Advanced Statistics

Analysis of Variance designs

Bernhard Angele

Goals of this lecture

- Learn about the relationship between dummy-coding and ANOVA
- Learn how to perform and report various types of ANOVA analyses
- Learn about different sum of square types
- Learn about the role of different contrasts (and how to make your own).

Dummy-coded variables and ANOVA

- Essentially, the Analysis of Variance is simply a multiple regression with only discrete variables coded as dummy contrasts
- As we've seen, we can perform F-tests comparing the full model with the dummy contrasts to a null model without the contrasts
- This gives us an overall test of whether there are *any* differences between the levels of the discrete variable, rather than a specific comparison..

Analysis of Variance

- The idea behind the analysis of variance is simple: We want to split the total variance in our data into variance explained by our grouping factor and random noise (error) variance.
- If the grouping factor explains more of the variance than we would expect based on random noise, then we can conclude that the grouping factor *significantly* improves our model (because yes, an ANOVA is a very simple statistical model)
- In other words, we can conclude that at least two of the factor levels are significantly different

Which contrasts should we use?

- Simplest way: Dummy contrasts (0 vs. 1)
  - Just like we did last time:
    - Here's a regression model with just one discrete predictor (let's assume for now that it has two levels only):

$$Y_i = \alpha + \gamma_1 D_i + \varepsilon_i$$

ANOVA as dummy contrast regression

- We want to compare the full model to the null model without the discrete predictor (intercept-only)
  - Full model: $Y_i = \alpha + \gamma_1 D_i + \varepsilon_i$
  - Null model: $Y_i = \alpha + \varepsilon_i$
- Remember, we need to get the regression and residual sums of squares for both models:

$$F_0 = \frac{(RegSS_1 - RegSS_0)/q}{RSS_1/(n-k-1)}$$

  - Also remember that $RSS = SS_{Total} - RegSS$

Calculating the sums of squares

- With a model this simple the sums of squares are easy enough to calculate
  - These equations are directly taken from our lecture on regression, so they are not new!

$$SS_{total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$SS_{model_1} = RegSS_1 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

- In a model with just discrete predictors, the predictions are just the group means!

The regression sums of squares

- Let's call the group means $\bar{A}_1$ and $\bar{A}_2$ and substitute them for $\hat{Y}_i$:

$$RegSS_1 = \sum_{i=1}^{i}(\hat{Y}_i - \bar{Y})^2 = \sum_{j=1}^{m} n_j(\bar{Y}_j - \bar{Y})^2$$

- m is the number of levels, so 2 in this case. j is the current level. n_j is the number of observations in the current level. $\bar{Y}_j$ is the mean of Y for the current level. In effect, For each observation, we put in the corresponding group mean and subtract the overall mean.
  - (actually, we subtract the mean of group means, but that's only important if our groups are of different sizes)

The regression sums of squares (2)

- The null model predicts every observation with the overall mean $\bar{Y}$, so actually it explains **no** variance (every prediction is the same!)

$$RegSS_0 = \sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2 = n \cdot (\bar{Y} - \bar{Y})^2 = 0$$

- For the error sum of squares RSS_1, remember that the error is simply the difference between the observed and the predicted Y-values are just the group means $\bar{Y}_j$:

$$RSS_1 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \hat{Y}_{ij})^2 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2$$

Contrasts and means

- Now let's plug in $RegSS_0$, $RegSS_1$, and RSS_1 into the equation for the F-value:

$$\begin{aligned} F_0 &= \frac{(RegSS_1 - RegSS_0)/q}{RSS_1/(n-k-1)} \\ &= \frac{(RegSS_1 - 0/k}{RSS_1/(n-k-1)} = \frac{RegSS_1/k}{RSS_1/(n-k-1)} \end{aligned}$$

- q is the number of extra predictors. In our example with two groups, null model has 0 predictors, and the full model has one dummy predictor, so $q = k - 0 = k$
- $k = m - 1$ is the number of predictors of the full model (i.e. the number of group means m minus 1).
- This F-value has degrees of freedom of $df_1 = m - 1$ and $df_2 = n - (m-1) - 1 = n - m$ or $n - k - 1$ (same thing!)

