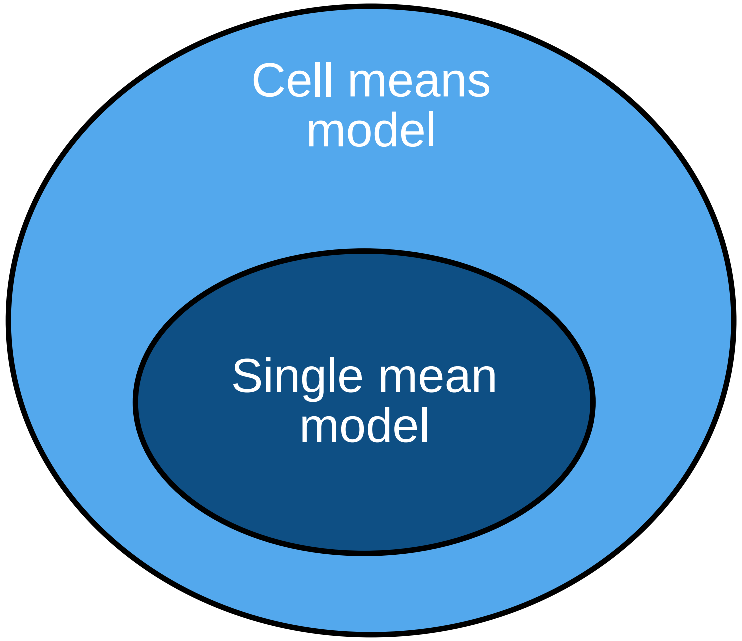
- Model reduces to: $Y_{ij} \text{ i.i.d.} \sim N(\mu, \sigma^2)$

- Or equivalently: $Y_{ij} = \mu + \epsilon_{ij}, \;\; \epsilon_{ij} \text{ i.i.d.} \sim N(0, \sigma^2)$.

- This is the so called **single mean** model.

Comparison of Models

- Note: Models are “nested”, the single mean model is a **special case** of the cell means model.
- Or: The cell means model is **more flexible** than the single mean model.
- Which one to choose? Let a **statistical test** decide.



Analysis of Variance (ANOVA)

- Classical approach: Decompose "**variability**" of response into different "**sources**" and **compare them.**
- More modern view: **Compare** (nested) **models** (model selection problem).
- In both approaches: Use statistical test with **global** null hypothesis

