Analysis of Variance (ANOVA)

Janette Walde

janette.walde@uibk.ac.at

Department of Statistics
University of Innsbruck

Outline I

Outline II

Problems/Questions

- Do fertilizer have different effects on different kind of wheat?
- React females differently on anti-cancer drugs as males?
- Does water evaporation of soil depend on the kind of vegetation growing, controlling for climate conditions?
- A new treatment meant to help those with chronic arthritis pain was developed and tested for its long-term effectiveness. Participants in the experiment rated their level of pain on a 0 (no pain) to 9 (extreme pain) scale at three-month intervals. Was the treatment effective?
- Does the exposure of plants to various amounts of CO2 affect characteristics of the plant?

ANOVA - Preliminaries

- **AN**alysis **O**f **VA**riance.
- Use variances and variance like quantities to study the equality or non-equality of population means.
- So, although it is analysis of variance we are actually analyzing means, not variances.
- There are other methods which analyze the variances between groups.
- Partitions the observed variance based on explanatory (independent) variables.
- Compares partitions to test significance on explanatory variables.

Basics

- **One-way ANOVA** is used when
  - Only testing the effect of one explanatory variable.
  - Each subject has only one treatment or condition. Thus, a between-subject design.
- Used to test for differences among two or more independent groups (in order to 'avoid' the multiple testing problem).
- Gives the same results as two sample t-tests if explanatory variable has to levels.

Model Description - Notation

- If we have K groups denote the means of the groups as $\mu_1, \mu_2, ..., \mu_K$.
- Subject i in group j has observation y_{ij}:
  - $y_{ij} = \mu_j + \epsilon_{ij}$
  - where ϵ_{ij} are independently distributed $N(0, \sigma^2)$.
  - Can combine this and say that subjects from group j have distribution $N(\mu_j, \sigma^2)$.
- With random assignment the sample mean for any treatment group is representative of the population mean for that group.

Model Assumptions

1. The errors ϵ_{ij} are normally distributed.
2. Across the conditions the errors have equal spread, referred to as equal variances.
   - Rule of thumb: the assumption is met if the largest variance is less than twice the smallest variance.
   - If unequal variances need to make a correction. If deviations are small this is usually $\alpha/2$.
3. The errors are independent from each other.

Typical exploratory analysis include

- Tabulation of the number of subjects in experimental group.
- Side-by-side box plots.
- Statistics about each group.

Application: Medley & Clements (1998) [MC1998]

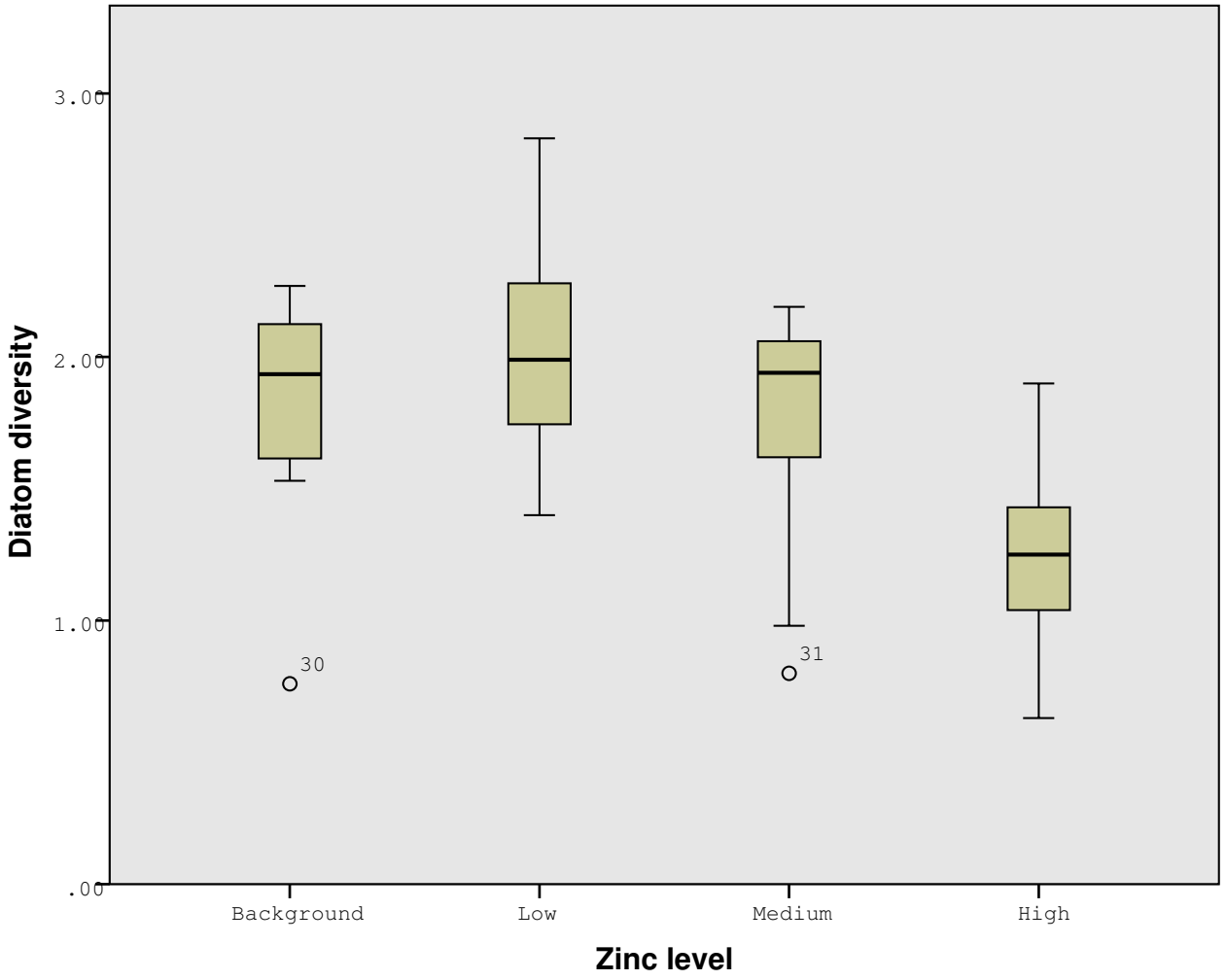
Medley & Clements (1998) samples a number of stations (between four and seven) on six streams known to be polluted by heavy metals in the Rocky Mountain region of Colorado, USA. They recorded zinc concentration, and species richness and species diversity of the diatom community and proportion of diatom cells that were the early-successional species, Achanthes minutissima.