Textbook ANOVA

- In undergraduate textbooks, usually a slightly different terminology is used.
- Often, people use the index j to denote the factor level and the index i to denote the i^{th} observation (e.g. the i^{th} person) within each factor level. Here the dependent variable is (somewhat confusingly) called x and the group means are called $A_1, A_2, \ldots$ up to A_j, while the observations within each group are called $x_{i1}, x_{i2}, \ldots$ up to x_{ij}.
- Let's also keep calling the number of factor levels m and the number of observations per level n_j.

Textbook ANOVA (2)

-These are the exact same formulas, though! I just mention this so you don't get confused when you look back into your old textbooks.

$$SS_{total} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2$$

$$SS_{model} = RegSS_1 = \sum_{j=1}^{m} n \cdot (\bar{A}_j - \bar{x})^2 = n \cdot \sum_{j=1}^{m}(\bar{A}_j - \bar{x})^2$$

$$SS_{error} = RSS_1 = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{A}_i)^2$$

Textbook ANOVA (3)

-The rest of the ANOVA works just like before (MS stands for mean squares):

$$df_{total} = n - 1$$

$$df_{model} = m - 1$$

$$df_{error} = n - m$$

$$MS_{model} = \frac{SS_{model}}{df_{model}}$$

$$MS_{error} = \frac{SS_{error}}{df_{error}}$$

$$F_{df_{model},df_{error}} = \frac{MS_{model}}{MS_{error}}$$

Data matrix for an ANOVA

- Using "textbook" terminology
- Columns are factor levels, rows are observations within a level
- j = factor level, m = number of factor levels
- i = observation, n_j = number of observations per factor level

$$\begin{array}{cccccc}
x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\
x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{n_1 1} & x_{n_j 2} & \dots & x_{n_j j} & \dots & x_{n_j m}
\end{array}$$

Reminder: What is an F-value?

- If you have followed our discussion of how **variance estimates** of random variables that come from a normal distribution are always χ^2 distributed, you may not be too surprised by this.
- An F-value is the *quotient* of two random variables following the χ^2 distribution, each divided by their degrees of freedom:
$$F_{(n_1,n_2)} = \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2} = \frac{\chi^2_{n_1}}{\chi^2_{n_2}} \cdot \frac{n_2}{n_1}$$
- F-values inherit both of the χ^2's degrees of freedom, so that they have both a numerator and a denominator degree of freedom.

(You remember that a quotient – or a ratio–is the result of a division, right?)

Reminder: The F distribution

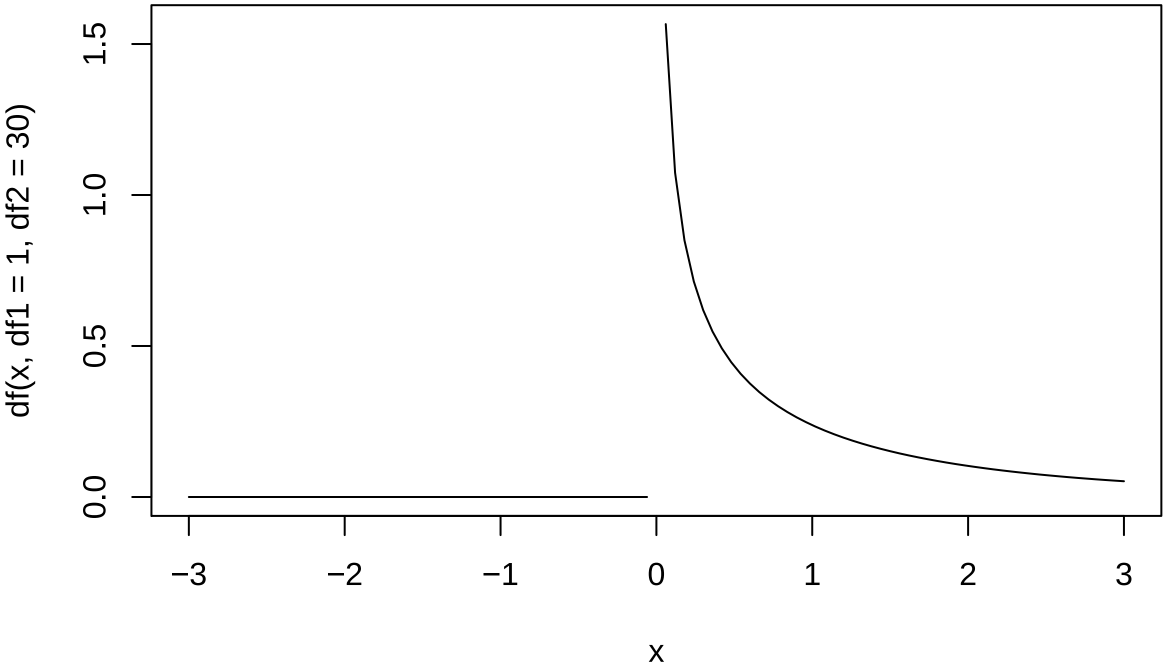
- It turns out that the ratio between model and error variance follows a specific distribution
- If there is no actual effect (i.e. the groups are just assigned at random) and
- As long as certain assumptions are valid (more on that later)
- This distribution is called the F-distribution
- Occasionally you will get a high MS_{model} simply by chance, but such occurrences are quite rare
- The F-distribution is the probability density function for different values of the variance ratio, i.e. the F-value.
- We essentially want to test if the F-value we get is extreme enough that it could only have occurred by chance 5% of the time (our α level)