$$\boxed{H_0 : \mu_1 = \mu_2 = \cdots = \mu_g}$$

versus the alternative

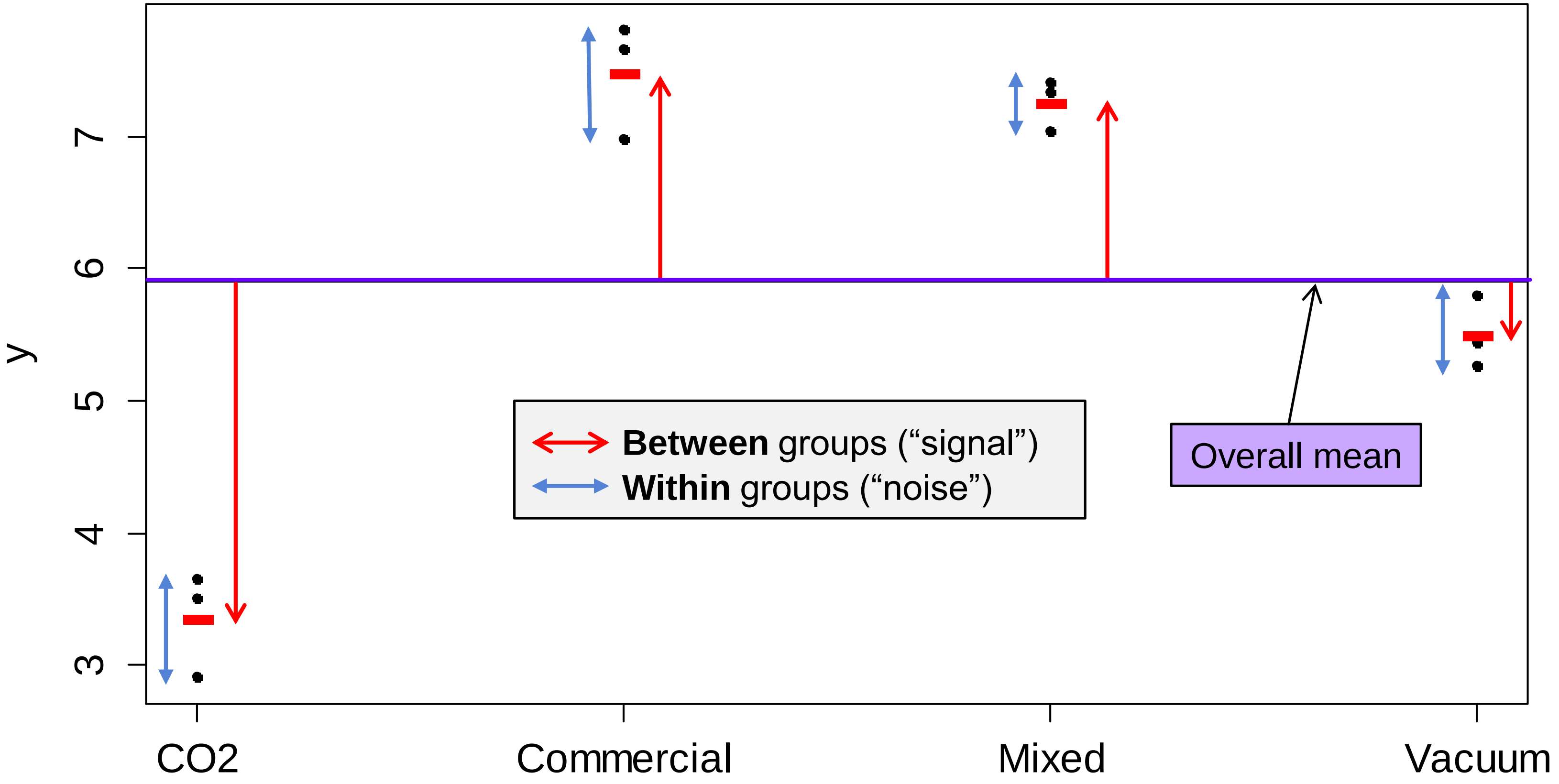
$$\boxed{H_A : \mu_k \neq \mu_l \text{ for } \textbf{at least one pair } k \neq l}$$

- H_0 says that the single mean model is sufficient to model the data.
- H_0 is equivalent to $\alpha_1 = \alpha_2 = \ ... = \alpha_g = 0$.

Decomposition of Total Variability

- See blackboard.

Illustration of Different Sources of Variability



ANOVA Table

- Typically, different sources of variation are presented in a so called **ANOVA table**:

Source	*df*	*Sum of squares (SS)*	*Mean Squares (MS)*	*F-ratio*
Treatments	$g-1$	SS_{Trt}	$MS_{Trt} = \frac{SS_{Trt}}{g-1}$	$\frac{MS_{Trt}}{MS_E}$
Error	$N-g$	SS_E	$MS_E = \frac{SS_E}{N-g}$	

- Use $\boldsymbol{F}$**-ratio** (last column) to construct a statistical test.
- **Idea**: Variation **between groups** should be **substantially** larger than variation **within groups** in order to reject H_0.
- This is a so called **one-way ANOVA**.

because only **one** factor involved

More Details about the F-Ratio

- It can be shown that $E[MS_{Trt}] = \sigma^2 + \sum_{i=1}^{g} n_i \alpha_i^2 / (g-1)$.
- Hence under H_0: MS_{Trt} is also an estimator for σ^2
  (contains **no "signal" just "error"**).
- Therefore, under H_0: $F = \frac{MS_{Trt}}{MS_E} \approx 1$.
- If we observe a value of F that is "**much larger**" **than 1**, we will **reject** H_0.
- What does "much larger" mean here?
- We need to be more precise: We need the **distribution** of F under H_0.

F-Distribution

- Under H_0 it holds that F follows a so called **$\boldsymbol{F}$-distribution** with $g-1$ and $N-g$ degrees of freedom: $F_{g-1,\,N-g}$.

- The **$\boldsymbol{F}$-distribution** has **two degrees of freedom parameters**: One from the numerator and one from the denominator mean square (treatment and error).

