Faculty of Health Sciences

Day 4: Analysis of Variance

Paul Blanche

Section of Biostatistics, University of Copenhagen

February 25, 2026

Outline/Intended Learning Outcomes (ILOs)

Simple pairwise comparisons
- ILO: to perform pairwise comparisons and draw rational conclusions

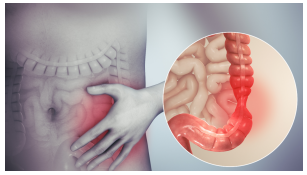
Analysis of Variance (ANOVA): one-way
- ILO: to describe the model, its parameters and assumptions
- ILO: to explain why all assumptions are not all equally important
- ILO: to interpret standard results

Analysis of Variance (ANOVA): two-way
- ILO: to contrast one- and two-way ANOVA
- ILO: to explain why all assumptions are not all equally important
- ILO: to interpret standard results

Case: Irritable Bowel Syndrome Dose Response

- Data from $n = 198$ women.
- Randmomized (double-blind) to:
  - Placebo ($n = 50$, "dose 0")
  - Dose 1 ($n = 54$)
  - Dose 2 ($n = 49$)
  - Dose 3 ($n = 45$)

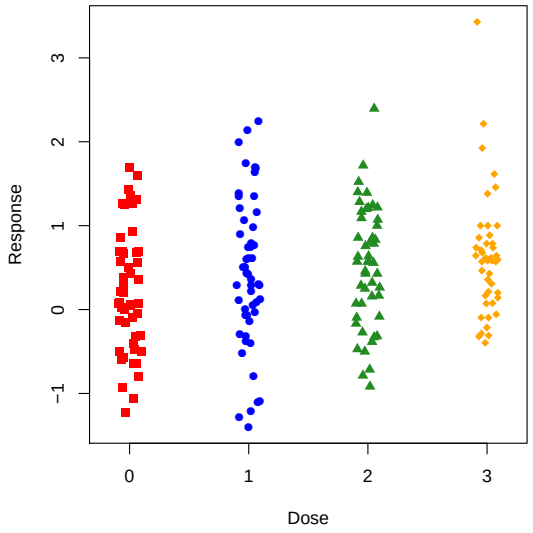
(Doses are blinded for confidentiality)



Outcome: baseline adjusted abdominal pain score at end of follow-up (12 weeks), approximately continuous variable, the larger the better.

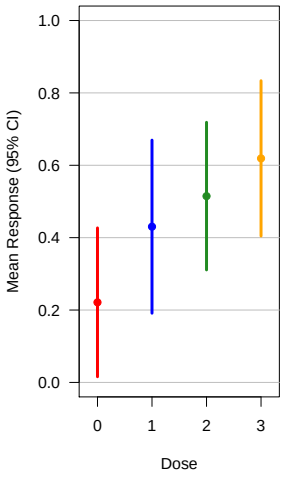
Research questions:
- Does the drug work?
- Are there differences between doses?

Outcome data



Dose	Mean	SD	n
0	0.22	0.72	50
1	0.43	0.88	54
2	0.51	0.71	49
3	0.62	0.71	45

Pairwise Welch's t-test: results



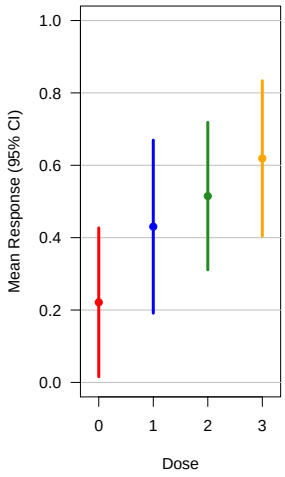
p-values from pairwise t-tests:

Dose	0	1	2
1	0.19		
2	**0.04**	0.59	
3	**0.01**	0.24	0.48

Note: the y-axis has changed!

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Pairwise Welch's t-test: results



p-values from pairwise t-tests:

Dose	0	1	2
1	0.19		
2	**0.04**	0.59	
3	**0.01**	0.24	0.48

Have we not reported all relevant results? What is left to worry about?

Note: the y-axis has changed!

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Interpretations (1/3)

Results include:

- Dose 0 not significantly different from dose 1.
- Dose 1 not significantly different from dose 2.
- But dose 0 significantly different from dose 2.

Are the results self-contradicting?

[1] Altman & Bland (1995). BMJ, 311(7003), 485.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Interpretations (1/3)

Results include:

- Dose 0 not significantly different from dose 1.
- Dose 1 not significantly different from dose 2.
- But dose 0 significantly different from dose 2.

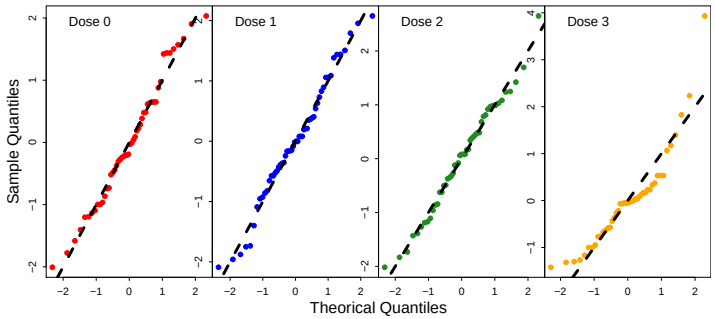
Are the results self-contradicting?

- No! This is just due to statistical uncertainty because of "small" sample sizes.
- Again, *"Absence of evidence is not evidence of absence"* [1]
- Dose 1 may have a similar effect to either dose 0 or dose 2.

[1] Altman & Bland (1995). BMJ, 311(7003), 485.

Interpretations (2/3)

What about the **assumptions**? Can we trust the results?



- QQplots for doses 0, 1, 2 look good but not so good for dose 3.
- However nothing "very" bad and decent sample size ($\geq$ 45 per group), so it seems fine.

Interpretations (3/3)

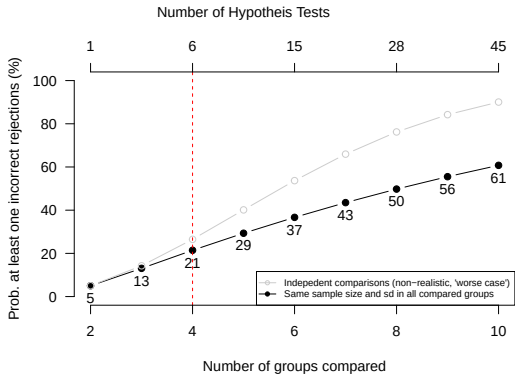
Beware of the **multiple testing** issue!
We are trying to answer the **same question** *"Are there differences between doses?"* **using six hypothesis tests**. If we find a difference between any two doses, we want to conclude that "there are differences between doses". The reasoning is good, except for the fact that the risk of making at least one false conclusion that two different doses lead to different mean outcome is $\geq$ 5%. Hence we do not control the risk of falsely concluding that "there are differences between doses" at 5%.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Interpretations (3/3)

Beware of the **multiple testing** issue!
We are trying to answer the **same question** *"Are there differences between doses?"* **using six hypothesis tests**. If we find a difference between any two doses, we want to conclude that "there are differences between doses". The reasoning is good, except for the fact that the risk of making at least one false conclusion that two different doses lead to different mean outcome is $\geq$ 5%. Hence we do not control the risk of falsely concluding that "there are differences between doses" at 5%.