Descriptive Analysis for [MC1998], I

Diatom diversity

Zinc level	mean	n	standard deviation
Background	1.7975	8	.48526
Low	2.0325	8	.44500
Medium	1.7178	9	.50301
High	1.2778	9	.42687
Total	1.6941	34	.52464

Descriptive Analysis for [MC1998], II



Hypothesis Testing

- $H_0 : \mu_1 = \mu_2 = ... = \mu_K$
- H_1: The μ's are not all equal.

- The null hypothesis is called the overall null and is the hypothesis tested by ANOVA.
- If the overall null is rejected you must do more specific hypothesis testing to determine which means are different, often referred to as contrasts or post-hoc analysis.

Theoretical Background and Terminology

- The 'sample' variance is the sum of the squared deviations from the mean divided by the number of observations minus 1

$$s^2 = \frac{\sum(y_i - \bar{y})^2}{n - 1}$$

- A mean square (MS) is a variance like quantity calculated as the sum of the squared deviations (SS) divided by the degrees of freedom (df)

$$MS = \frac{SS}{df}$$

Within versus Between

- In one-way ANOVA we work with two mean square quantities
  - MS_{within} ... the mean square within-groups
  - $MS_{between}$ ... the mean square between-groups
- For each individual group we have

$$\frac{SS_i}{df_i} = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}{n_i - 1}$$

- So the estimate of MS_{within} is

$$MS_{within} = \frac{SS_{within}}{df_{within}} = \frac{\sum_{i=1}^{K} SS_i}{N - K}$$

- And the estimate of $MS_{between}$ is

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_i - \bar{y})^2}{K - 1}$$

Mean Squares - What do these values mean?

- MS_{within} is considered a true estimate of σ^2 that is unaffected by whether the null or alternative hypothesis is true.
- $MS_{between}$ is considered a good estimate of σ^2 only when the null hypothesis is true. If the alternative is true, values of $MS_{between}$ tend to be inflated.
- Thus, we can look at the ratio of the two mean square values to evaluate the null hypothesis.

Testing the Hypothesis

- The F-test looks at the variation among the group means relative to the variation within the sample

$$F = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between}/df_{between}}{SS_{within}/df_{within}} = \frac{SS_{between}/(K-1)}{SS_{within}/(N-K)}$$

- The F-statistic tends to be larger if the alternative hypothesis is true than if the null hypothesis is true.

- The test statistic F has an $F(K-1, N-K)$ distribution.

What does the F ratio tell us?

$$F = MS_{between} / MS_{within}$$

- The denominator is always an estimate of σ^2 (under both the null and alternative hypotheses).
- The numerator is either another estimate of σ^2 (under the null) or is inflated (under the alternative).
- If the null is true, values of F are close to 1.
- If the alternative is true, values of F are larger.
- Large values of F depend on the degrees of freedom.

The ANOVA table

When running an ANOVA, statistical packages will return an ANOVA table summarizing the SS, MS, df, F-statistic, and p-value:

	SS	df	MS	F	Sig
Group (Treatment, between)	$SS_{bet.}$	$df_{bet.}$	$MS_{bet.}$	$\frac{MS_{bet.}}{MS_{within}}$	p-value
Residual (Error, within)	SS_{within}	df_{within}	MS_{within}		
Total	$SS_{bet.} + SS_{within}$	$df_{bet.} + df_{within}$			

Back to our application [MC1998]

ONEWAY ANOVA

Diatom diversity

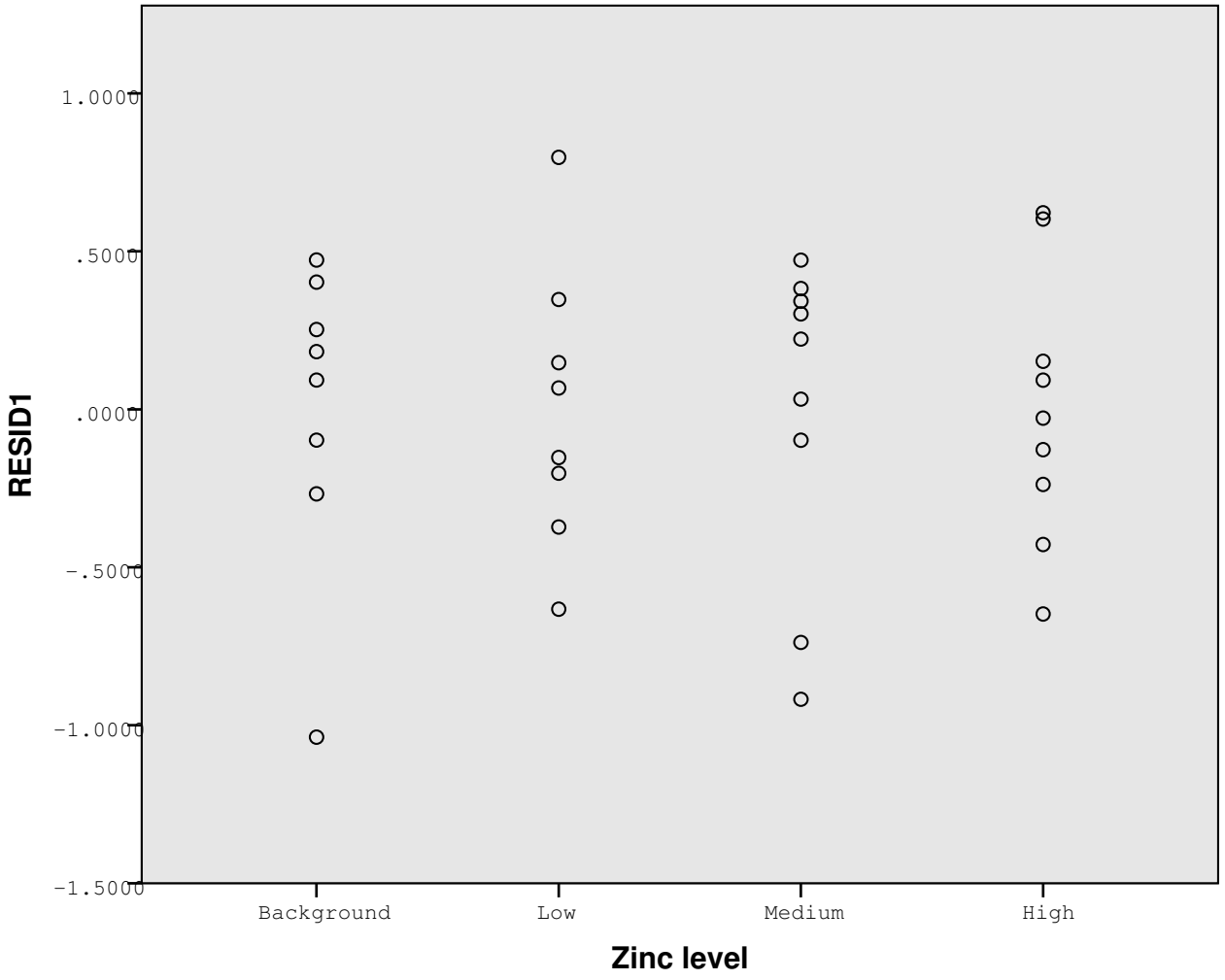
	SS	df	MS	F	Significance
Between groups	2.567	3	.856	3.939	.018
Within groups	6.516	30	.217		
Total	9.083	33			