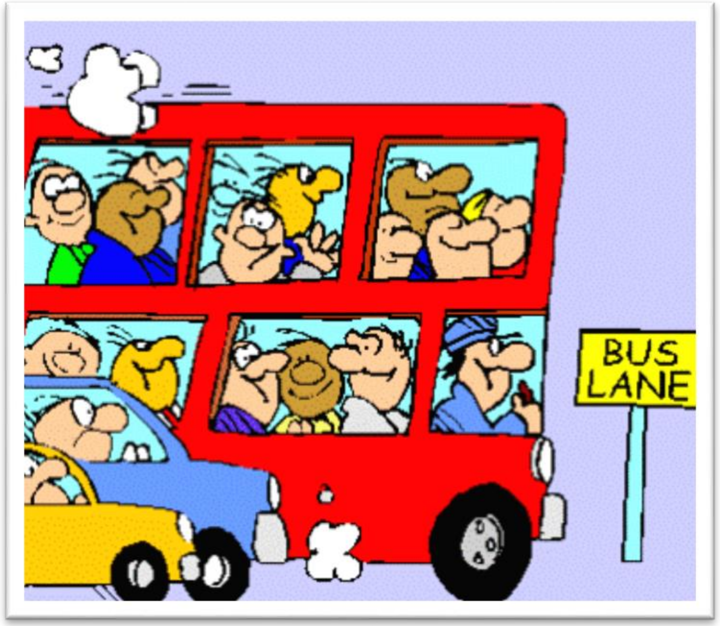
- Technically: $F_{n,\,m} = \frac{\frac{1}{n}(X_1^2+\cdots+X_n^2)}{\frac{1}{m}(Y_1^2+\cdots+Y_m^2)}$ where X_i, Y_j are i.i.d. $N(0,1)$.

- Illustration and behavior of quantiles: See R-Code.

- We reject H_0 if the corresponding **$\boldsymbol{p}$-value** is small enough or if F is larger than the corresponding quantile (the F-test is always a **one-sided** test).

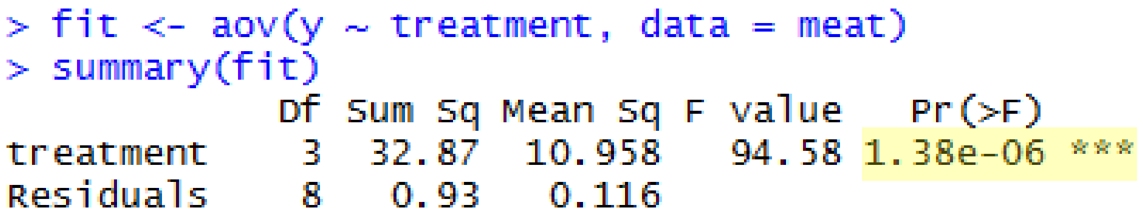
More on the F-Test

- It holds that $F_{1,n} = t_n^2$ (= the square of a t_n-distribution).
- It can be shown that the F-test for the $g = 2$ case is nothing else than the squared t-test.
- The F-test is also called an **omnibus test** (Latin for “for all”) as it compares **all group means simultaneously**.



Analysis of Meat Storage Data in R

- Use function `aov` to perform "**a**nalysis **o**f **v**ariance".
- When calling `summary` on the fitted object, an ANOVA table is printed out.

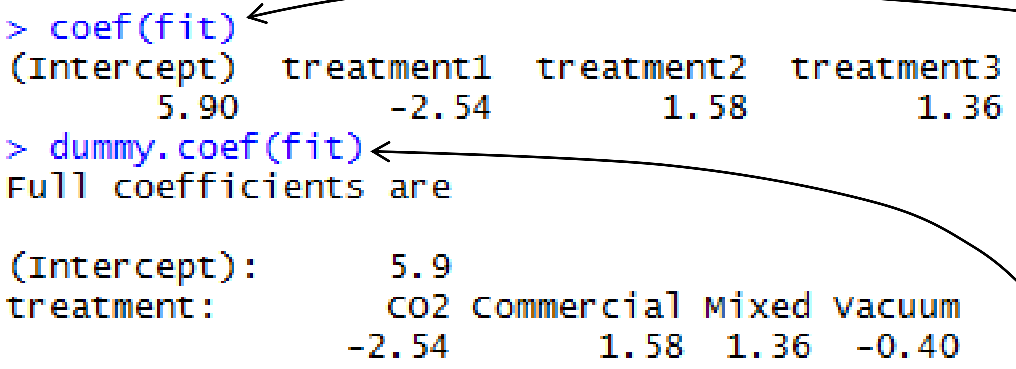


```
> fit <- aov(y ~ treatment, data = meat)
> summary(fit)
            Df Sum Sq Mean Sq F value   Pr(>F)    
treatment    3  32.87  10.958   94.58 1.38e-06 ***
Residuals    8   0.93   0.116                     
```