How bad can this be?



What statistical method should we use?

We want to control the FWER[2] at 5%.

Using Bonferroni? No.
It's a conservative (i.e. sub-optimal) approach which ignores the (strong) correlation between the comparisons (as may comparisons use data from the same groups, e.g. 0 vs 1 and 0 vs 2 both use data for dose 0)

[2] Family-wise error rate (FWER): probability of making one or more false discoveries when performing multiple hypotheses tests (Lecture 2).
[3] Although, strictly speaking, they rely on "minor" large sample size approximations.
[4] That is why the method are still "new" and underused.

What statistical method should we use?

We want to control the FWER[2] at 5%.

Using Bonferroni? No.
It's a conservative (i.e. sub-optimal) approach which ignores the (strong) correlation between the comparisons (as may comparisons use data from the same groups, e.g. 0 vs 1 and 0 vs 2 both use data for dose 0)

More modern alternative? Yes.
Use specific method and software for multiple correction that do not make any additional assumptions.[3] The details of the method and computation are more complicated[4] but not the interpretation and user-friendly software exist.

[2] Family-wise error rate (FWER): probability of making one or more false discoveries when performing multiple hypotheses tests (Lecture 2).
[3] Although, strictly speaking, they rely on "minor" large sample size approximations.
[4] That is why the method are still "new" and underused.

Recommended analysis (see R-demo for code)

Statistical methods:
Comparisons between groups were made with a heteroscedastic ANOVA model (not assuming equal variances). P-values and 95% confidence intervals were adjusted for multiple testing using the max-t test method (aka min-p method) as implemented in the multcomp-package [ref.[5]] of the statistical software R [ref.[6]] and described in [ref.[7]].

Results (adjusted for multiple testing):

Comparison	Est. Diff	95% CI	p-value
1 - 0	0.209	[-0.202; 0.620]	0.552
2 - 0	0.293	[-0.083; 0.670]	0.185
3 - 0	0.398	[0.011; 0.784]	0.041
2 - 1	0.084	[-0.325; 0.494]	0.951
3 - 1	0.189	[-0.230; 0.607]	0.647
3 - 2	0.104	[-0.281; 0.489]	0.896

Note: p-values $\leq$ 6 times the non-adjusted ones (Bonferroni).

[5] Hothorn, Bretz & Westfall (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal 50(3), 346–363.
[6] R Core Team (2026). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[7] Herberich, Sikorski & Hothorn. "A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs." PloS one 5.3 (2010): e9788.

Why the "min-p" name? Which interpretation?

- To answer *"Yes!"* to the research question *"Are there differences between doses?"*, it is sufficient to show that a difference exists between any two of the doses. Hence one significant adjusted p-value is enough evidence and the **minimum** of these adjusted p-value can be used to conclude.[8]

[8] Obviously, we have at least one significant p-value if the minimum of all the adjusted p-values is significant.

Why the "min-p" name? Which interpretation?

- To answer *"Yes!"* to the research question *"Are there differences between doses?"*, it is sufficient to show that a difference exists between any two of the doses. Hence one significant adjusted p-value is enough evidence and the **minimum** of these adjusted p-value can be used to conclude.[8]
- This means that the p-value corresponding to the global null hypothesis

$$H_0 : \textit{"the mean response is the same for all doses"}$$

  can simply to be computed as the minimum of the adjusted p-values computed for all the pairwise comparisons.
- Note: "minimum p-value" is equivalent to "maximum t-statistic". Hence the two equivalent names, "min-p" or "max-t test", for the method.

Conclusion for our case: we found a statistically significant association between the dose of medication and the mean response (p=0.041).

[8] Obviously, we have at least one significant p-value if the minimum of all the adjusted p-values is significant.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Take home message/generic example (beyond doses example)

- This modern min-P method can be used to compute a p-value corresponding to the global null hypothesis

  $$H_0 : \text{"the mean response is the same for all groups"}$$

  no matter the type of categorical variables that defines the groups (e.g., blood cancer type). You can then write a conclusion sentence such as:

  "A significant association was found between the [variable defining the group, e.g. cancer type] and the mean [response, e.g., antibiotics treatment days] (p=[min-P found, e.g. 0.01])."

- You can then complement[9] this result by presenting the result for the comparison of any two groups, e.g.:

  "Especially, the mean difference in [response, e.g., antibiotics treatment days] between [this group, e.g. AML] and [this group, e.g. ALL] was [mean diff, 95-CI and p found, e.g., 4.5 days (95%-CI 2.4 to 6.6, p=0.005)]".

12 / 59 [9] unlike with commonly used alternative statistical methods, there will never be inconsistencies between "overall" results and the results of all pairwise comparisons; see next slides.

What if we only want the comparisons to placebo?

Instead of the research question *"Are there differences between doses?"*, we sometimes want to answer *"Does the drug work/do something?"*, which in other words corresponds to, *"Is there a dose for which the average outcome is different, compared to placebo?"*

Hence we only want all the comparisons to one (reference) group.

This is known as the **many-to-one** comparisons case (Dunnett), by contrast to the **all pairwise** comparisons case (Tukey).[10]

The method for this case is similar and we can use the **same software**.

Case results (adjusted for multiple comparisons to placebo):

Comparison	Est. Diff	95% CI	p-value
1 - 0	0.209	[-0.168; 0.586]	0.418
2 - 0	0.293	[-0.052; 0.639]	0.115
3 - 0	0.398	[0.043; 0.752]	0.023

Note: p-values $\leq$ 3 times the non-adjusted ones (Bonferroni).

13 / 59 [10] Dunnett & Tukey were among the first statisticians who proposed specific, powerful, methods for these specific cases.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Pre-specification matters

The same comparison, e.g. Dose 3 versus Placebo (Dose 0), leads to the estimated mean difference 0.398, but different 95% confidence intervals (CI) and p-values (after adjusting for multiple testing) when we consider either:

- All-pairwise (6) comparisons: 95% CI=$[0.011; 0.784]$, p= 0.041.
- Many-to-one (3) comparisons: 95% CI=$[0.043; 0.752]$, p= 0.023.