Reminder: The F-distribution (2)

- Like the *t*-distribution, the shape of the F-distribution varies depending on sample size (degrees of freedom).
- Remember that F is the *ratio* of two variances (both chi^2-distributed).
- Because of this, the F distribtion has *two* degrees of freedom parameters
  - df_1, also called $df_{numerator}$
  - df_2, also called $df_{denominator}$

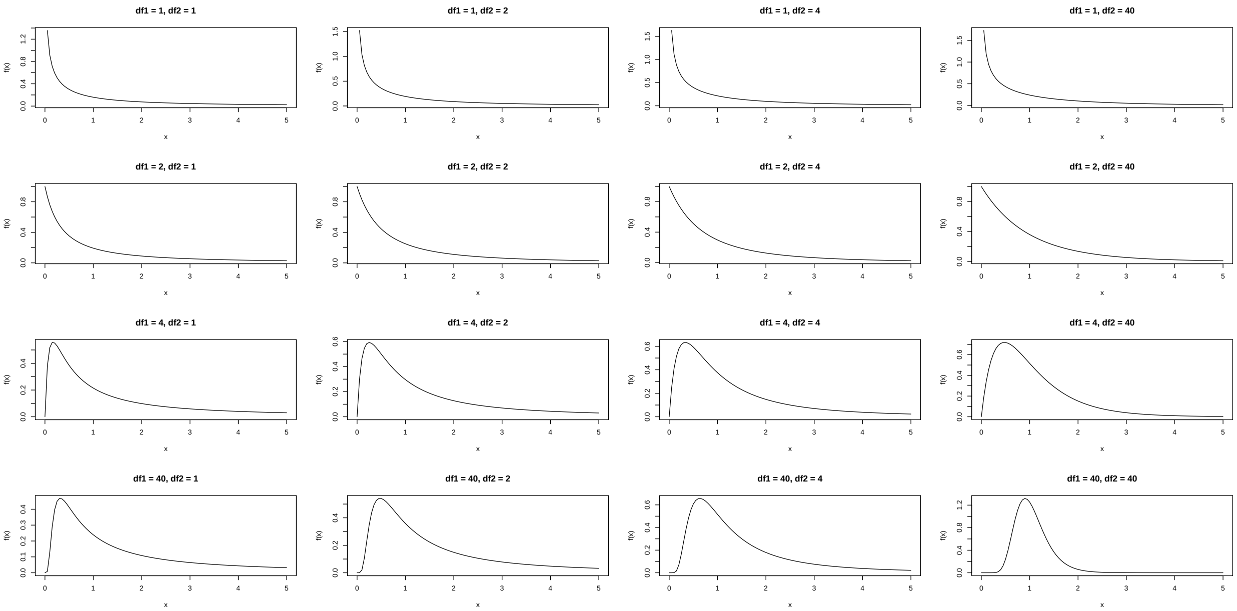
Plotting the F distribution (1)

- Let's take a look:



- F can't be negative - this makes sense: it's the quotient of two χ^2. You may remember that a square of a number can never be negative!