Reject H_0 because p-value is very small

Analysis of Meat Storage Data in R

- **Coefficients** can be extracted using the function `coef` or `dummy.coef`.



```
> coef(fit)
(Intercept)  treatment1  treatment2  treatment3
       5.90       -2.54        1.58        1.36
> dummy.coef(fit)
Full coefficients are

(Intercept):       5.9
treatment:        CO2 Commercial Mixed Vacuum
                -2.54       1.58  1.36  -0.40
```

Useless if encoding scheme unknown. Interpretation for computer trivial. For you?

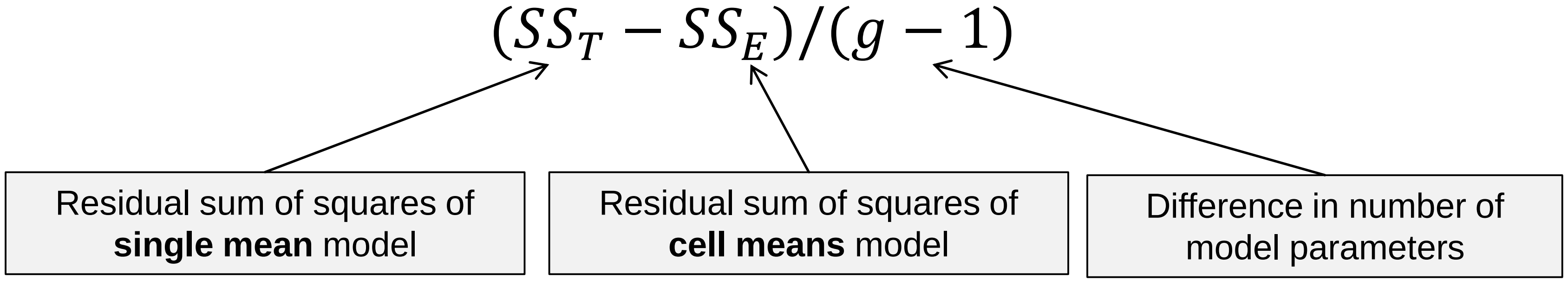
Coefficients in terms of the **original** levels of the factor rather than the “coded” variables.

$$\begin{aligned}
\mu_{\mathrm{CO_2}} &= 5.9 - 2.54 = 3.36 \\
\mu_{\mathrm{Commercial}} &= 5.9 + 1.58 = 7.48 \\
\mu_{\mathrm{Mixed}} &= 5.9 + 1.36 = 7.26 \\
\mu_{\mathrm{Vacuum}} &= 5.9 - 0.40 = 5.50
\end{aligned}$$

- Compare with fitted values (see R-Code).

ANOVA as Model Comparison

- Because $SS_T = SS_{Trt} + SS_E$, we can rewrite the numerator of the F-ratio as



$$(SS_T - SS_E)/(g-1)$$

Residual sum of squares of **single mean** model

Residual sum of squares of **cell means** model

Difference in number of model parameters

- Or in other words, SS_{Trt} is the **reduction in residual sum of squares** when going from the single mean to the cell means model.

- If we reject the F-test, we conclude that we really need the more complex cell means model, hence the group means are different.

Checking Model Assumptions

- Statistical inference (e.g., F-test) is only valid if the **model assumptions** are fulfilled.

- Need to check:
  - Are the errors **normally distributed**?
  - Are the errors **independent**?
  - Is the **error variance constant**?

- We don't observe the errors but we have the residuals as proxy.