$H_0 : \mu_{Background} = \mu_{Low} = \mu_{Medium} = \mu_{High}$

Checking the assumptions

- Use the residuals which are the estimates of ϵ_{ij}.
  1. Look at normal probability plot.
  2. Look at residual versus fitted plot.
  3. Independence is hard to check and often just assumed from study design.
- For mild violations of the assumptions there are options for correction.
- When the assumptions are NOT met the p-values are simply wrong.

Checking the assumptions for [MC1998], Ia



Checking the assumptions for [MC1998], Ib

Test for homogeneity of variances

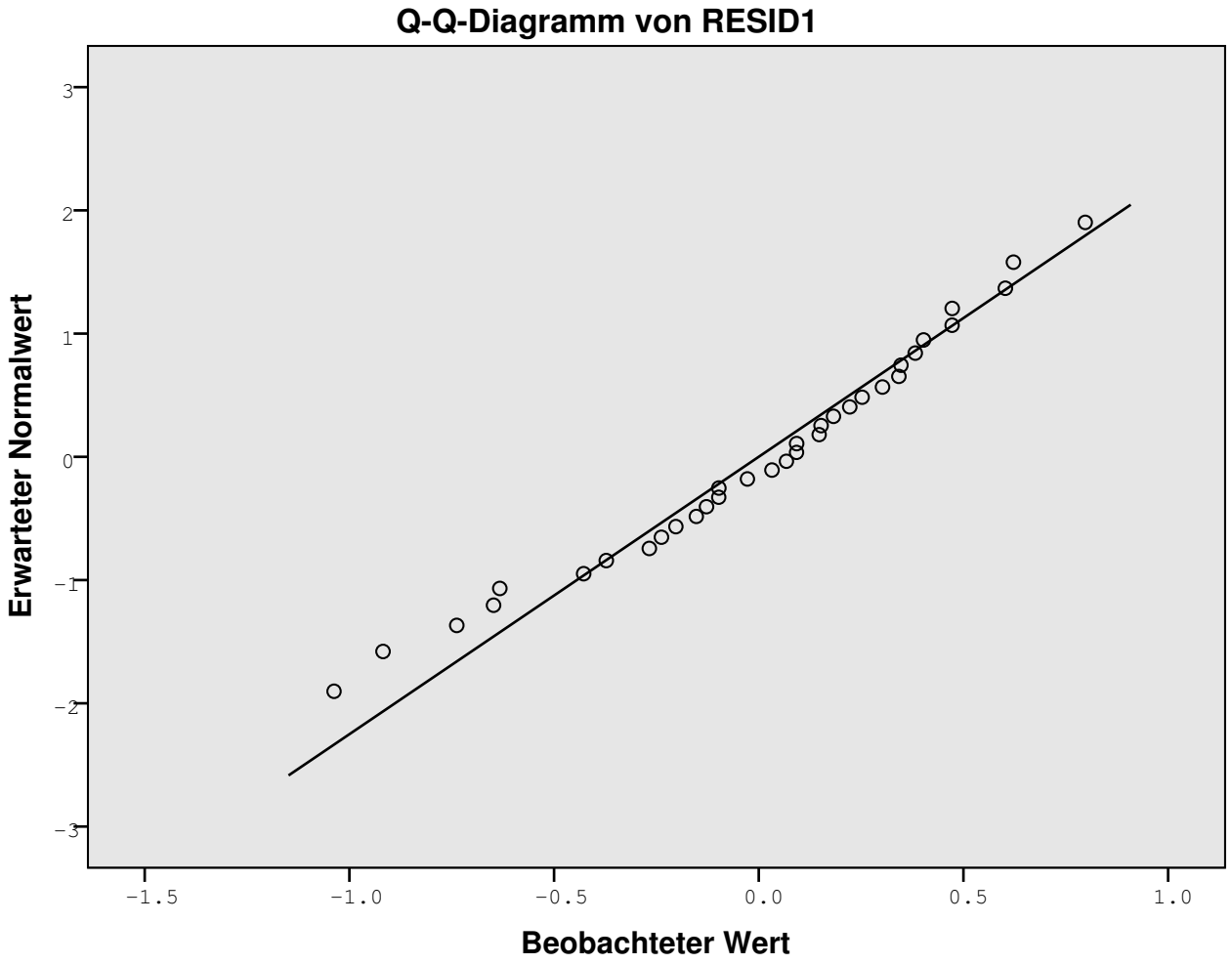
Diatom diversity

Levene statistic	df1	df2	significance
.087	3	30	.967

$H_0 : \sigma^2_{Background} = \sigma^2_{Low} = \sigma^2_{Medium} = \sigma^2_{High}$

Assumption of homogeneity of variances is appropriate.

