CHAPTER 8

Finding differences between several groups

Sign language, linguistic relativity and ANOVA

What you will learn from this chapter:

This chapter introduces ANOVA (analysis of variance), a special case of linear regression with binary or categorical independent variables. This method is widely used in experimental linguistics, when the researcher compares several groups of experimental objects that undergo different treatments. In this chapter you will learn several types of ANOVA: one-way ANOVA with one factor as an independent variable, factorial ANOVA with two or more categorical independent variables, and repeated-measures and mixed ANOVA. The methods are illustrated by three case studies. The first two focus on grammatical features of an emergent sign language. The third case study deals with cross-linguistic differences in time conceptualization, which are interpreted as evidence in favour of the linguistic relativity hypothesis.

8.1 What is ANOVA?

In Chapter 5, we discussed the *t*-test and its non-parametric equivalents, which can be used for comparison of two groups. ANOVA (analysis of variance) is a family of methods that enable one to investigate the differences between any number of groups, which can be specified by one or more categorical variables. Although some researchers are not aware of this fact, ANOVA can be regarded as a special case of regression analysis. Due to historical reasons, ANOVA has been used mostly in experimental research, whereas regression has been mostly employed in correlational research based on observations (e.g. corpus data), surveys or questionnaires.

A typical application of ANOVA is when the researcher wants to find if there are differences between several groups, which undergo different experimental treatments. For example, imagine that you want to test the usefulness of video subtitles for learning English as a foreign language. In that case, you might compare three different conditions. In one group, subtitles are in English. In a second group, subtitles are in the students' native language. A third group is the control group where no subtitles are used at all.

In this chapter you will learn how to perform three major types of ANOVA:

- **independent one-way ANOVA**, which is used for comparison of three and more groups, as in the example above. The situations when it can be used are similar to the ones when the independent *t*-test is applied. However, as has been said already, the latter is limited to two groups only.
- **independent factorial ANOVA**, which involves two or more categorical independent variables. Imagine that native speakers of two different languages are exposed to two different experimental conditions, e.g. two types of stimuli, and their reaction times are measured. Factorial ANOVA allows one to measure the individual effects of the variables, as well as their possible interactions. In this example, an interaction may emerge, for instance, if the speakers of one language react to the different types of stimuli differently, and the speakers of the other language react to all stimuli in the same way.
- **repeated-measures and mixed ANOVA**, which are used for the same purposes as the above-mentioned types of ANOVA in situations when observations are not independent, e.g. when every subject is tested more than once, or every stimulus is presented more than once. These types of ANOVA test the differences between groups while taking into account individual variation.

There are two important concepts related to ANOVA, namely, **within-subject** and **between-group design**. In between-group design, which is also called between-subject design, different groups of subjects are assigned to different experimental conditions. For example, one group may be the one where the subjects are administered a new drug. The other group may be the control group, which receives a placebo. In this example, the type of medical treatment is a between-group variable. Independent one-way and factorial ANOVAs contain only between-group variables.

In within-subject design, the same subjects are tested in several experimental conditions. Imagine you conduct a lexical decision experiment. You ask your subjects to decide whether a word does or does not exist in the language. The stimuli are a mixture of real and nonce words. Every subject reacts to both types of stimuli. The type of stimuli is considered then a within-subject variable. One can also say that this variable is tested within one subject. If all subjects are tested in all possible conditions, this means that all variables are within-subject variables. In that case, you will need **repeated-measures ANOVA**. If your design contains both within-subject and between-group variables, you will have to perform **mixed ANOVA**. ANOVA with within-subject variables can also be one-way, two-way (factorial), and so on, depending on the number of independent variables in the model.