Plotting the F-distribution (2)



Effect size

- Just like for t-tests, we can get an estimate of effect sizes (called η^2, "eta-squared")
- $\eta^2 = \frac{SS_{model}}{SS_{total}}$
- η^2 is an estimate of the relationship between variance explained by the ANOVA model and total variance in the data
- You can use η^2 for power estimates with GPower.
- Compare this to Cohen's d, another estimate of effect size that we used for t-tests:
- $d = \frac{\bar{x_1} - \bar{x_2}}{s}$
- Cohen's d is an estimate of how large a difference in means is (in sample standard deviations)

Another way of thinking about ANOVA designs and contrasts

- Some people have asked me where the -1 and 1 contrast coding came from, and when we use it
- Very good question! This will take a couple of slides to explain
- Let's start out by re-writing the regression model for the one-way ANOVA so that it contains the group effects explicitly

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

- Here, μ is the (population) grand mean of the dependent variable, and α_1, α_2, $\ldots$, α_j are the (population) group effects.

Too many parameters

- There is a problem with this: we have more parameters than we have actual group means. For example, for $m = 3$, we have:

$$\mu_1 = \mu + \alpha_1$$

$$\mu_2 = \mu + \alpha_2$$

$$\mu_3 = \mu + \alpha_3$$

- We'd be estimating four coefficients (μ, α_1, α_2, and α_3), but we only have three group means.
- As we've seen in the previous lecture, you can only have $m - 1$ predictors. Any more than that, and they become perfectly collinear.
- In other words, the model is *overspecified* and *underdetermined.*

Reducing the number of parameters

- We could set α_1 to be 0. That way we get:

$$\mu_1 = \mu$$

$$\mu_2 = \mu + \alpha_2$$

$$\mu_3 = \mu + \alpha_3$$

- In this case, the first group mean μ_1 becomes the baseline, represented by the intercept parameter μ. This is the 0 vs. 1 dummy coding scheme!
- But other restrictions are possible, too.
  - We could constrain the α parameters so that they sum to 0:

$$\sum_{j=1}^{m} \alpha_j = 0$$

Parameters that sum to 0

- If we want parameters which sum to 0, the μ parameter becomes the mean of the group means (if the design is balanced, the grand mean):

$$\mu = \frac{\sum \mu_j}{m}$$

- The α_j parameters become the differences between the grand mean and the group means (for the first $m - 1$ group means):

$$\alpha_j = \mu_j - \mu$$

$$\mu_j = \mu + \alpha_j$$

- The last group mean is the difference between the grand mean and the sum of the other group means:

$$\mu_m = \mu - \sum_{j=1}^{m-1} \alpha_j$$

Deviation or sum contrasts

- These contrasts are called **deviation** contrasts in SPSS

S1	S2
1	0
0	1
-1	-1

- Here is how this works for three levels:

$$\mu_1 = \mu + 1 \times \alpha_1 + 0 \times \alpha_2 = \mu + \alpha_1$$

$$\mu_2 = \mu + 0 \times \alpha_1 + 1 \times \alpha_2 = \mu + \alpha_2$$

$$\mu_3 = \mu + -1 \times \alpha_1 + -1 \times \alpha_2 = \mu - \alpha_1 - \alpha_2$$

Making deviation or sum contrasts by hand

- Each contrast S_j is defined like this:
  - For each group ($j = 1, 2, 3...m$) set the row that coresponds to the current group j to 1
  - Set the row that corresponds to group m (the last group) to -1
  - Set all other values to 0.
- The F-values etc are just the same!

Using SPSS to perform a one-way ANOVA

- I took this data set from Andy Johnson, since he has made a great video explaining exactly how to analyse it.
- We are investigating the effect of swearing on pain tolerance (see Stephens et al., 2009)
- Three groups: continuous use of swear word, neutral word, or no word whilst hand in cold water (DV = time until participant can't stand the pain and pulls hand from water)
- Get the SPSS data file `Swearing and Pain Data.sav` from myBU.
- Watch Andy Johnson's video and follow along.