Take home messages:

- The more comparisons the wider the 95% CI and the higher the p-values.
- Do not investigate more comparisons than interesting/possible (power $\downarrow$).
- The choice of investigating "all-pairwise" versus "many-to-one" comparisons should be done before seeing the data, i.e. pre-specified. Rigorously adjusting for multiple testing is not possible otherwise and how much we can trust the results without pre-specification is most unclear.

14 / 59

Digression: prespecified vs post hoc analyses

It is completely fine and often useful to performed post hoc[11] analyses as long as:

- they are reported as such in publications[12],

  "Distinguish prespecified from exploratory analyse"[13]

- conclusions based on them are not too strong.

"The main analyses should concentrate on the primary research questions to reduce the amount of testing of data-generated hypotheses. However, science would not proceed if analyses of questions not stated in the protocol were not allowed so, obviously, new ideas generated from the data can be pursued as long as conclusions based on such additional analyses are suitably calibrated."[14]

[11] A post hoc analysis is an analysis specified after the data were seen.

[12] Otherwise this is "data fishing", "data snooping" or "p-hacking" and this is considered as something in between "questionable research practice" and "scientific dishonesty and research misconduct"; see KU course "Responsible Conduct of Research".

[13] "Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals, Updated January 2025", www.icmje.org/icmje-recommendations.pdf

15 / 59 [14] Andersen & Skovgaard, *Regression with linear predictors* , page 473 (Springer, 2010).

Appendix: power and sample size calculation

When planning several comparisons, say K, with a FWER control at α , one can:

1. Define an adjusted type-I error $\alpha' = \alpha/K$.
2. Perform sample size and power calculation for each comparison as in the case of a unique comparison, using this adjusted type-I error α' as input of the formula instead of α.

Note: this is a "slightly" conservative approach[15].

[15] This is because this simple calculation assumes that the (less powerful) Bonferroni correction will be used.

Outline/Intended Learning Outcomes (ILOs)

Simple pairwise comparisons
- ILO: to perform pairwise comparisons and draw rational conclusions

Analysis of Variance (ANOVA): one-way
- ILO: to describe the model, its parameters and assumptions
- ILO: to explain why all assumptions are not all equally important
- ILO: to interpret standard results

Analysis of Variance (ANOVA): two-way
- ILO: to contrast one- and two-way ANOVA
- ILO: to explain why all assumptions are not all equally important
- ILO: to interpret standard results

What is ANOVA about?

ANOVA stands for "**AN**alysis **O**f **VA**riance", but this is a useful method to compare means (via the comparisons of variances.).

Useful for answering research questions such as:

- Is this continuous outcome associated with this categorical variable?
- Is the mean outcome the same for all levels of this categorical outcome?

Examples:

- **Outcome:** weight, blood pressure, concentration, pain score ...
- **Categorical variable:** BMI group, age group, dose level ...

This is a very commonly used, well-known and "old" method.

ANOVA model (one-way)

The j-th observation from the i-th group is described as:

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

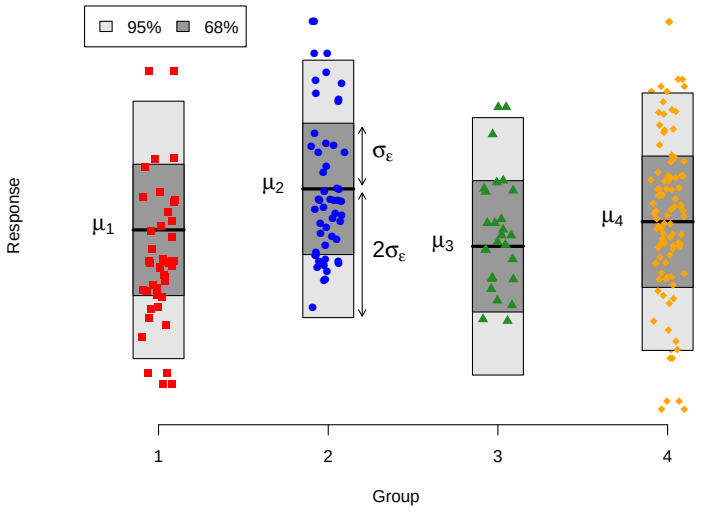
- μ_i is the (population) mean for group i.
- ε_{ij}'s are individual 'error' terms ("random/unexplained deviation from the mean") assumed normally distributed with zero mean and the same variance σ^2_ε regardless of group.[16]

Model assumptions (1-2 important, 3-4 not always):

1. Observations from different groups are independent.
2. Individual observations within each group are independent.
3. 'Error' terms are normally distributed.
4. The variance of 'error' terms is the same for all groups (homogeneity).

[16] ε_{ij} refects *"general differences between subjects [...] as well as sometimes differences from occasion to occasion within subjects, since, although we are only assuming that we measure a given subject once, this is often only one of many occasions on which we might have measured the subject"* (Senn, Statist. Med. 2004; 23:3729–3753).

Visual interpretation ($Y_{ij} = \mu_i + \varepsilon_{ij}$, hypothetical data)



- σ_ε tells us how **vertically spread** are the points above and below each group mean μ_i, for each group.

ANOVA (one-way): why and how?

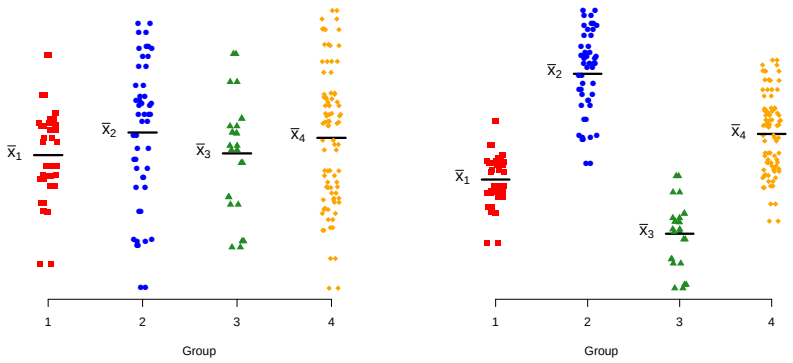
Why using a "traditional" ANOVA analysis, aka an "F-test"?
To test the global null hypothesis "H_0: The mean of all (K) groups are all equal", that is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K.$$

How does it work?
By using a F-test which compares the between-group variability to the within-group variability. If the between-group variability is large enough relative to the within-group variability, then we reject H_0.

- Hence the name ANOVA: we analyze variances
- Computation possible by hand, hence the method became popular during the pre-computer age.