8.2 Motion events in Nicaraguan Sign Language: Independent one-way ANOVA

8.2.1 Theoretical background and data

For this case study you will need a number of add-on packages, which should be first installed, if they have not been installed previously, and then loaded.

```
> install.packages(c("car", "coin", "nparcomp"))
> library(Rling); library(car); library(coin); library(nparcomp)
```

This case study focuses on Nicaraguan Sign Language (NSL). It is a very young language, which has been developing over the last thirty years. Many of its originators are still around today, and researchers can observe how the language has changed since its earliest days. In a series of studies, e.g. Senghas & Coppola (2001) and Senghas et al. (2004), Ann Senghas and her collaborators have demonstrated that NSL has been developing 'from analog to digital', that is, from holistic continuous expressions to discrete elements, which resemble words and morphemes. For example, the early expressions of motion in NSL were holistic movements, which express the path and manner simultaneously. As an illustration, consider a rolling event. One can express it gesturally in a single holistic movement, which combines the manner and the path. The manner can be expressed as a wiggling movement, and the path as a movement to the speaker's right or left. Senghas and her colleagues, however, observed that younger NSL speakers prefer to express the manner and the path sequentially, in two separate signs. For a rolling event, the manner would be expressed by circling, and the path would be indicated as a sign that shows the trajectory to the signer's side. This finding is not surprising. According to Talmy's well-known theory (1985), languages typically represent motion events by separate linguistic elements that correspond to path and manner (e.g. *roll* [*MANNER*] *down* [*PATH*] *the hill*).

For illustration, we will explore a fictitious dataset that resembles the data in Senghas et al. (2004). The data are available in the data frame `NSL` in the `Rling` package. The data frame contains two columns. The first one, *MannerPath*, displays the proportions of separate expressions of manner and path by NSL signers when they were asked to describe a set of different motion events. Each participant was shown an animated cartoon that included those events, and was asked to narrate this story to a peer. The proportions range from 0 (no separate expressions, only holistic analog expressions) to 1 (only separate expressions, no holistic analog expression). The second column, *Cohort*, shows one of the three generational groups to which each signer belongs: Cohort 1 includes the very first NSL signers, who joined the community and learnt to sign in the late 1970s – early 1980s; Cohort 2, which arrived in the mid- to late 1980s; and Cohort 3, which arrived from 1990 to the moment of data collection. Each cohort is represented by nine participants. Although the cohorts are represented by numbers, this variable is a factor.

```
> data(NSL)
> head(NSL)
Manner  Path   Cohort
1       0.44   1
2       0.24   1
3       0.41   1
4       0.03   1
5       0.32   1
6       0.18   1
> str(NSL)
'data.frame': 27 obs. of 2 variables:
$ MannerPath: num 0.44 0.24 0.41 0.03 0.32 0.18 0.25 0.42 0.17 0.77
…
$ Cohort: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 2 …
```

The earlier a signer joined the community, the more 'ancient' level of NSL he or she represents. We expect thus that the mean proportion of separate expressions of manner and path will be higher for the more recent Cohorts than for the earlier ones. We will use ANOVA to test this hypothesis. *MannerPath* will be the response variable, and *Cohort* will be a between-group exploratory variable. Since each subject was tested only once, the design is independent.

8.2.2 Exploring the data

We will begin with a graphical exploration and will create a box plot that represents the distribution of scores in each cohort:

```
> boxplot(NSL$MannerPath ~ NSL$Cohort, xlab = "Cohort", ylab =
"Proportion of separate expressions", main = "Path and motion in
NSL")
```

Figure 8.1 displays the graph (for information about how to make a `ggplot2` version, see Chapter 3; see also Chapter 5 about how to create a bar plot with 95% confidence intervals). The box plot shows clearly that the first cohort of NSL signers use separate expressions of manner and path much less frequently than the second and the third cohorts. However, we do not see a clear difference between the second and third cohorts.

Next, we will compute the mean proportions of separate expressions in each cohort with the help of `tapply()`. This useful function can compute a statistic of interest for several groups at the same time:

```
> tapply(NSL$MannerPath, NSL$Cohort, mean)
1           2           3
0.2733333   0.7555556   0.7344444
```