Checking the assumptions for [MC1998], IIa



Checking the assumptions for [MC1998], IIb

Normality Test

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	statistic	df	sig.	statistic	df	sig.
RESID1	.090	34	.200	.969	34	.430

[a] Correction of significance using the approach of Lilliefors.

Post-Hoc Analysis

- If H_0 is rejected, we conclude that not all the μ's are equal.
- We would like to make statements about where there are differences.
- Can use planned or unplanned comparisons (or contrasts).
  * Planned comparisons are interesting comparisons decided on before analysis.
  * Unplanned comparisons occur after seeing the results. Be careful not to go fishing for results!

Post-Hoc Analysis, cont.

- What if we notice a possible interesting difference when looking at the results?
- Can do comparisons but need to adjust the α-level to control for Type-1 error.
- Bonferroni correction for the number of comparisons done:

$$\alpha^* = \frac{\alpha}{\text{number of comparisons}}$$

(Bonferroni-Holm correction).

Post-Hoc Analysis for [MC1998]

Multiple comparisons

Diatom diversity, Bonferroni

(I) Zinc level	(J) Zinc level	mean diff. (I-J)	s.e.	sig.	95%-CI lower limit	95%-CI upper limit
B.ground	Low	-.23500	.23303	1.000	-.8932	.4232
	Medium	.07972	.22647	1.000	-.5600	.7194
	High	.51972	.22647	.173	-.1200	1.1594
Low	B.ground	.23500	.23303	1.000	-.4232	.8932
	Medium	.31472	.22647	1.000	-.3250	.9544
	High	.75472*	.22647	.014	.1150	1.3944
Medium	B.ground	-.07972	.22647	1.000	-.7194	.5600
	Low	-.31472	.22647	1.000	-.9544	.3250
	High	.44000	.21970	.326	-.1806	1.0606
High	B.ground	-.51972	.22647	.173	-1.1594	.1200
	Low	-.75472*	.22647	.014	-1.3944	-.1150
	Medium	-.44000	.21970	.326	-1.0606	.1806

* difference of means is significant at the 0.05 level.

Application

The data are gathered from a plant physiology experiment which investigated the effect of various sugar on the growth of peas. Growth or length is measured in 'ocular units'. Five groups are analyzed, a control group and four groups varying in the kind of sugar and its amount. In each groups there are 10 measurements.

sub-jects	control group	2% glucose	2% fructose	1% glucose + 2% saccharose	1% fructose
1	71	57	58	58	62
2	68	58	61	59	66
3	70	60	56	58	65
⋮	⋮	⋮	⋮	⋮	⋮

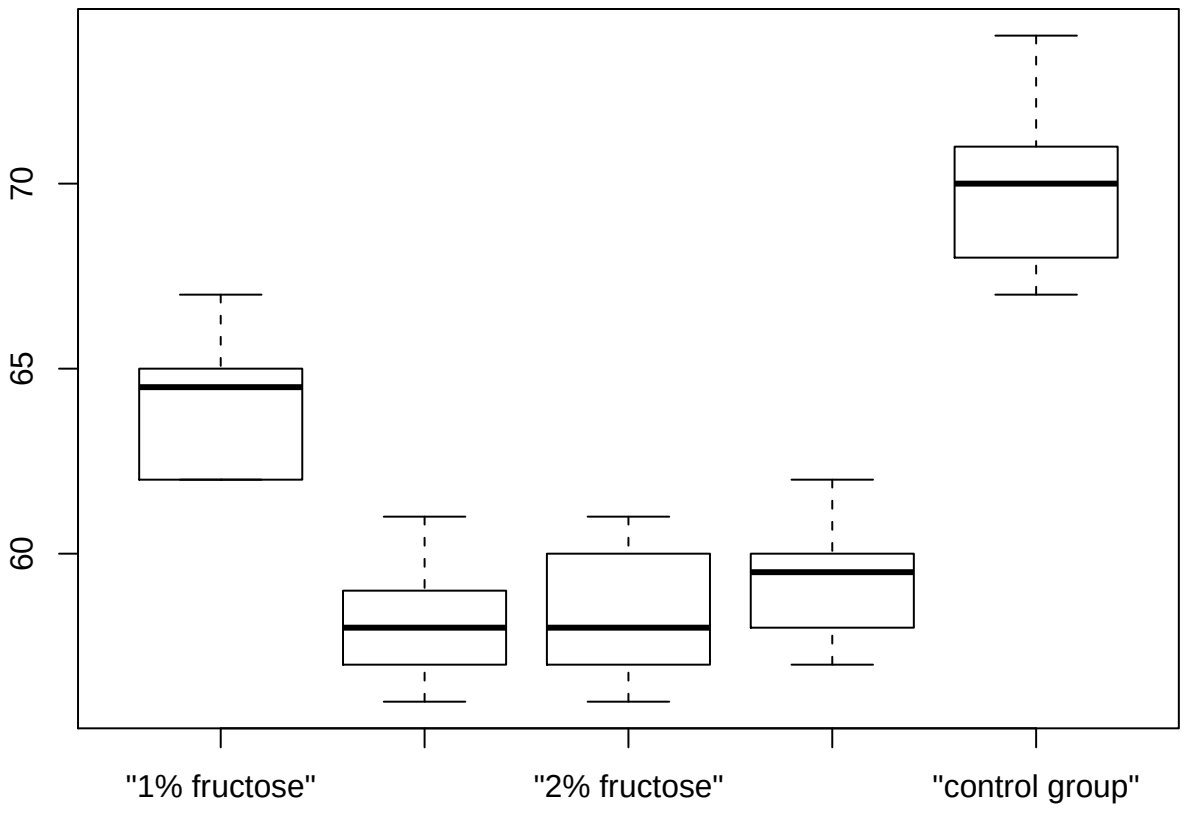
Example

- We want to know whether the means of the variable 'length' differ significantly across the groups, i.e. does the supply with sugar influence the growth of the peas?
- $K = 5$: control group, group 1, group 2, group 3, group 4.
- H_0 : Growth is independent of the sugar support.