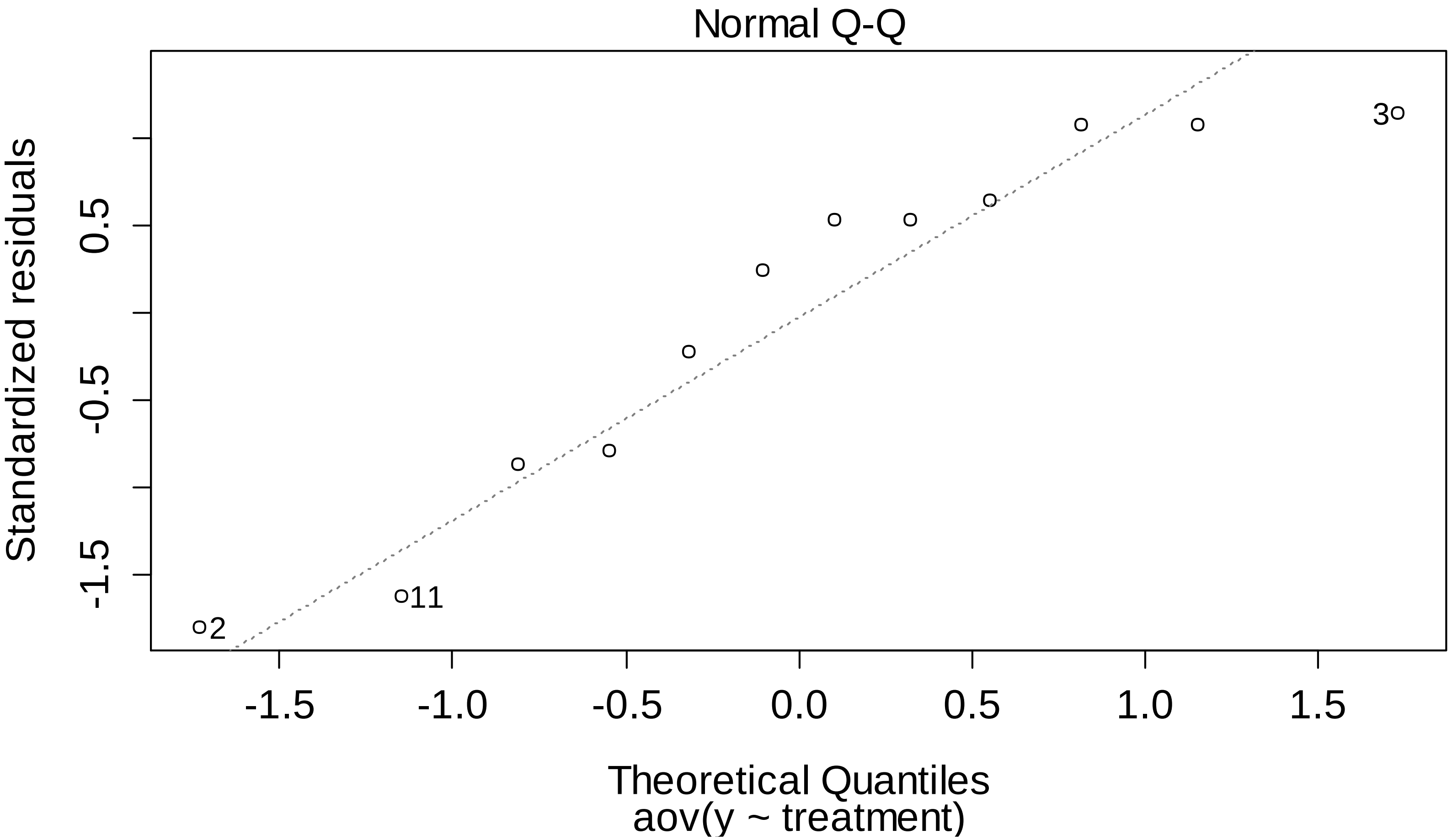
- Will use **graphical assessments** to **check assumptions**.
  - QQ-Plot
  - Tukey-Anscombe plot (TA plot)
  - Index plot
  - …

QQ-Plot (is normal distribution good approximation?)

- Plot **empirical quantiles of residuals** vs. **theoretical quantiles (of standard normal distribution)**.
- Points should lie more or less on a **straight line** if residuals are normally distributed.
- R: `plot(fit, which = 2)`
- If unsure, compare with (multiple) simulated versions from normal distribution with the same sample size

  ```
  qqnorm(rnorm(nrow(data)))
  ```
- **Outliers** can show up as isolated points in the “corners”.