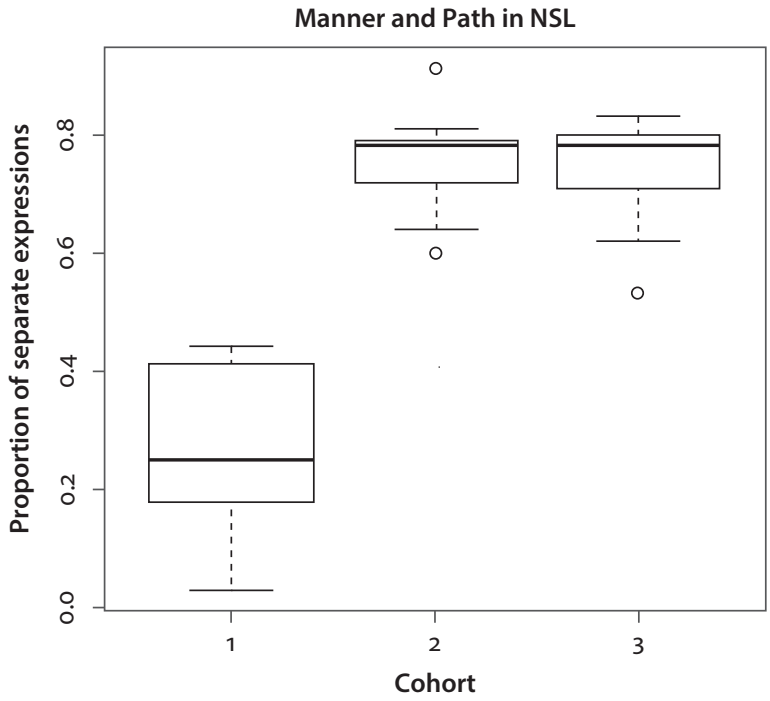


Figure 8.1. Box plots of proportions of separate expression of manner and path in three cohorts of Nicaraguan Sign Language

The first argument of `tapply()` is the quantitative variable, `NSL$MannerPath`. It is followed by the grouping variable, `NSL$Cohort`. Finally, we specify the function that computes the statistic, `mean`. The result is a one-dimensional table with three means. Again, we see that the mean proportion of separate expressions in the first cohort is much smaller than those in the second and third cohorts. Interestingly, there is even a small decrease in the third cohort. To see whether these differences are statistically significant, one needs one-way ANOVA.

Why bother with ANOVA?

Indeed, why not perform several *t*-tests instead? The answer to this question lies in the expression 'inflation of surprise' (Baayen 2008: 114). To put it simply, when one performs multiple comparisons on the same data, the probability that he or she finds some surprising results goes up. As the number of hypotheses to test increases, the chances to commit a Type I error (i.e. reject the null hypothesis when it is in fact true) increase, as well. This is the main reason for using ANOVA instead of multiple *t*-tests.

8.2.3 Assumptions of one-way parametric ANOVA

The alternative hypothesis of one-way ANOVA is that at least two group means are different from one another. The test is therefore non-directional. To obtain reliable results with the help of traditional parametric ANOVA, the following assumptions should be met:

- *The observations in the samples are independent from one another.*
- *The response variable is at least interval-scaled.*
- *Each sample is drawn from a normally distributed population and/or the sample sizes are equal* (Field et al. 2012: 413).
- *The variance is homogeneous, or homoscedastic* (cf. Chapter 6). In other words, the variances of the populations represented by the groups should be equal. The opposite of homogeneity is heterogeneity, or heteroscedasticity.

If any of these assumptions is saliently violated, other versions of ANOVA should be used. They will be discussed in the next subsection.

Are these assumptions met in the data? There are no reasons to assume that the observations are dependent. The response variable is ratio-scaled. It is difficult to evaluate the normality, since the samples are very small. However, the sample sizes are equal. For didactic purposes, we provide the code that enables one to run the Shapiro test three times simultaneously and obtain the *p*-value for each of the three groups:

```
> aggregate(MannerPath ~ Cohort, data = NSL, function(x) shapiro.
test(x)$p.value)
Cohort  MannerPath
1       1 0.5614787
2       2 0.5210358
3       3 0.1266828
```

None of the *p*-values, which are shown in the right-hand column, is smaller than 0.05. Therefore, we do not have reasons to believe that the assumption of normality is not met. However, these results should be taken with caution because the chances that the Shapiro test will detect non-normality depend on the sample size.

