Completely Randomized Designs (CRD)
One-Way ANOVA
One-way ANOVA

Experimental units
  homogeneous
  inhomogeneous
  one-way ANOVA

Block Designs
  one block f., two (more)
  block size
    large
    small
  (B)IBD
  Latin Squares
  Youden Squares

CRD
  fixed effects, global test, contrasts, ...

RCB
  random effects, variance components, ...

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

RCB with factorial treatment structure, ...

Random effects, mixed effects models, nested factor structure, ...

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

Experimental units
  homogeneous
  inhomogeneous

One treatment factor

Multiple treatment factors

Youden

Similar to Lawson (2015)
Example: Meat Storage Study (Kuehl, 2000, Example 2.1)

- 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
  - 1% CO, 40% O₂, 59% N
  - 100% CO₂

- **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, $i = 1, \ldots, g$ (of course: $n_1 + n_2 + \ldots + n_g = N$).
- 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 $n_1, \ldots, n_g$ 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 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**: 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. $\sim N(\mu_X, \sigma^2), i = 1, \ldots, n$
  - $Y_j$ i.i.d. $\sim N(\mu_Y, \sigma^2), j = 1, \ldots, m$
  - $X_i, Y_j$ 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 **test** or 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 $j$th observation in the $i$th treatment group, $i = 1, \ldots, g; j = 1, \ldots, n_i$.
- **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 **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/](https://gallery.shinyapps.io/anova_shiny_rstudio/)
- Why **cell means**? Have a look at meat storage data:

<table>
<thead>
<tr>
<th>Commercial</th>
<th>Vacuum</th>
<th>Mixed</th>
<th>CO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.66</td>
<td>5.26</td>
<td>7.41</td>
<td>3.51</td>
</tr>
<tr>
<td>6.98</td>
<td>5.44</td>
<td>7.33</td>
<td>2.91</td>
</tr>
<tr>
<td>7.80</td>
<td>5.80</td>
<td>7.04</td>
<td>3.66</td>
</tr>
</tbody>
</table>
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 $\epsilon_{ij}$ i.i.d. $\sim N(0, \sigma^2)$.

- The $\epsilon_{ij}$’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 $\mu_i = \mu + \alpha_i, i = 1, \ldots, g$.

- E.g., think of $\mu$ as a “global mean” and $\alpha_i$ as the corresponding deviation from the global mean.

- $\alpha_i$ is also called the $i$th treatment effect.

- This looks like a needless complication now, but will be very useful later (with factorial treatment structure).

- Unfortunately this model is not identifiable anymore.

- Reason: $g + 1$ parameters ($\mu, \alpha_1, \ldots, \alpha_g$) for $g$ different means ($\mu_1, \ldots, \mu_g$)
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 treatments effect are allowed to vary freely. We also say that the treatment effect has \( g - 1 \) degrees of freedom (df).
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: $\mu, \alpha_1, \ldots, \alpha_{g-1}$ to parametrize the $g$ group means $\mu_1, \ldots, \mu_g$.

The interpretation of the parameters $\mu, \alpha_1, \ldots, \alpha_{g-1}$ 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 model fits the data “well”.

- Criterion: Choose parameter estimates such that

$$
\sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\alpha}_i)^2
$$

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

- The 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

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_i \cdot )</td>
<td>Sum of all values in group ( i )</td>
<td>( y_i \cdot = \sum_{j=1}^{n_i} y_{ij} )</td>
</tr>
<tr>
<td>( \bar{y}_i \cdot )</td>
<td>Sample average in group ( i )</td>
<td>( \bar{y}<em>i \cdot = \frac{1}{n_i} \sum</em>{j=1}^{n_i} y_{ij} = \frac{1}{n_i} y_i \cdot )</td>
</tr>
<tr>
<td>( y_. )</td>
<td>Sum of all observations</td>
<td>( y_. = \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} )</td>
</tr>
<tr>
<td>( \bar{y}. )</td>
<td>Grand mean</td>
<td>( \bar{y}. = \frac{y_.}{N} )</td>
</tr>
</tbody>
</table>

Rule: If we replace an index with a dot (“\( \cdot \)”) it means that we are summing up values over that index.
Parameter Estimates, the Other Way Round

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

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \bar{y}_i - \bar{y}.$$ 

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

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

Estimator for $\sigma^2$ is $MS_E$, **mean squared error**, i.e.