  $H_0 : \mu_{controlgroup} = \mu_{group1} = \mu_{group2} = \mu_{group3} = \mu_{group4}$
- H_1 : Growth varies across groups.

  H_1 : At least one of the means is different.

Box plots



Summary

The largest variance is less than twice the smallest variance ($2.2 < 2 \cdot 1.4 = 2.8$). Use $\alpha = 0.05$.

Groups	n_i	Mean	Variance
Control group	10	70.1	2.2
Group 4	10	64.1	1.8
Group 3	10	58.0	1.4
Group 2	10	58.2	1.9
Group 1	10	59.3	1.6

Sample Output

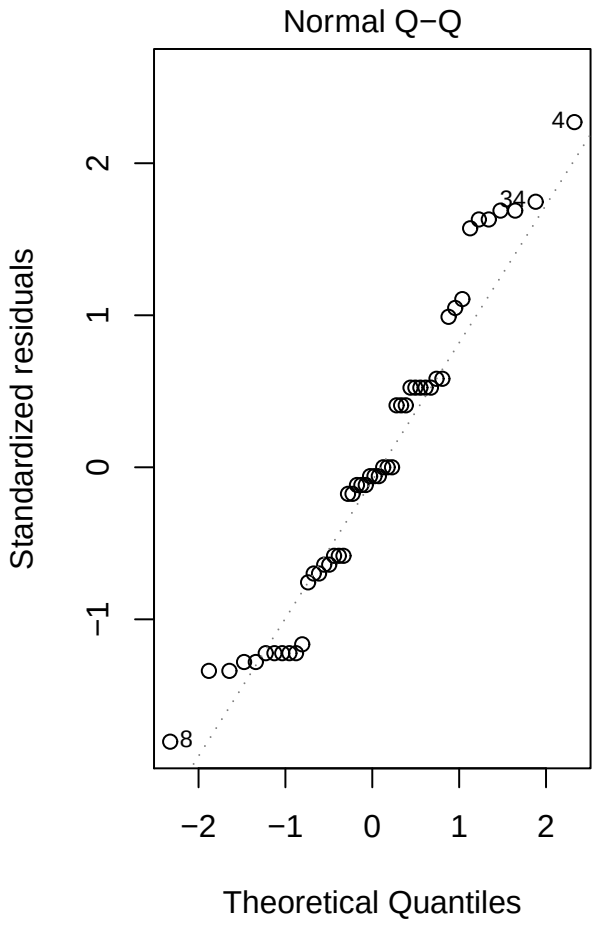
	SS	*df*	*MS*	*F*	*Sig*
Group (Treatment, between)	1077.3	4	269.330	82.168	0.000
Residual (Error, within)	147.5	45	3.278		
Total	1224.8	49			

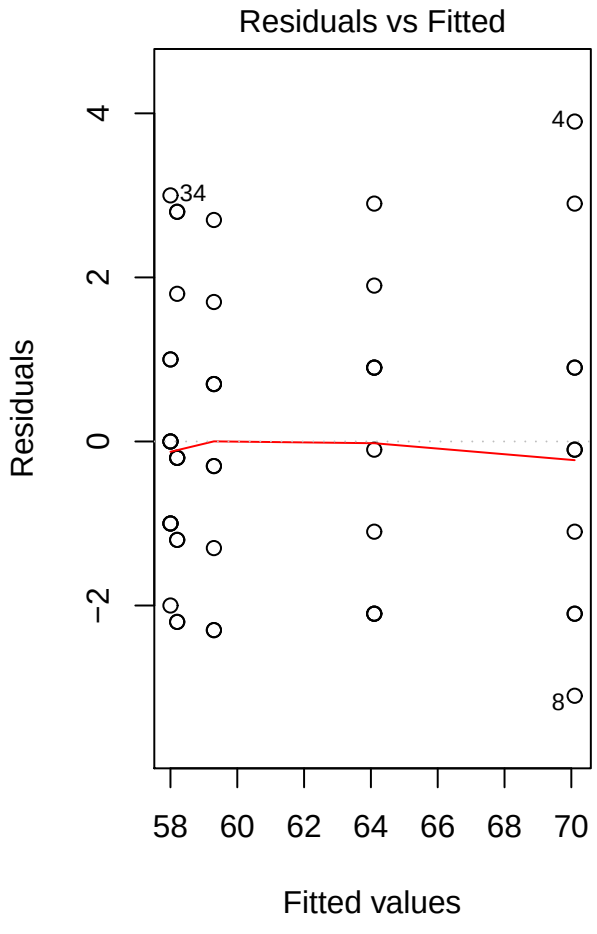
- Our estimate of σ^2 is approximately 3.3.
- The numerator $MS_{between} = 269.330$ and appears to be highly inflated.

Results

- F-statistic $= 22.1$.
- p-value: < 0.05.
- *Conclusion* - the growth differs for at least one of the groups.
- To make stronger statements need to do further testing.

Checking the assumptions





Contrasts

- A simple contrast hypothesis compares two population means:
  - $H_0 : \mu_1 = \mu_5$
- A complex contrast hypothesis has multiple population means on either side:
  - $H_0 : (\mu_1 + \mu_2)/2 = \mu_3$
  - $H_0 : (\mu_1 + \mu_2)/2 = (\mu_3 + \mu_4 + \mu_5)/3$

Comparison to Regression Analysis

- The conclusions about the overall null hypothesis will be the same.
- In regression can make statements comparing groups to baseline.
- To make more conclusive statements will need to do more analysis.
- ANOVA and either planned or post-hoc comparisons will do the same and is often easier.