ANOVA: intuition of the F-test (hypothetical data)

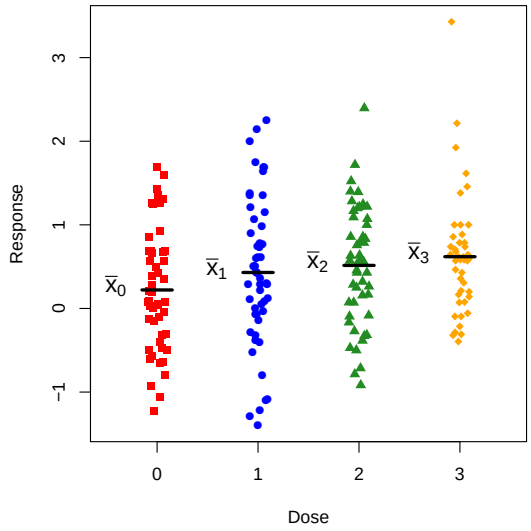


- **Left:** the between-group variance (i.e. the variance of sample means $\bar{x}_i$) is **small** relative to the within-group variability: **do not reject** H_0.
- **right:** the between-group variance is **large** relative to the within-group variability: **reject** H_0.

Note: of course "small"/"large" is also relative to the sample size.

Case: "traditional" ANOVA analysis, aka F-test

- $\bar{x}_0 = 0.22$
- $\bar{x}_1 = 0.43$
- $\bar{x}_2 = 0.51$
- $\bar{x}_3 = 0.62$
- $\hat{\sigma}_\varepsilon = 0.76$
- p-value=0.07
- **Do not reject!**

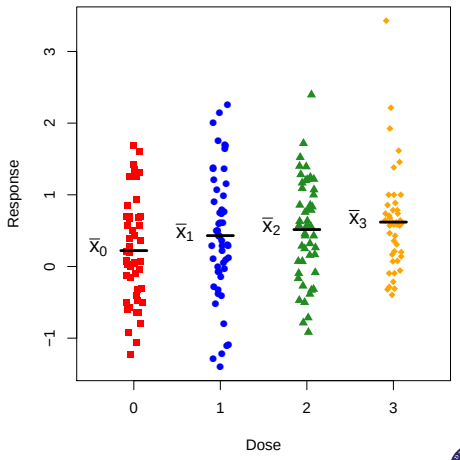


This is **not consistent** with the results from the all-pairwise comparisons (seen in the first slides)...

Case: similar F-test, but without assuming homogeneity

Let's use a more flexible model, which does not require assumption 4 (homogeneity).

- $\bar{x}_1 = 0.22$, $\hat{\sigma}_1 = 0.72$
- $\bar{x}_2 = 0.43$, $\hat{\sigma}_2 = 0.88$
- $\bar{x}_3 = 0.51$, $\hat{\sigma}_3 = 0.71$
- $\bar{x}_4 = 0.62$, $\hat{\sigma}_4 = 0.71$
- p-value=0.055
- Do not reject!



Still **not consistent** with the results from the all-pairwise comparisons.. although the modeling assumptions are similar ! (only the hypothesis test method is different)

Critics of the F-test and recommendations (1/2)

- When the F-test is significant we can conclude to differences between the groups but we do not know between which groups! (frustrating....). Historically, people used to proceed in two steps: 1) test the global null $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$, 2) if H_0 is rejected, proceed to make pairwise comparisons, but why not directly start with pairwise comparisons, especially because...
- F-test and pairwise comparisons are inconsistent: either may find a significant difference the other doesn't. When it happens it is frustrating and hard to explain.
- When the F-test is not significant it is difficult to know whether it is due to lack of effect or lack of evidence, because there are no corresponding confidence intervals.

Critics of the F-test and recommendations (2/2)

Recommendations:

- Unless you are not interested in the pairwise comparisons or have another specific reason in mind (e.g. power), prefer the all-pairwise comparisons and min-p (aka max-t test) approach to the F-test.
- If you are not interested in the pairwise comparisons but **only** in knowing whether a continuous outcome is associated with a categorical, and if you want to keep the analysis as "simple and common" as possible, then you may prefer the F-test to the more "modern" min-p approach.

Appendix: Power of F-test vs all-pairwise comparisons

We can test the global null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ using:

- F-test.
- min-p test (aka max-t test): perform all-pairwise comparisons and compute the p-value for H_0 as the minimum of the multiplicity adjusted p-values of all the comparisons (hence the "min-p" name). The rational is that we can safely reject H_0 if there is at least one significant difference after adjusting for multiple testing.

Which approach is the most powerful?

- F-test when the means of all groups are different although there is no particularly large difference between any two groups.
- "min-p test" when there exists a particularly large difference between two of the groups.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

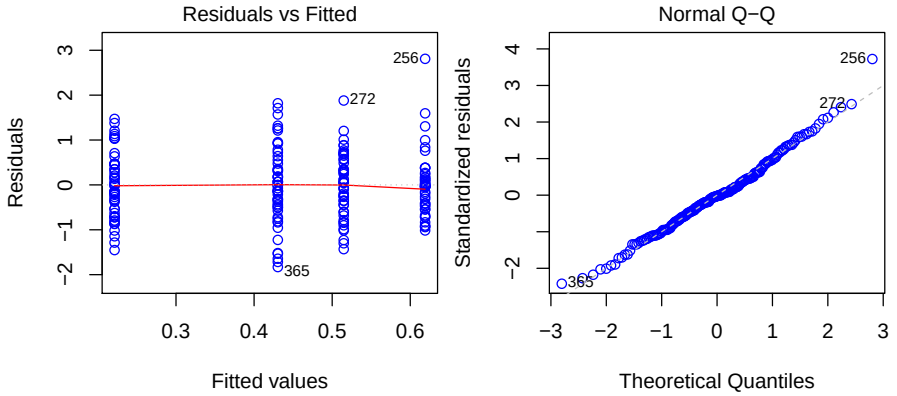
Checking the ANOVA model assumptions

The F-test and ANOVA model are still very much used. What should we know about checking their modeling assumptions?

- Assumptions 1-2 (independence):
  - rely on the study design.
- Assumptions 3 (homogeneity of variances):
  - check with residual plots or compute sd in each group (best).
  - can be relaxed if needed (see R-demo for code).
  - log-transforming the data might help to obtain homogeneity of the variances. [17]
- Assumptions 4 (normality):
  - check with qqplot.
  - not needed with large sample sizes in each group.[18]

[17] It should ideally be prespecified. If not, this is a posthoc analysis.
[18] As for the t-test, for the same reason: the central limit theorem.