Multiway ANOVA

- This works just like the one-way ANOVA in terms of dummy/deviation/etc coding.
- For brevity, I'll just talk about the "textbook" version of the ANOVA here.
- What if we have two independent variables, A and B?
- We can still split the total variance into $SS_{total} = SS_{model} + SS_{error}$
- But now SS_{model} is composed of multiple terms: $SS_{model} = SS_A + SS_B + SS_{A \times B}$, so that $SS_{total} = SS_A + SS_B + SS_{A \times B} + SS_{Error}$
- Each of these terms has degrees of freedom: $df_{Total} = df_A + df_B + df_{A \times B} + df_{Error}$

Multiway ANOVA (2)

- For each term, you can compute mean squares and F values, e.g. $F_A = \frac{MS_A}{MS_{Error}}$
- What is $SS_{A \times B}$? It's the **interaction** between A and B
  - A *main effect* (A or B) is a difference between means
  - An *interaction* is a difference between differences

Multiway ANOVA (2)

- At this point, doing the analysis by hand is getting really tedious. Leave this to SPSS!
- As a little taster, I'll show you the formula for the total sums of squares:

$$SS_{total} = \sum_{j=1}^{m} \sum_{k=1}^{l} \sum_{i=1}^{n_{jk}} (x_{ijk} - \bar{x})^2,$$

  where j is the level of factor A, m is the total number of levels of factor A, k is the level of factor B, l is the total number of levels of factor B, i denotes the current observation number within its cell (i.e. the i^{th} observation within that combination of A and B), and n_{ijk} is the total number of observations within each cell.
- Nice and simple, right? This is why people started writing software to do this!

Multiway ANOVA (2)

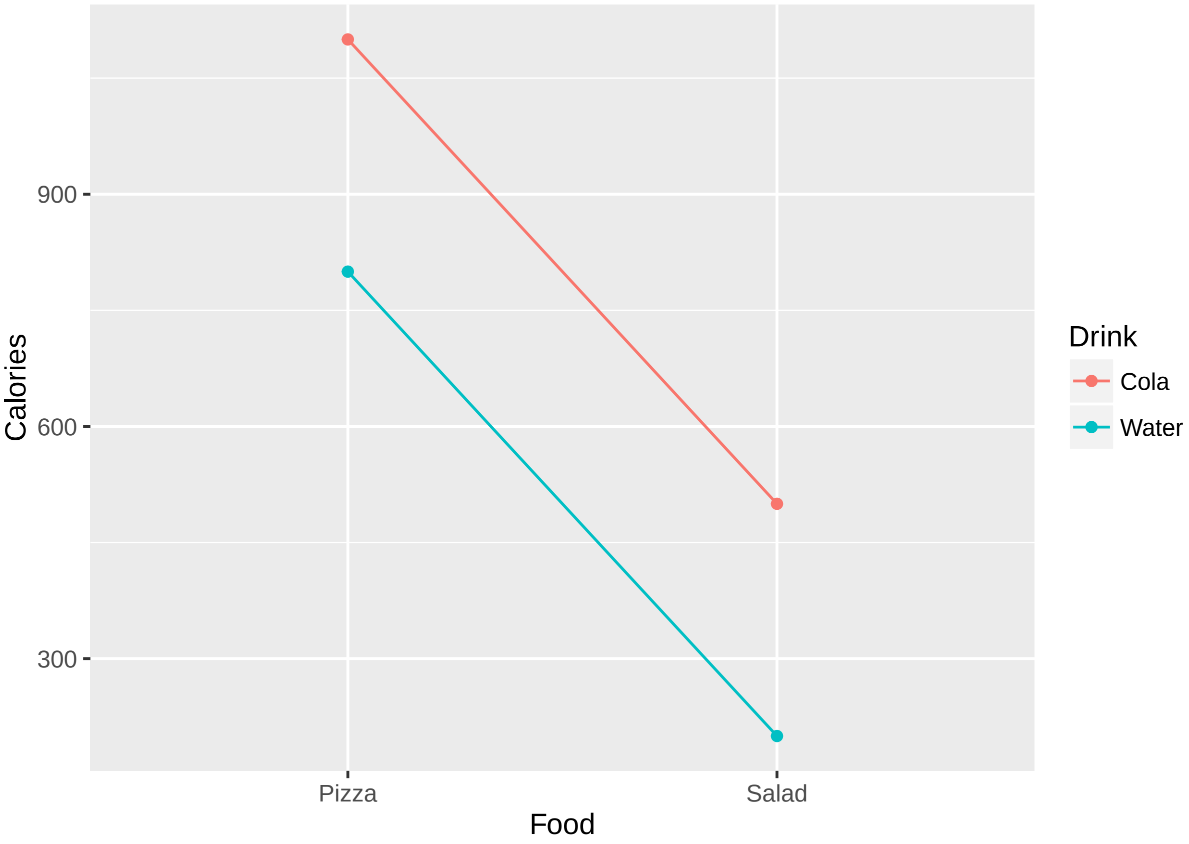
- How to compute the dfs:
  - For main effects, just like in the oneway ANOVA: $df_A = k_A - 1$, where k_A is the number of groups or *levels* of that variable
  - For interactions, it's the product of the dfs of the corresponding main effects:
  - $df_{A \times B} = df_A \cdot df_B$
  - Just as a reminder: df_{Total} is still $N - 1$, where $N = m * l * n_{ijk}$ is the total number of subjects or observations in your study (across all variables)
  - And as before, if you subtract all the other dfs from df_{Total}, you get df_{Error}
  - $df_{Error} = df_{Total} - df_A - df_B - df_{A \times B}$