ANOVA as regression output

Coefficients:	Estimate	Std. Error	t value	Sig.
(Intercept)	64.1	0.5725	111.961	0.000
"1% glu, 2% sac"	-6.1	0.8097	-7.534	0.000
"2% fructose"	-5.9	0.8097	-7.287	0.000
"2% glucose"	-4.8	0.8097	-5.928	0.000
control group"	6.0	0.8097	7.410	0.000
"1% fructose"	0^b			

[b] This parameter is set to zero, it is redundant.

$R - Squared = .880$ (adjusted $R - Squared = .869$)

F-statistic: 82.17 on 4 and 45 df, Sig.: 0.000

Bonferroni-Holm correction for application

Multiple Comparisons, Bonferroni

(I)group	(J)group	mean difference (I-J)	standard error	Sig.	95%-CI lower limit	95%-CI upper limit
0	1	10.80*	.810	.000	8.41	13.19
	2	11.90*	.810	.000	9.51	14.29
	3	12.10*	.810	.000	9.71	14.49
	4	6.00*	.810	.000	3.61	8.39
1	0	-10.80*	.810	.000	-13.19	-8.41
	2	1.10	.810	1.000	-1.29	3.49
	3	1.30	.810	1.000	-1.09	3.69
	4	-4.80*	.810	.000	-7.19	-2.41
2	0	-11.90*	.810	.000	-14.29	-9.51
	1	-1.10	.810	1.000	-3.49	1.29
	3	.20	.810	1.000	-2.19	2.59
	4	-5.90*	.810	.000	-8.29	-3.51
3	0	-12.10*	.810	.000	-14.49	-9.71
	1	-1.30	.810	1.000	-3.69	1.09
	2	-.20	.810	1.000	-2.59	2.19
	4	-6.10*	.810	.000	-8.49	-3.71
4	0	-6.00*	.810	.000	-8.39	-3.61
	1	4.80*	.810	.000	2.41	7.19
	2	5.90*	.810	.000	3.51	8.29
	3	6.10*	.810	.000	3.71	8.49

*. The mean difference is significant at the 5% level.

Post Hoc Power Analysis - [MC1998]

Between Subject Effects

Dependent Variable: Diatom diversity

Source	SS Typ III	df	MS	F	sig.	power[b]
Corrected model	2.567[a]	3	.856	3.939	.018	.778
Constant	98.657	1	98.657	454.195	.000	1.000
ZINC	2.567	3	.856	3.939	.018	.778
Residual	6.516	30	.217			
Total	106.664	34				
Corrected total variation	9.083	33				

[a] $R - squared = .283$ (adjusted $R - squared = .211$)

[b] $\alpha = .05$

A-Priori Power Analysis

- Two different TOEFL prep. courses charge $1200 for a two month course. An (unethical) experiment would be to randomize students into one of the two courses or take no course.
- What information is needed to calculate power for this one-way ANOVA?
  * Sample size
  * Within group variance (σ^2)
  * Estimated or minimally interesting outcome means for each group.

Estimate of σ^2

Based on previous years, we know that 95% of the student scores on TOEFL fall between 900 and 1500:

- $\sigma = (1500 - 900)/4 = 150$
- $\sigma^2 = 150^2$

Minimally interesting outcome

- What is the minimally average benefit, in points gained, that would justify the program?
  The minimally interesting outcome is based on previous knowledge.
- For this example we'll try several different values.

Computing the Power

- Different applets will define things slightly different (*http* : *//www.epibiostat.ucsf.edu/biostat/sampsize.html*).
- For the applet I used ('nQuery'), they require 'sd[treatment]'. From their definition this is calculated as:

$$\text{sd[treatment]} = \sqrt{\frac{\sum_{i=1}^{K}(\mu_i - \overline{\mu})^2}{K}}$$

μ_i ... mean of group i

K ... number of groups

- Ready to go to power applet.

Computing the Power, Cont.

- Let $\sigma = 150$, $n = 50$, effect $= 50$ points
  Power $= 38\%$
- Let $\sigma = 150$, $n = 100$, effect $= 50$ points
  Power $= 68\%$
- Let $\sigma = 150$, n $= 50$, effect $= 100$ points
  Power $= 94\%$
- Let $\sigma = 150$, n $= 50$, effect $= 25$ points
  Power $= 12\%$
- Let $\sigma = 100$, n $= 50$, effect $= 50$ points
  Power $= 73\%$
- Let $\sigma = 100$, n $= 100$, effect $= 50$ points
  Power $= 96\%$
- Let $\sigma = 100$, n $= 50$, effect $= 100$ points
  Power $= 99\%$
- Let $\sigma = 100$, n $= 50$, effect $= 25$ points
  Power $= 23\%$

Moving past One-way ANOVA

- What if we have two categorical explanatory variables?
- What if we have categorical and quantitative explanatory variables?
- What if subjects have more than one treatment?
- What if there is more than one response variable?
- And many other combinations ...

Two-way ANOVA (Factorial Design)

Two-way (or multi-way) ANOVA is an appropriate analysis method for a study with a quantitative outcome and two (or more) categorical explanatory variables.

Suppose we now have two categorical explanatory variables:

- Is there a significant X_1 effect?
- Is there a significant X_2 effect?
- Are there significant interaction effects?
- If the interaction is significant, the model is called an interaction model.
- If the interaction is not significant, the model is called an additive model.

If X_1 has k levels and X_2 has m levels, then the analysis is often referred to as a 'k by m ANOVA' or '$k \times m$ ANOVA'.

Model and Assumptions

Model

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \\ SS_{Total} &= SS_A + SS_B + SS_{AB} + SS_{Residual} \end{aligned}$$

where factor A has p groups ($i = 1, ..., p$), factor B has q groups ($j = 1, ..., q$) crossed with each level of A and there are n_c replicates ($k = 1, ..., n_c$) within each combination of A and B categories, i.e. each cell.