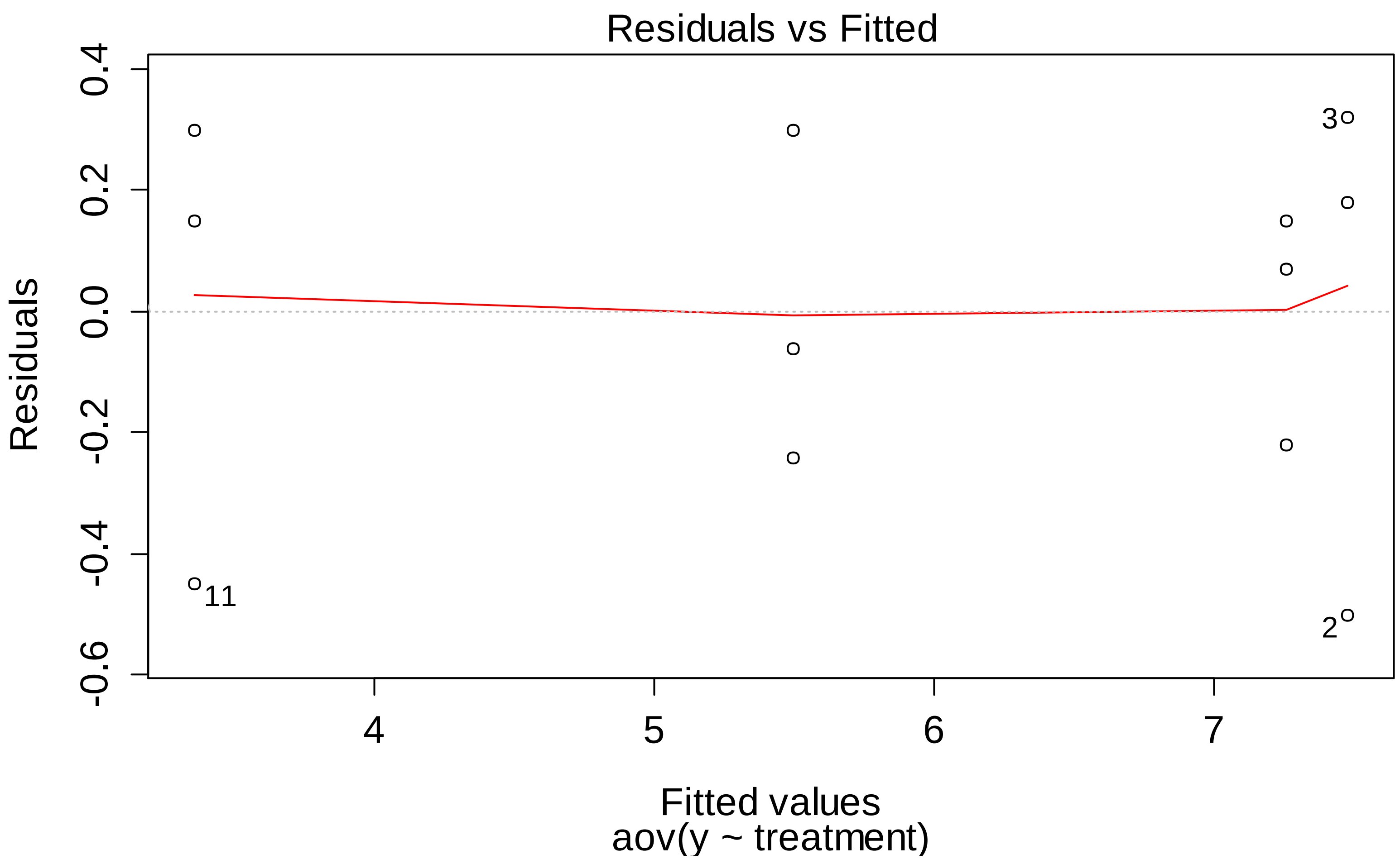
QQ-Plot (Meat Storage Data)