Interactions

- Main effects are additive
- For example, this table shows the (fictional) total calories that you might have for lunch given two different food choices and two different drink choices:

Food	Drink	Calories
Pizza	Water	800
Pizza	Cola	1100
Salad	Water	200
Salad	Cola	500

Additive effects

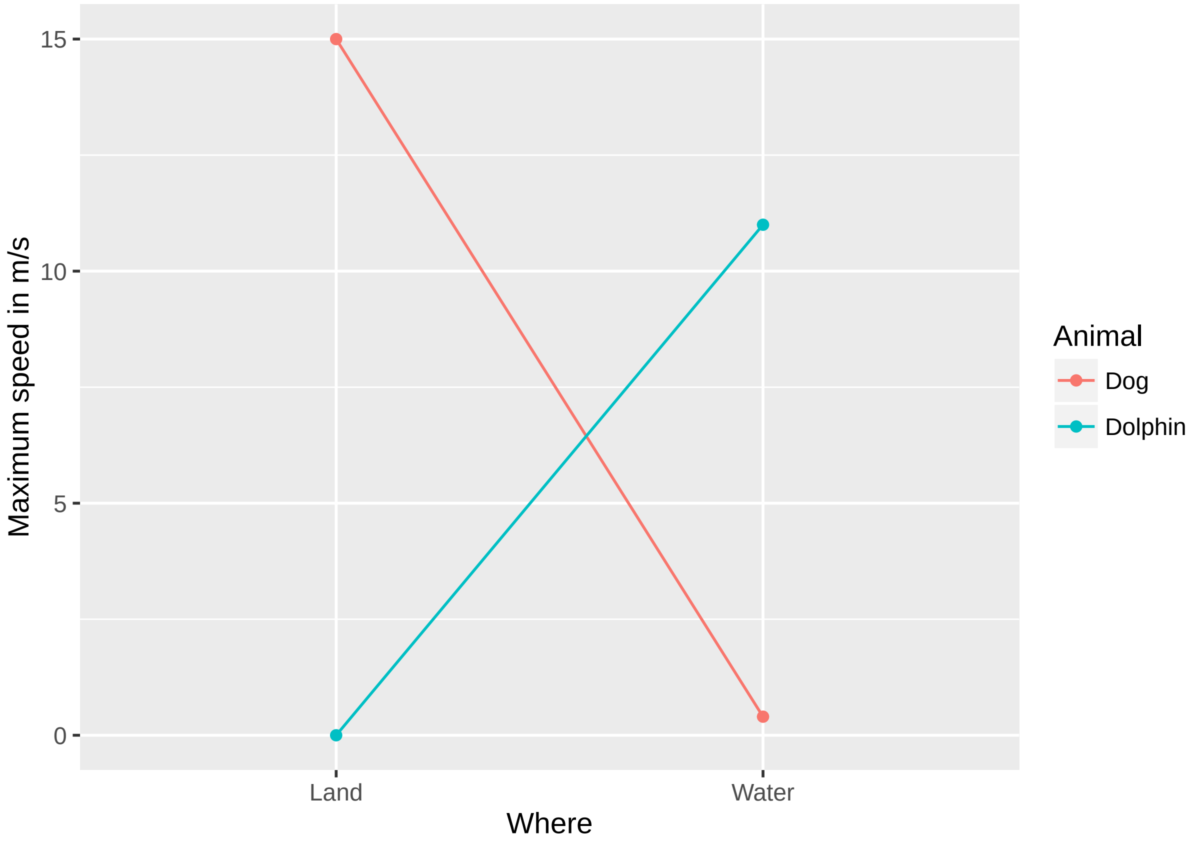


Non-additive effects

- Example: Animal and maximum movement speed (meters/s) on land and in water (mostly non-fictional, based on a very quick Wikipedia search)