The assumptions are the same as in One-way ANOVA:

1. The errors ϵ_{ijk} are normally distributed.
2. Across the conditions, the errors have equal spread. Often referred to as equal variances.
3. The errors are independent from each other.

Model Results

- Results are again displayed in an ANOVA table
- Will have one line for each term in the model. For a model with two factors, we will have one line for each factor and one line for the interaction. We will also have a line for the error and the total.
- Interactions measure whether the effect of one factor depends on the levels of the other factor and vice versa.
- See next page.

The ANOVA table

	SS	df	MS	F	$Sig.$
Factor 1		$k-1$			
Factor 2		$m-1$			
Interaction		$(k-1)(m-1)$			
Error		$N-k\cdot m$	$\star$		
Total		$N-1$			

- The MS(error), denoted by $\star$ in the above table, is the true estimate of σ^2.
- The MS in each row is that row's SS/df.
- The F-statistic is the MS/MS(error).

Interaction Plot

- Table of means
- Interaction Plot
  * An interaction plot is a way to look at outcome means for two factors simultaneously.
  * A plot with parallel lines suggests an additive model.
  * A plot with non-parallel lines suggests an interaction model.
  * Note that an interaction plot should **NOT** be the deciding reason in whether or not to run an interaction model.

Application

Do anti-cancer drugs have different effects on males and females? Three types of different drugs are given patients having cancer. The diameter of the tumor is measured.

- X_1: Kind of drug - 3 levels
- X_2: Gender - 2 levels
- Response: Tumor diameter

We will fit a 3 by 2 ANOVA.

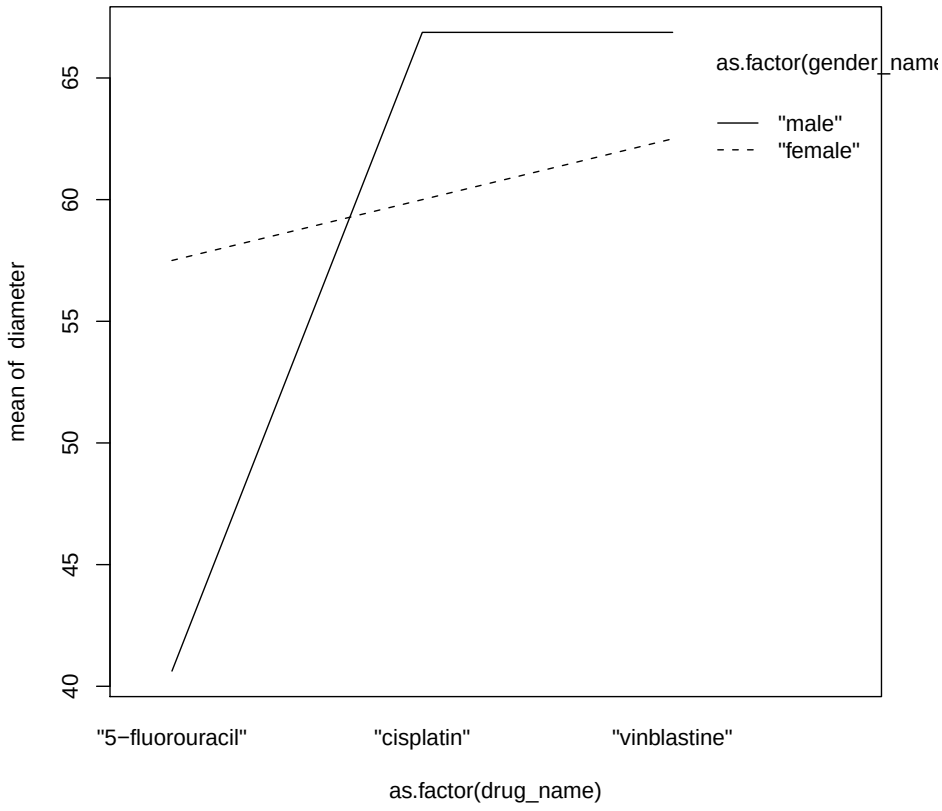
Table of means and counts

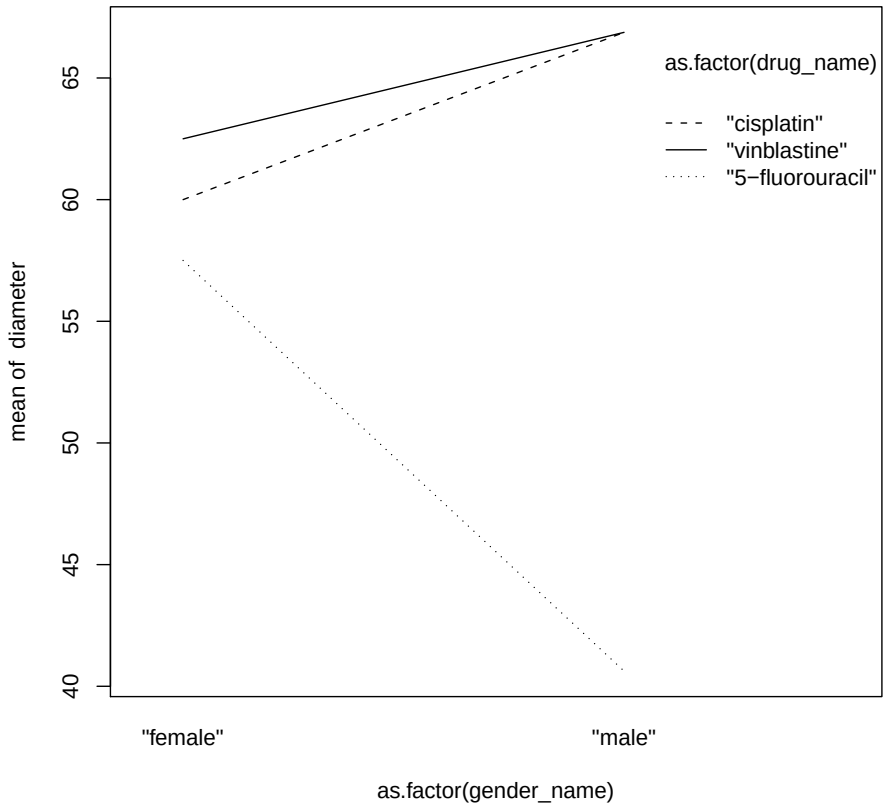
	Male	Female	Overall
Cisplatin	66.875	60.0	63.4375
Vinblastine	66.875	62.5	64.6875
5-fluorouracil	40.625	57.5	49.0625
Overall	58.125	60.0	59.0625

Note, this table should also include the standard error of each of the means.