Case: model checking "default" plots



Notes:

- These plots are similar to those used to check the linear models (see Lecture 3).
- The 'default' left plot does not use "jitter", which substantially complicates the interpretation... Comparing the numerical values of the SD in each group can be more informative (and this is simple to do)

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

ANOVA: usual software parametrization (1/4)

R code for ANOVA

```
fitlm <- lm(resp~dosefact, data=d)
summary(fitlm)
```

which returns (among other things)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2213     0.1079   2.051   0.0416 *
dosefact1     0.2091     0.1498   1.396   0.1643
dosefact2     0.2935     0.1534   1.913   0.0572 .
dosefact3     0.3977     0.1568   2.537   0.0120 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7631 on 194 degrees of freedom
Multiple R-squared:  0.03513,Adjusted R-squared:  0.02021
F-statistic: 2.354 on 3 and 194 DF,  p-value: 0.07335
```


ANOVA: usual software parametrization (2/4)

- `(Intercept)`: est. mean in the reference group: $\bar{x}_0 = 0.2213$.
- `dosefact1`: est. mean difference: $\bar{x}_1 - \bar{x}_0 = 0.2091$.
- `dosefact2`: est. mean difference: $\bar{x}_2 - \bar{x}_0 = 0.2935$.
- `dosefact3`: est. mean difference: $\bar{x}_3 - \bar{x}_0 = 0.3977$.

- `F-statistic`: provides F-test p-value: 0.07335.
- `Residual standard error`: estimate of σ_ε: 0.7631

ANOVA: usual software parametrization (2/4)

- `(Intercept)`: est. mean in the reference group: $\bar{x}_0 = 0.2213$.
- `dosefact1`: est. mean difference: $\bar{x}_1 - \bar{x}_0 = 0.2091$.
- `dosefact2`: est. mean difference: $\bar{x}_2 - \bar{x}_0 = 0.2935$.
- `dosefact3`: est. mean difference: $\bar{x}_3 - \bar{x}_0 = 0.3977$.

- `F-statistic`: provides F-test p-value: 0.07335.
- `Residual standard error`: estimate of σ_ε: 0.7631

- p-values for the mean differences are not adjusted for multiple testing.

ANOVA: usual software parametrization (2/4)

- `(Intercept)`: est. mean in the reference group: $\bar{x}_0 = 0.2213$.
- `dosefact1`: est. mean difference: $\bar{x}_1 - \bar{x}_0 = 0.2091$.
- `dosefact2`: est. mean difference: $\bar{x}_2 - \bar{x}_0 = 0.2935$.
- `dosefact3`: est. mean difference: $\bar{x}_3 - \bar{x}_0 = 0.3977$.

- `F-statistic`: provides F-test p-value: 0.07335.
- `Residual standard error`: estimate of σ_ε: 0.7631

- p-values for the mean differences are not adjusted for multiple testing.

- "default" summary presents only comparisons between the reference group and others (3 out of 6 possible). This is arbitrary!
- Note that if **Dose 1** had been chosen as **the reference group**, among the 3 differences shown in the output none would be significant.

31 / 59

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

ANOVA: usual software parametrization (3/4)

R code for ANOVA when the reference Dose is now Dose 1.

```
d$dosefact <- relevel(d$dosefact,ref="1")
fitlm <- lm(resp~dosefact, data=d)
summary(fitlm)
```

which returns (among other things)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4304     0.1038   4.145 5.08e-05 ***
dosefact0    -0.2091     0.1498  -1.396    0.164
dosefact2     0.0844     0.1505   0.561    0.576
dosefact3     0.1887     0.1540   1.225    0.222
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7631 on 194 degrees of freedom
Multiple R-squared:  0.03513,Adjusted R-squared:  0.02021
F-statistic: 2.354 on 3 and 194 DF,  p-value: 0.07335
```

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

ANOVA: usual software parametrization (4/4)

The ANOVA model is actually a specific kind of linear model. [19] The mean of each group is described by the regression formula:

$$\mu_i = \alpha + \beta_1 \cdot \mathrm{I}(\text{group}_i\text{=Dose 1}) + \beta_2 \cdot \mathrm{I}(\text{group}_i\text{=Dose 2}) + \beta_3 \cdot \mathrm{I}(\text{group}_i\text{=Dose 3})$$

where I() is the indicator function:

$$\mathrm{I}(\text{condition}) = \begin{cases} 1, \text{ if condition is true} \\ 0, \text{ if condition is false} \end{cases}$$

Group	Dose 0	Dose 1	Dose 2	Dose 3
Mean	α	$\alpha + \beta_1$	$\alpha + \beta_2$	$\alpha + \beta_3$
Estimate	0.2213	0.2213 + 0.2091	0.2213 + 0.2935	0.2213 +0.3977

[19] more on linear model in Lecture 7; it explains the use of the `lm` function in R.

Outline/Intended Learning Outcomes (ILOs)

Simple pairwise comparisons
- ILO: to perform pairwise comparisons and draw rational conclusions

Analysis of Variance (ANOVA): one-way
- ILO: to describe the model, its parameters and assumptions
- ILO: to explain why all assumptions are not all equally important
- ILO: to interpret standard results

Analysis of Variance (ANOVA): two-way
- ILO: to contrast one- and two-way ANOVA
- ILO: to explain why all assumptions are not all equally important
- ILO: to interpret standard results

Two-way ANOVA

What is it about?
Analysis the mean of a continuous outcome depending on two categorical variables.

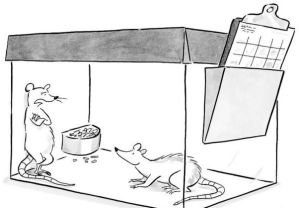
Why and when is it useful?
1. to increase power and precision of the estimates in randomized experiments.
2. to correct/adjust for differences between the groups that we primarily aim to compare in observational studies (e.g. to adjust for baseline differences, i.e., confounding; to get closer to "causal" conclusions [20]).
3. to (sometimes) better handle missing data.

Note: points 2 and 3 are closely related.[21]

[20] More on Lecture 7.
[21] Essentially, missing data are problematic when they make the the groups that we compare "different", i.e., not comparable .

Case: weight gains in rats

- Data from $n = 40$ rats, fed on four diets.
- Two amounts of protein (low and high)
- Two sources of protein (beef and cereal)
- Randomized, Factorial, Balanced experiment.

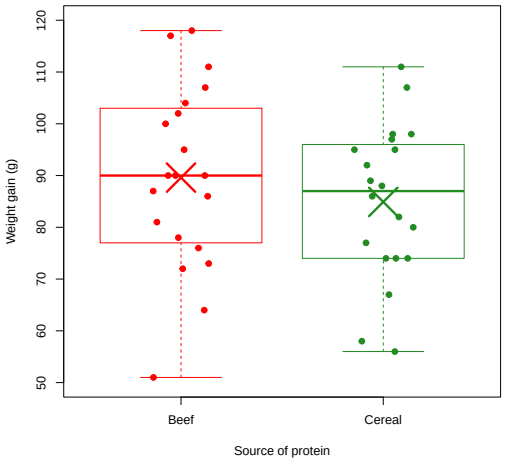


Outcome: weight gain in grams.

Research question: Does one of the two sources of proteins lead to larger weight gains (in average)?