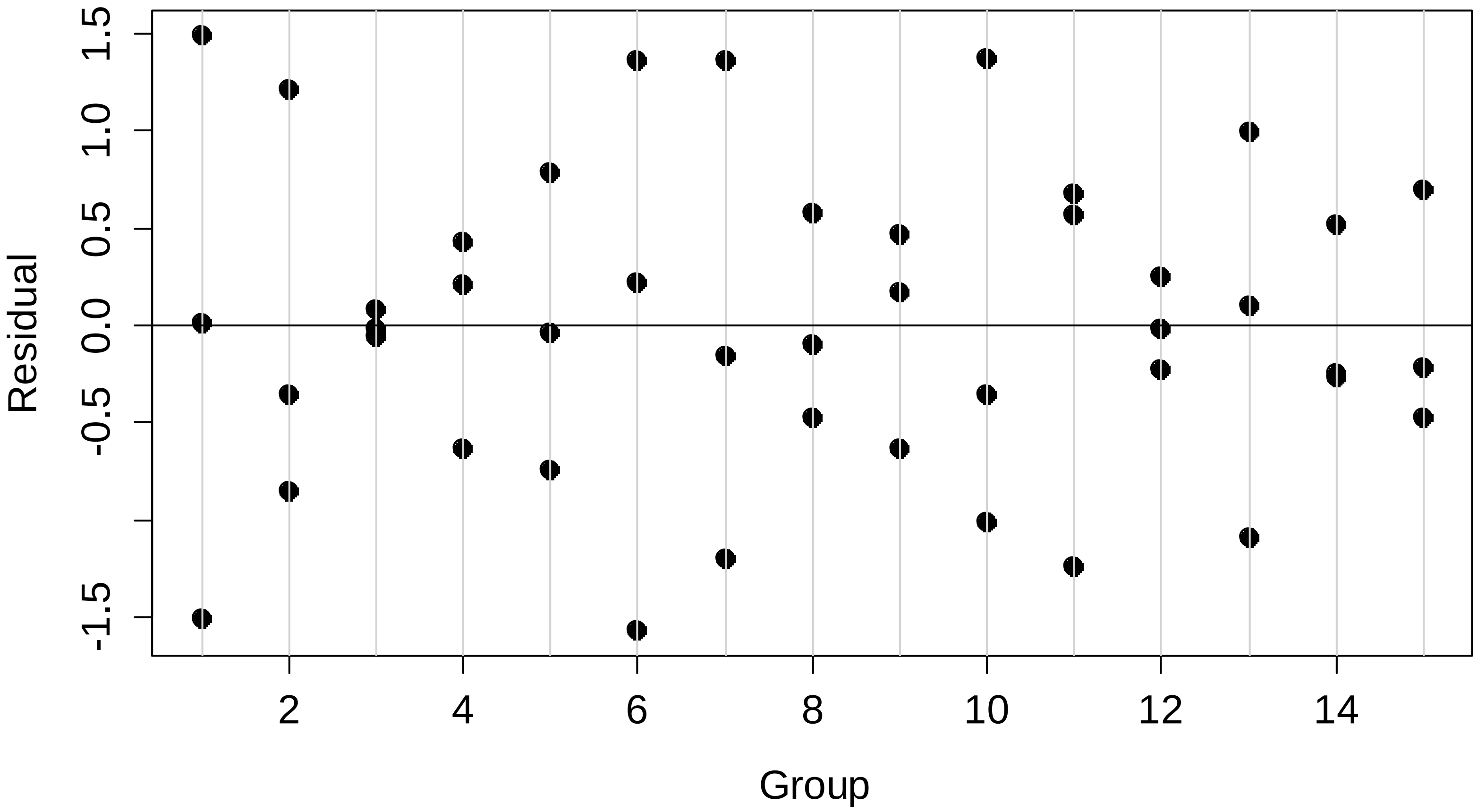
Tukey-Anscombe Plot (TA-Plot)

- Plot **residuals** vs. **fitted values**.
- Checks **homogeneity of variance** and **systematic bias** (here not relevant yet, why?)
- R: `plot(fit, which = 1)`
- “Stripes” are due to the data structure (g different groups).

Tukey-Anscombe Plot (Meat Storage Data)



Constant Variance?



Index Plot

- Plot residuals against **time** index to check for potential serial correlation (i.e., dependence with respect to time).
- Check if residuals close in time are too similar / dissimilar?
- Similarly for potential **spatial** dependence.

Fixing Problems

- **Transformation of response** (square root, logarithm, …) to improve QQ-Plot and constant variance assumption.
- Carefully **inspect potential outliers**. These are very interesting and informative data points.
- Deviation from normality less problematic for large sample sizes (reason: central limit theorem).
- **Extend model** (e.g., allow for some dependency structure, different variances, etc.)
- Many more options…
- More details: Exercises and Oehlert (2000), Chapter 6.