	Male	Female
Cisplatin	8	8
Vinblastine	8	8
5-fluorouracil	8	8

Interaction plots





Interaction plots

- There are two ways to do an interaction plot. Both are legitimate. Ease of interpretation is the final criteria of which to do.
- If one explanatory variable has more levels than the other, interpretation is often easier if the explanatory variable with more levels defines the x-axis.
- If one explanatory variable is quantitative but has been categorized and the other is categorical, interpretation is often easier if the categorized quantitative variable defines the x-axis.

Results

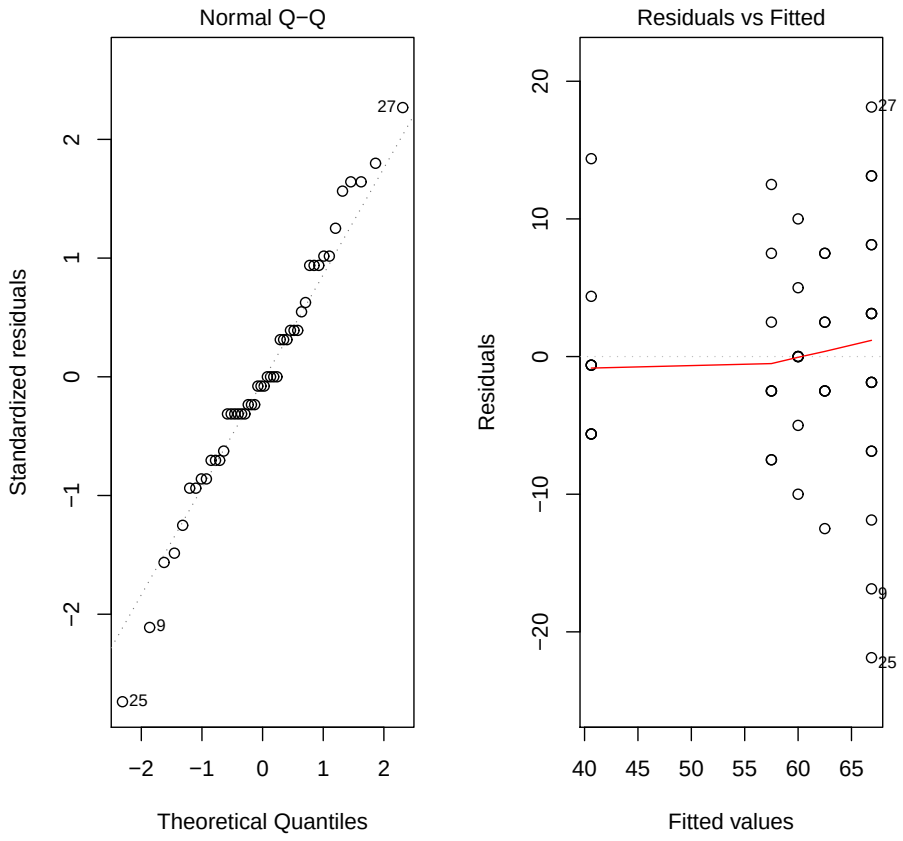
Between Subjects Effects

Source	SS Typ III	df	MS	F	Sig.
corrected model	3817.187[a]	5	763.437	10.459	.000
constant	167442.187	1	167442.187	2294.009	.000
gender	42.187	1	42.187	.578	.451
drug	2412.500	2	1206.250	16.526	.000
gender * drug	1362.500	2	681.250	9.333	.000
error	3065.625	42	72.991		
total	174325.000	48			
corrected total variation	6882.812	47			

[a] $R - squared = .555$ (adjusted $R - squared = .502$)

The column denoted as *Sig.* contains the *p*-values

Checking the assumptions



Notes

- The main effects should always be kept if the interaction is significant.
- Note that due to the groups, you will see vertical lines in the residual versus predicted plot. This is due to the fact that all subjects with a particular combination of the factors will have the same predicted value.

Post-hoc Comparisons

3. gender * drug

gender	drug	mean	standard error	95%-CI	
				lower limit	upper limit
0	1	66.875	3.021	60.779	72.971
	2	66.875	3.021	60.779	72.971
	3	40.625	3.021	34.529	46.721
1	1	60.000	3.021	53.904	66.096
	2	62.500	3.021	56.404	68.596
	3	57.500	3.021	51.404	63.596

Further Model Extensions I

- Nested (hierarchical) designs: additional factors are included that are nested within the main factor of interest. E.g., Caselle and Warner (1996) looked at recruitment densities of a coral reef fish at five sites on the north shore of the US Virgin Islands, with six random transects within each site and replicate observations of density of recruits along each transect.

- **An**alysis of **Cova**riance
  * At least one quantitative and one categorical explanatory variable are included in the model.
  * In general, the main interest is the effects of the categorical variable and the quantitative variable is considered to be a control variable.

Further Model Extensions II

* It is a blending of regression and ANOVA.

- Multivariate designs: MANOVA/MANCOVA

- Medley, C.N. and Clements, W.H. (1998). Responses of diatom communities to heavy metals in streams: the influence of longitudinal variation. *Ecological Applications* 8: 631-644.
- Caselle, J. E. and Warner, R.R. (1996). Variability in recruitment of coral reef fishes: the importance of habitat at two spatial scales. *Ecology* 77: 2488-2504.
- Quinn, G.P. and Keough, M. J. (2010). *Experimental Design and Data Analysis for Biologists*. Cambridge University Press. Cambridge.