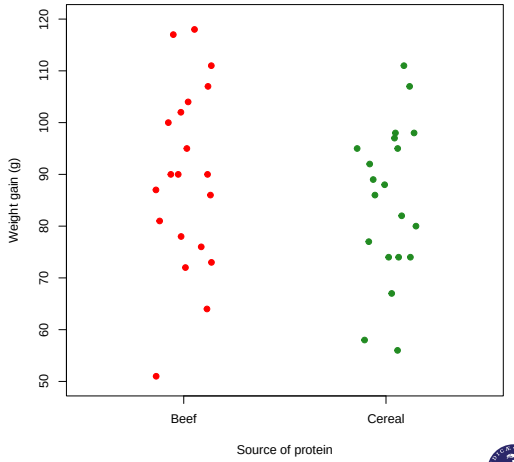
Simple comparison: t-test

- $\bar{x}_1 = 89.6$, $\hat{\sigma}_1$=17.7
- $\bar{x}_2 = 84.9$, $\hat{\sigma}_2$=15.0
- p-value of t-test=0.37
- Difference in mean 4.70, 95% CI = [-5.81;15.21]



Note: crosses show the means.

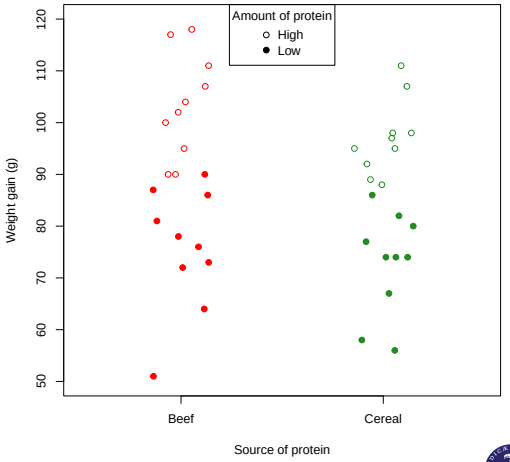
The data



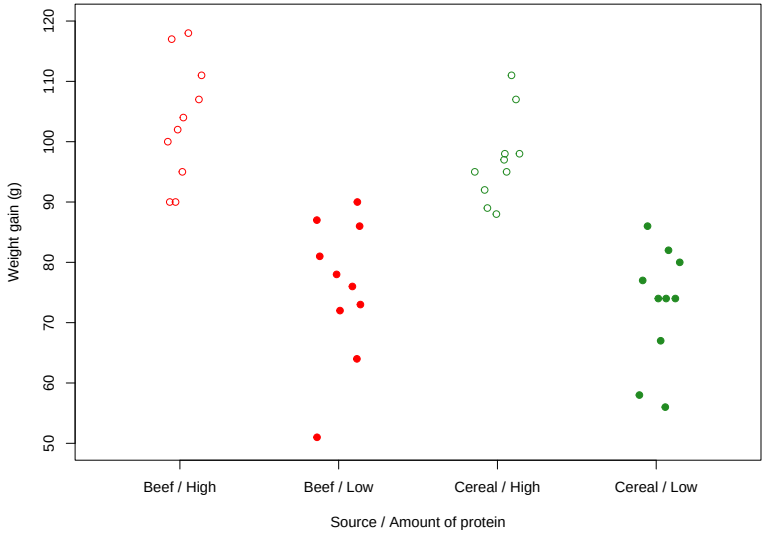
Adjusting increases power & precision

Theory (maths) shows that adjusting on a variable (strongly) associated with the outcome will generally increase the power of the analysis and the precision of the estimates (i.e. smaller s.e., narrower CI).

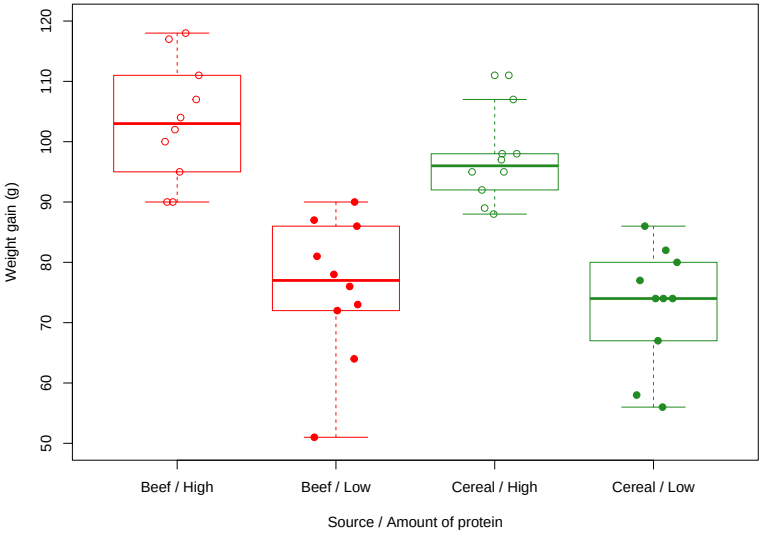
To get the intuition, let's imagine this hypothetical situation...



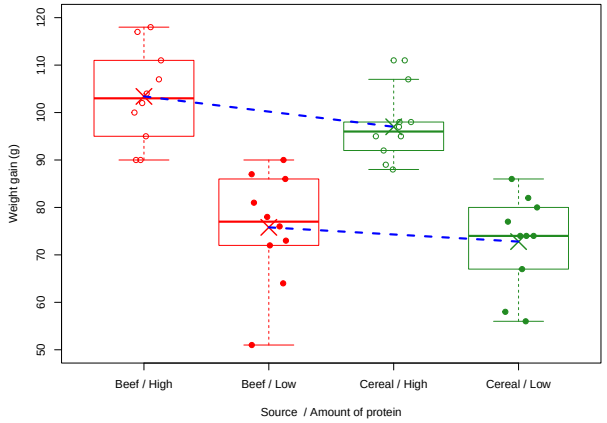
Unfolding the four groups (hypothetical) (1/3)



Unfolding the four groups (hypothetical) (2/3)



Unfolding the four groups (hypothetical) (3/3)



- Some evidence in each subgroup (High and Low)
- The evidence of each subgroup is based on less subjects ($\searrow n \Rightarrow \searrow$ evidence/power) but based on data with less variability ($\searrow \sigma \Rightarrow \nearrow$ evidence/power), which somehow "balances out".
- So, overall there is more evidence (i.e. smaller s.e. and more power).

The two-way ANOVA model (without interaction)

The k-th observation from the (i, j)-th combination group (e.g. source of protein i and amount of protein j) is described as:

$$\begin{aligned} Y_{ijk} &= \mu_{ij} + \varepsilon_{ijk} \\ &= \gamma_i + \eta_j + \varepsilon_{ijk} \qquad \text{(assuming no interaction)}. \end{aligned}$$

- $\mu_{ij} = \gamma_i + \eta_j$ is the mean for the (i, j)-th combination group.
- ε_{ijk}'s are individual 'error' terms ("random/unexplained deviation from the mean") assumed normally distributed with zero mean and the same variance σ^2_ε regardless of group.