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

This is an **unbiased estimator** for $\sigma^2$ (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)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimator</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$\bar{y}.\cdot$</td>
<td>$\sigma/\sqrt{N}$</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>$\bar{y}_i.\cdot$</td>
<td>$\sigma/\sqrt{n_i}$</td>
</tr>
<tr>
<td>$\alpha_i$</td>
<td>$\bar{y}_i. - \bar{y}.\cdot$</td>
<td>$\sigma \sqrt{1/n_i - \frac{1}{N}}$</td>
</tr>
<tr>
<td>$\mu_i - \mu_j = \alpha_i - \alpha_j$</td>
<td>$\bar{y}_i. - \bar{y}_j.\cdot$</td>
<td>$\sigma \sqrt{1/n_i + 1/n_j}$</td>
</tr>
</tbody>
</table>

Therefore, a 95% confidence interval for $\alpha_i$ 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 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 treatment has **no effect** on the response.

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

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

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

- This is the so called **single mean** model.
Comparison of Models

- Note: Models are “nested”, single mean model is a special case of cell means model.
- Or: Cell means model is more flexible than 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

\[ H_0: \mu_1 = \mu_2 = \ldots = \mu_g \]

versus the alternative

\[ H_A: \mu_k \neq \mu_l \text{ for at least one pair } k \neq l \]

- \( H_0 \) says that the single mean model is sufficient to model the data.
- \( H_0 \) is equivalent to \( \alpha_1 = \alpha_2 = \ldots = \alpha_g = 0 \).
Decomposition of Total Variability

- See blackboard.
Illustration of Different Sources of Variability

- **Between groups** ("signal")
- **Within groups** ("noise")

Grand mean
ANOVA Table

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

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of squares (SS)</th>
<th>Mean Squares (MS)</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>$g - 1$</td>
<td>$SS_{Trt}$</td>
<td>$MS_{Trt} = \frac{SS_{Trt}}{g-1}$</td>
<td>$MS_{Trt}$/$MS_E$</td>
</tr>
<tr>
<td>Error</td>
<td>$N - g$</td>
<td>$SS_E$</td>
<td>$MS_E = \frac{SS_E}{N - g}$</td>
<td></td>
</tr>
</tbody>
</table>

- Use **$F$-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 $\sigma^2$ (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 **$F$-distribution** with $g - 1$ and $N - g$ degrees of freedom: $F_{g-1, N-g}$.

- The **$F$-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 **$p$-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 “analysis of variance”
- When calling `summary` on the fitted object, an ANOVA table is printed out.

```r
> fit <- aov(y ~ treatment, data = meat)
> summary(fit)
                Df Sum Sq Mean Sq F value    Pr(>F)
 treatment     3 32.875  10.958  94.588 1.38e-06 ***
 Residuals     8  0.930  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`.

```r
> 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 coefficients rather than the “coded” variables.

\[
\begin{align*}
\mu_{CO2} & = 5.9 - 2.54 = 3.36 \\
\mu_{Commercial} & = 5.9 + 1.58 = 7.48 \\
\mu_{Mixed} & = 5.9 + 1.36 = 7.26 \\
\mu_{Vacuum} & = 5.9 - 0.40 = 5.50
\end{align*}
\]

- 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

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

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

---

**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

```r
qqnorm(rnorm(nrow(data)))
```

**Outliers** can show up as isolated points in the “corners”.
QQ-Plot (Meat Storage Data)

Theoretical Quantiles
Standardized residuals
aov(y ~ treatment)
Normal Q-Q

2
11
3
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 results close in time 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, …)

- Many more options…