Where	Animal	Speed
Land	Dog	15.0
Land	Dolphin	0.0
Water	Dog	0.4
Water	Dolphin	11.0

Non-additive effects (2)



- Crossover interaction

Marginal effects

- In the presence of a significant interaction, main effects are much harder to interpret
  - Better to call them marginal effects (although few people do, even in publications!)
- What does it mean that the marginal speed of a dolphin is 7.5 m/s (when averaging over the water and land conditions)?
  - Not much! The mean is nearly meaningless here…

Marginal effects (2)

- In some cases, you will still be interested in the marginal effects
  - For example, your anxiety treatment might differ in its effectiveness for male and female participants, but the marginal effects show that overall, everyone benefits from it at least a little.
  - Of course, if males get a little more anxious and females get a lot less anxious (a crossover interaction), the positive marginal effect still doesn't mean you should give males this treatment!

Types of sums of squares for the F-test

- Type I: Compare a model containing the predictor **and all other predictors** entered *so far* with a model only containing the predictors entered **so far** (order matters).
  - This is what the `anova` command in R does.
- Type II: Compare a model containing the predictor *along with* all the other predictors to a model containing *all the other predictors* **except** the predictor in question **and its interactions**.
- Type III: Compare a model containing the predictor *along with* all the other predictors to a model containing *all the other predictors* **except** the predictor in question, but **including its interactions**.
  - This is the equivalent of the *t*-tests.
  - This is the standard in SPSS.

Summary: Sums of square types

- ANOVA results (both classic ANOVA and regression model tests) can vary depending on which SS you use
  - Make sure that you know which ones you are using
  - If you are using SPSS, the answer is *probably* III.
- In standard ANOVA (with only discrete predictors), all SS types give the same result *as long as your design is balanced*
  - An unbalanced design will lead to differing sums of squares.
- In multiple regression, all SS types give the same result *as long as your predictor variables are not correlated*

Trying different contrasts

- Back to the dog data!

Mea	n number of objects known
Beagle	8.93
Border Collie	59.80
Terrier	12.33

Trying different contrasts (2)

- We can try some different contrast coding schemes to see how they work
- We can do this here because there are fake data and we know the actual means
- With real data, you need to plan your contrasts **before** you analyse your data (ideally, before you even collect them)
  - That's why they are called **planned** contrasts as opposed to **post hoc**.
- You can't even look at the means first!
- Otherwise, you're cheating. This is far worse than a small violation of normality or homoscedasticity!

Using contrasts in SPSS

- To be honest, SPSS is terrible with contrasts – it's really inconsistent
- You can specify your own contrasts using the Univariate ANOVA module, but you can't do that for the General Linear Model module
  - Instead, you have to pick a number of standard contrasts
- In theory, you could define your own contrasts, but it's a bit tricky

Standard contrasts

- Simple
  - Compares each level to the reference level, the intercept is the grand mean
- Deviation
  - Compares each level to the overall mean of the dependent variable (the reference level is not compared)
- Helmert
  - Compares each level to the mean of the subsequent ones
- Difference (reverse Helmert)
  - Compares each level to the mean of the previous ones
- Repeated (successive differences)
  - Compares each level to the subsequent level

Interpreting sum/deviation contrasts

- The intercept α is the grand mean of all the observations (28.33)
- β_1 is the difference between the grand mean and the mean of Beagle ($10 - 28.33 = -18.33$)
- β_2 is the difference between the grand mean and the mean of Border Collie ($60 - 28.33 = 31.67$)
- Terrier is never explicitly compared to the grand mean.
- In general: each level (except for the last level) is compared to the grand mean.