Two-way ANOVA assumptions (without interaction)

Model assumptions (1-4 similar to that of the one-way ANOVA):

1. Observations from different groups are independent.
2. Individual observations within each group are independent.
3. 'Error' terms are normally distributed.
4. The variance of 'error' terms is the same for all groups (homogeneity).
5. There is no interaction ($\rightarrow$).

Note: 1-2 and 5 are important[22], 3-4 not always (as for the one-way ANOVA).

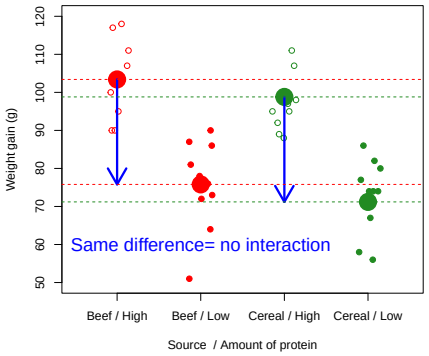
[22] Actually assumption 5 is not very important in the specific case of a randomized and "well balanced" experiment, when correctly interpreting the results as "marginal" differences in means.

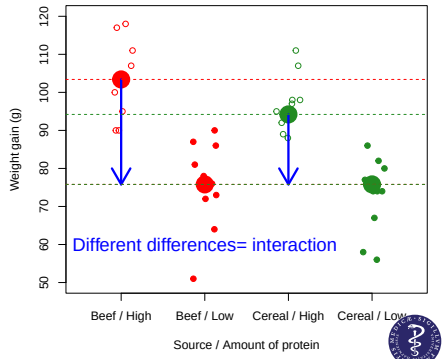
The meaning of "no interaction" (1/2)

No interaction models $\mu_{ij} = \gamma_i + \eta_j$ for the mean for the (i, j)-th combination group. In our example that means that:

The { source / amount } of protein "shifts" the mean of all { amounts / sources } (up or down) in the same way.

Example (hypothetical):



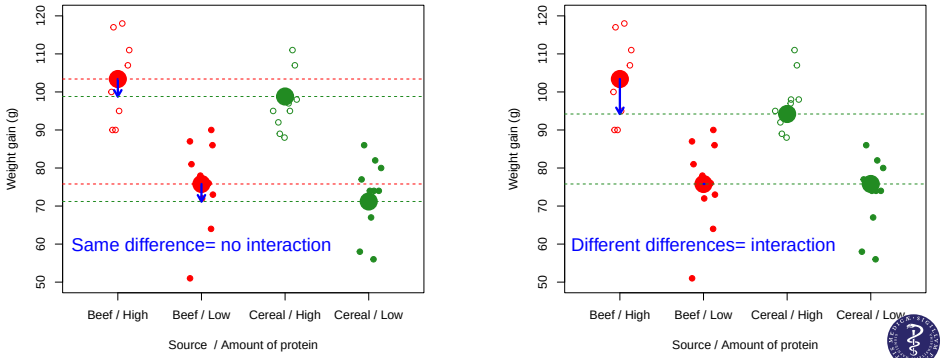


The meaning of "no interaction" (2/2)

No interaction models $\mu_{ij} = \gamma_i + \eta_j$ for the mean for the (i, j)-th combination group.
In our example that means that:

The { source / amount } of protein "shifts" the mean of all { amounts / sources } (up or down) in the same way.

Example (hypothetical):



Two-way ANOVA: usual software parametrization (1/4)

Now we come back to the real data and run the analysis!

One-way ANOVA (for comparison)

```
OneWayRes <- lm(weightgain~source,data=weightgain)
summary(OneWayRes)
```

which returns

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    89.600      3.669  24.419   <2e-16 ***
sourceCereal   -4.700      5.189  -0.906    0.371
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.41 on 38 degrees of freedom
Multiple R-squared:  0.02113,Adjusted R-squared:  -0.004628
F-statistic: 0.8203 on 1 and 38 DF,  p-value: 0.3708
```


Two-way ANOVA: usual software parametrization (2/4)

Two-way ANOVA

```
TwoWayRes <- lm(weightgain~type+source,data=weightgain)
summary(TwoWayRes)
```

which returns

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    95.300      4.255  22.396   <2e-16 ***
typeLow       -11.400      4.914  -2.320    0.026 *
sourceCereal   -4.700      4.914  -0.957    0.345
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.54 on 37 degrees of freedom
Multiple R-squared:  0.1455,Adjusted R-squared:  0.09926
F-statistic: 3.149 on 2 and 37 DF,  p-value: 0.05459
```

Note:

- Variable `type` in the data `weightgain` indicates the amount of protein.
- F-test p-value is not so interesting here (H_0: "neither effect of amount of protein nor of source of protein").
- Same estimated difference (-4.70) as in the one-way ANOVA because of the balanced factorial design.

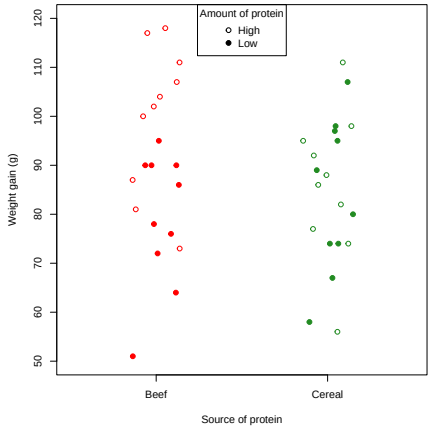
Two-way ANOVA: usual software parametrization (3/4)

- `(Intercept)`: est. mean in the reference group (High,Beef).
- `typeLow`: est. mean difference between amount (`type`) of protein, Low vs High, for the same source of protein (any).
- `sourceCereal`: est. mean difference between sources of protein, Cereal vs Beef, for the same amount (`type`) of protein (any).
- `F-statistic`: not so interesting (see previous slide).
- `Residual standard error`: estimate of σ_ε: 15.54

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Digression

In this example, the s.e. for the difference of interest is unfortunately only "a little bit" smaller with the two-way ANOVA than with the one-way ANOVA (4.914 vs 5.189). This is because the vertical spread of the observations of the weight gains (i.e. the standard deviation) is not largely different when looking at the entire data or within subgroups.



(plot of the real data)

Two-way ANOVA: usual software parametrization (4/4)

This ANOVA model is also a specific kind of linear model. The mean of each group is given by this regression formula:

$$\mu_{ij} = \alpha + \beta_1 \cdot \mathrm{I}(\text{amount}_j{=}\text{Low}) + \beta_2 \cdot \mathrm{I}(\text{source}_i{=}\text{Cereal})$$

Modeled means and estimates:

Source \ Amount	High	Low
Beef	α 95.3	$\alpha + \beta_1$ $95.3 + (-11.4) = 83.9$
Cereal	$\alpha + \beta_2$ $95.3 + (-4.7) = 90.6$	$\alpha + \beta_1 + \beta_2$ $95.3 + (-4.7) + (-11.4) = 79.2$

What if we compare more than two groups?

- In short, everything is very similar to what we have seen before.
- For illustration, let's artificially create additional data of 20 more observations from rats fed with Fish (again, 10 receive a Low amount of protein, 10 a High amount).

Two-way ANOVA (results with additional, artificial, data)

Two-way ANOVA

```
TwoWayRes <- lm(weightgain~type+source,data=weightgain)
summary(TwoWayRes)
```

which returns

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    95.250      3.786  25.157   <2e-16 ***
typeLow       -11.300      3.786  -2.984   0.0042 **
sourceCereal   -4.700      4.637  -1.014   0.3152
sourceFish    -10.050      4.637  -2.167   0.0345 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.66 on 56 degrees of freedom
Multiple R-squared:  0.1955,Adjusted R-squared:  0.1524
F-statistic: 4.537 on 3 and 56 DF,  p-value: 0.006454
```

F-test with two-way ANOVA

An F-test can be performed for the null hypothesis $\boxed{H_0 : \beta_2 = \beta_3 = 0}$, with model formula

$$\mu_{ij} = \alpha + \beta_1 \cdot \text{I(amount}_j\text{=Low)} + \beta_2 \cdot \text{I(source}_i\text{=Cereal)} + \beta_3 \cdot \text{I(source}_i\text{=Fish)}$$

that is

> H_0: "All sources of proteins give the same average weight gain, when comparing rats fed with the same amount of protein".

It compares the mean weight gain from all sources of proteins "adjusted" on the amount of protein received (i.e. within groups of rats receiving the same amount of proteins).

- This is a very commonly used, well-known and "old" method.
- Pros and cons: similar to that of F-test for one-way ANOVA.

R code and results

R code for the F-test (two-way ANOVA)

```
Full.lm <- lm(weightgain~type+source, data=weightgain) # "full" model (same as TwoWayRes)
Cons.lm <- lm(weightgain~type, data=weightgain)        # "constrained" model
anova(Cons.lm,Full.lm)                                 # F-test (compares the 2 models)
```

which returns

```
Analysis of Variance Table

Model 1: weightgain ~ type
Model 2: weightgain ~ type + source
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     58 13054
2     56 12042  2    1011.4 2.3517 0.1045
```

R code and results

R code for the F-test (two-way ANOVA)

```
Full.lm <- lm(weightgain~type+source, data=weightgain) # "full" model (same as TwoWayRes)
Cons.lm <- lm(weightgain~type, data=weightgain)        # "constrained" model
anova(Cons.lm,Full.lm)                                 # F-test (compares the 2 models)
```

which returns

```
Analysis of Variance Table

Model 1: weightgain ~ type
Model 2: weightgain ~ type + source
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     58 13054
2     56 12042  2    1011.4 2.3517 0.1045
```

Comments:

- F-test p-value=0.1045 is not significant.
- To avoid coding mistakes and misunderstandings of the R output, compare the two models and do not instead use "`anova(Full.lm)`", since the order of the variables in the formula would generally matter in that case.

Recommended analysis (see R-demo for code)

Statistical methods:
Comparisons between sources of proteins were made using a two-way ANOVA model (without interaction), to adjust for the amount of proteins received. P-values and 95% confidence intervals were adjusted for multiple testing using the max-t test method (aka min-p method) as implemented in the multcomp-package [ref.[23]] of R [ref.[24]] and described in [ref.[25]].

Results (adjusted for multiple testing):

Comparison	Est. Diff	95% CI	p-value
Cereal - Beef	-4.7	[-15.9; 6.5]	0.572
Fish - Beef	-10.1	[-21.2; 1.1]	0.086
Fish - Cereal	-5.4	[-16.5; 5.8]	0.486

We did not find a statistically significant association between the source of proteins and the mean weight gain (p=0.086).

Notes:

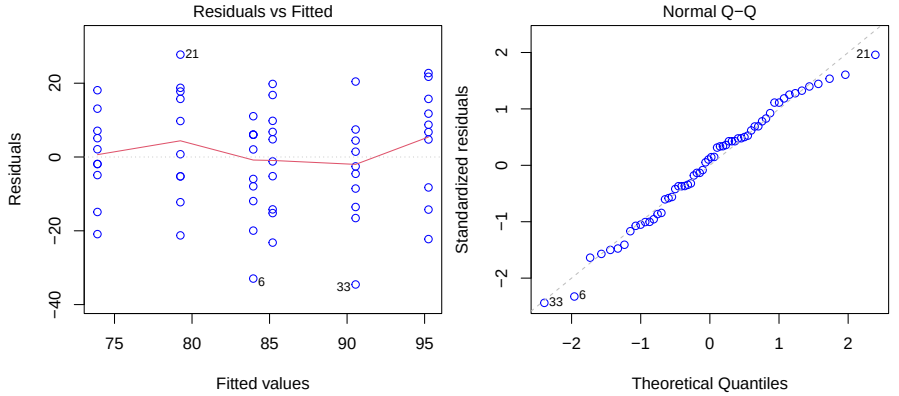
- p-values $\leq$ 3 times the non-adjusted ones obtained from the default summary (Bonferroni).
- results for comparisons, including "Fish - Cereal" (unlike in default summary)

[23] Hothorn, Bretz & Westfall (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal 50(3), 346–363.
[24] R Core Team (2026). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[25] Bretz, Hothorn, & Westfall (2016). Multiple comparisons using R. CRC Press.

About assuming "no interaction"

- This assumption can be important.
- It simplifies the interpretation of the results.
- It should be supported by subject-matter knowledge.
- This assumption can (most often) be checked with the data.
- Usually, the smaller the sample size the more assumptions we need to compensate. This applies to the assumption of no interaction.
- More on interactions in Lecture 7.
- not important in the specific case of a randomized experiment, when we correctly interpret the results as "marginal" differences in means.

Model checking (default) plots



Note: these are similar plots to those of the linear models.

(A few) Take home messages

- ANOVA is useful to compare the mean outcome in several groups.
- One-way: one categorical variable used in the analysis, Two-way: two categorical variables.
- A two-way ANOVA can be better than a one-way ANOVA (gain in power and precision), but not necessarily (more assumptions ussually required) .
- Pairwise comparisons are often relevant, but adjusting for multiple testing is necessary to control the risk of false discoveries.
- F-tests are often used, sometimes useful, but not always.
- All model assumptions are not equally important.
- Think about your research question first, then choose your statistical method (one- or two-way ANOVA, F-test, pairwise comparisons, which multiple testing adjustment: none, all-pairwise, many-to-one?)
- Check the main modeling assumptions, tune your conclusions appropriately.