Applying difference (reverse Helmert) contrasts

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.02	0.880	30.7	0
breed1	25.43	1.078	23.6	0
breed2	-7.34	0.622	-11.8	0

Interpreting (reverse) Helmert contrasts

- The intercept α is the grand mean of all the observations (28.33)
- β_1 is half of the difference between the mean of Beagle and the mean of Border Collie ($(60 - 10)/2 = 25$)

- β_2 is half of the difference between the joint mean of Beagle and Border Collie and the mean of Terrier ($(15 - (60 + 10)/2)/2 = -10$)
- In general: each level is compared to the mean of the previous levels

Make your own contrasts?

- **DANGER**: If you apply your contrasts directly as dummy variables, you must use the **inverse** of your contrast matrix
- If your contrasts are not orthogonal, and you don't use the inverse of your matrix, you won't be comparing what you think you're comparing.
- If you don't know what this means, don't use your own contrasts until you do.
- For more background information on regression and linear models, see John Fox's book (Warning: it does involve matrix algebra). Check Chapter 8 for information about how the contrasts work and why you need to be careful.

How to do a multiway ANOVA

- Example data: Attractiveness, music, and alcohol
- Again, we have a video made by Andy Johnson on how this works in SPSS
- Download the SPSS data file (`Music, beer, and courting.sav`) and follow along with the video.

Make a means table

Music	Alcohol	N	Mean	SD
Generic sexy pop tunes	No alcohol	10	49.3	5.50
Generic sexy pop tunes	1 Pint of Stella	10	52.2	5.88
Generic sexy pop tunes	4 pints of Stella	10	70.4	5.04
Quiet	No alcohol	10	47.9	4.91
Quiet	1 Pint of Stella	10	52.3	5.77
Quiet	4 pints of Stella	10	62.3	5.36

Do the ANOVA

Effect	df_n	df_d	F	p	$p < 0.5$	η_P^2
Music	1	54	5.01	0.029	*	0.085
Alcohol	2	54	59.79	0.000	*	0.689
Music:Alcohol	2	54	3.24	0.047	*	0.107

- All three terms are significant.
  - Why η_P^2 instead of η^2? We want an estimate of the **partial** effect of each predictor. The standard η^2 is still the comparison between SS_{model} and SS_{error}, which doesn't tell us much.
- Partial η^2 (i.e. η_P^2) only takes into account the SS for our effect and SS_{error}, e.g. for Factor A: $\eta_P^2 = \frac{SS_A}{SS_A + SS_{error}}$

Assumption tests:

- Levene's test:

df_n	df_d	SS_n	SS_d	F	p	$p < 0.5$
5	54	5.8	412	0.152	0.979	

- No problems with homogeneity of variances

Pairwise comparisons

- Easiest way: Use Tukey's HSD (I'll explain that with SPSS, since the output is a bit complicated)

Writing it up

A 2-factor (2x3) independent samples ANOVA was conducted where the first factor represents music exposure (quiet and music) and the second factor represents alcohol condition (no alcohol, 1-pint, and 4-pints). There was no evidence for a violation of the homogeneity of variance assumption. Overall, the ANOVA method should be robust to the slight deviation from normality that was observed. Attractiveness ratings were significantly higher with music exposure, $F(1{,}54) = 5.01$, $p = .03$, $\eta^2_G = .09$. The main effect of alcohol was also significant, $F(2{,}54) = 59.79$, $p < .001$, $\eta^2_P = .69$. A post hoc test (Tukey's HSD) indicated that participants who drank 4 pints of beer rated attractiveness as significantly higher than participants who had no alcohol ($p < .001$) and one pint ($p < .001$). There was no difference between the no alcohol and 1-pint groups ($p = .11$).

Writing it up (2)

The music by alcohol interaction was also significant, $F(2{,}54) = 3.24$, $p = .047$, $\eta^2_G = .11$. This indicates that alcohol had different effects under conditions of music exposure. Specifically, post-hoc comparisons showed that with no alcohol there was no difference in attractiveness ratings for music (M = 49.30, SD = 5.50) and no music (M = 47.90, SD = 4.91). Similarly, following 1-pint there was no difference in attractiveness ratings for music (M = 52.20, SD = 5.88) and no music (M = 52.30, SD = 5.77). However, following 4-pints attractiveness ratings were higher with music (M = 70.40, SD = 5.04) than without music (M = 62.30, SD = 5.36). This effect was significant ($p = .018$).