Finally, homogeneity of variance, or homoscedasticity, can be tested with the help of the Levene test, which is available as `leveneTest()` in the package `car`.

```
> leveneTest(MannerPath ~ Cohort, data = NSL) # equivalent to
leveneTest(…, center = median), which provides a more robust test
than the original Levene test with mean as the center.

Levene's Test for Homogeneity of Variance (center = median)
       Df  F value  Pr(>F)
group  2   0.9207   0.4118
       24
```

The null hypothesis of the Levene test is that the groups have equal variances. Since the *p*-value is greater than 0.05, the null hypothesis of equal variances cannot be rejected.

There exist many alternatives to the Levene test. One of them is the Fligner – Killeen median test, which is robust to departures from normality (Conover et al. 1981). The test can be carried out as follows:

```
> fligner.test(MannerPath ~ Cohort, data = NSL)

     Fligner-Killeen test of homogeneity of variances

data: MannerPath by Cohort
Fligner-Killeen:med chi-squared = 2.3505, df = 2, p-value = 0.3087
```

The large *p*-value again suggests that we cannot discard the null hypothesis that the variance is homogeneous across the groups.

8.2.4 Performing parametric one-way ANOVA

Since we have not found any violations of the test assumptions, we can perform the parametric one-way ANOVA. There are two options. First, we can use the general `lm()` function for linear regression:

```
> NSL.lm <- lm(MannerPath ~ Cohort, data = NSL)
> summary(NSL.lm)

Call:
lm(formula = MannerPath ~ Cohort, data = NSL)

Residuals:
Min       1Q        Median   3Q       Max
-0.24333  -0.06444  0.02444  0.06000  0.16667

Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  0.27333   0.03727     7.334     1.42e-07 ***
Cohort2      0.48222   0.05271     9.149     2.71e-09 ***
Cohort3      0.46111   0.05271     8.748     6.26e-09 ***
--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1118 on 24 degrees of freedom
Multiple R-squared: 0.8167,      Adjusted R-squared: 0.8014
F-statistic: 53.46 on 2 and 24 DF, p-value: 1.439e-09
```

The second option is to use `aov()`, which produces the traditional ANOVA output:

```
> NSL.aov <- aov(MannerPath ~ Cohort, data = NSL)
> summary(NSL.aov)
            Df   Sum Sq   Mean Sq    F value   Pr(>F)
Cohort      2    1.337    0.6684     53.47     1.44e-09 ***
Residuals   24   0.300    0.0125
--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The main test statistic in both versions is the *F*-ratio. In ANOVA, it is usually interpreted as the ratio of the average between-group variability and the average within-group variability. Between-group variability measures variance of group means. In the `aov()` output, you can find it in the `Cohort` row, in the column `Mean Sq`, which represents the mean between-group variability. This value is 0.6684, and it is equal to the sum of squares (`Sum Sq`) 1.337 divided by two degrees of freedom (`DF`). Within-group variability is variability that can be attributed to chance factors. It is also called error or residual variability. The mean residual variability in our example is 0.0125 (0.3 divided by 24), as one can see in the `Residuals` row. The greater the average between-group variability in comparison with the average within-group variability, the greater the *F*-ratio. In this example, it is $0.6684/0.0125 \approx 53.47$, with 2 and 24 degrees of freedom. The alternative hypothesis of the test is that there is at least one pair of groups with different means, or, in other words, that at least one difference between groups means is different from zero. Since $p < 0.001$, we can conclude that at least one pair of groups has different means.

8.2.5 Alternative tests

There exist a number of alternative tests, which can be used when one or more assumptions are violated. There is still controversy about the use of some tests, but the following guidelines seem to represent the consensus:

- `oneway.test()` can be used when the variance is not homogeneous, but the other assumptions hold. To perform the test, one can follow the example:

```
> oneway.test(MannerPath ~ Cohort, data = NSL)

One-way analysis of means (not assuming equal
variances)

data: MannerPath and Cohort
F = 41.2291, num df = 2.000, denom df = 15.648,
p-value = 5.787e-07
```

- The Kruskal-Wallis one-way ANOVA by ranks can be used when the response variable is on the ordinal scale or when the samples come from markedly non-normal

distributions. However, at least one of the two assumptions should still be met: homogeneous variances of the ranks and/or equal sample sizes (Sheskin 2011: 1002). The test is an extension of the Mann-Whitney test (*U*-test), which was introduced in Chapter 5. It can also be used on interval- or ratio-scaled data when one needs to reduce the impact of outliers. Consider an example:

```
> kruskal.test(MannerPath ~ Cohort, data = NSL)

Kruskal-Wallis rank sum test

data: MannerPath by Cohort
Kruskal-Wallis chi-squared = 17.4208, df = 2, p-value
= 0.0001649
```

– Non-parametric ANOVA based on bootstrapping or permutation (see Chapters 7 and 14) can be used when all assumptions concerning the distribution are violated, except for the assumption of independence. There are many packages that offer this option. For example, one can use ANOVA based on permutation that is implemented in `oneway_test()` from the package `coin`:

```
> oneway_test(MannerPath ~ Cohort, data = NSL, distribution =
approximate(B = 9999))

Approximative K-Sample Permutation Test

data: MannerPath by Cohort (1, 2, 3)
maxT = 4.6046, p-value < 2.2e-16
```

One can also use Wilcox' (2005) functions based on the comparison of medians and trimmed means with bootstrap (Field et al. 2012: 441–443).

– repeated-measures and mixed ANOVA should be used when observations are dependent (see Section 8.4).

8.2.6 Post hoc tests

The *F*-ratio test is an omnibus test. That is, it can tell us that there is some significant difference somewhere, but it does not say where exactly. To find out which groups differ significantly, one can perform a post hoc test. There is a variety of post hoc tests for one-way ANOVA. In this section, we will consider only two, the parametric Tukey Honest Significant Differences test and a non-parametric multiple comparison test.

The Tukey Honest Significant Differences (HSD) test is available as `TukeyHSD()`. This function requires an `aov` object. The function returns the adjusted 'honest' *p*-values, hence the name of the test, 'Honest Significant Differences'. The test is quite robust to violations of the normality assumption. However, there are two assumptions that should

be met: homogeneous variances and independence of observations. The test can be performed as follows:

```
> TukeyHSD(NSL.aov)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = MannerPath ~ Cohort, data = NSL)

$Cohort
      diff          lwr          upr         p  adj
2-1   0.48222222    0.3505939    0.6138506   0.00000
3-1   0.46111111    0.3294828    0.5927394   0.00000
3-2   -0.02111111   -0.1527394   0.1105172   0.91568
```

The function returns the differences between the group means for three pairs of groups. The greatest difference is between Cohorts 2 and 1. It is positive because the mean response in Cohort 2 is greater than that in Cohort 1. It is also statistically significant, judging from a very small *p*-value, which can be found in the rightmost column. The results for Cohorts 3 and 1 are very similar. The negative difference between Cohorts 3 and 2 indicates that the mean response in Cohort 3 is smaller than that in Cohort 2. However, this difference is not statistically significant.

In addition, the function returns the 95% confidence intervals of the differences between the groups. The column `lwr` gives us the lower end points of the intervals, whereas the `upr` column tells about the upper boundary. If a confidence interval includes 0, we cannot be sure that the groups are truly different. The differences and their confidence intervals can be visualized as follows: (see the result in Figure 8.2)

```
> plot(TukeyHSD(NSL.aov))
```

If the assumption of equal variances is violated, one can use a non-parametric test. Here we will discuss `nparcomp()` from the eponymous package:

```
> npar <- nparcomp(MannerPath ~ Cohort, data = NSL, type = "Tukey")
#---Nonparametric Multiple Comparisons for relative contrast
effects---#
[output omitted]

> npar$Analysis
Comparison   Estimator   Lower   Upper   Statistic    p.Value
1 p(1, 2)    0.999       0.996   1.000   11.3376072   0.0000000
2 p(1, 3)    0.999       0.996   1.000   11.3376072   0.0000000
3 p(2, 3)    0.463       0.176   0.776   -0.2546148   0.9918819
```

The output provides a lot of details, most importantly, which groups are compared (the leftmost column) and the *p*-value (the rightmost column). The results confirm that

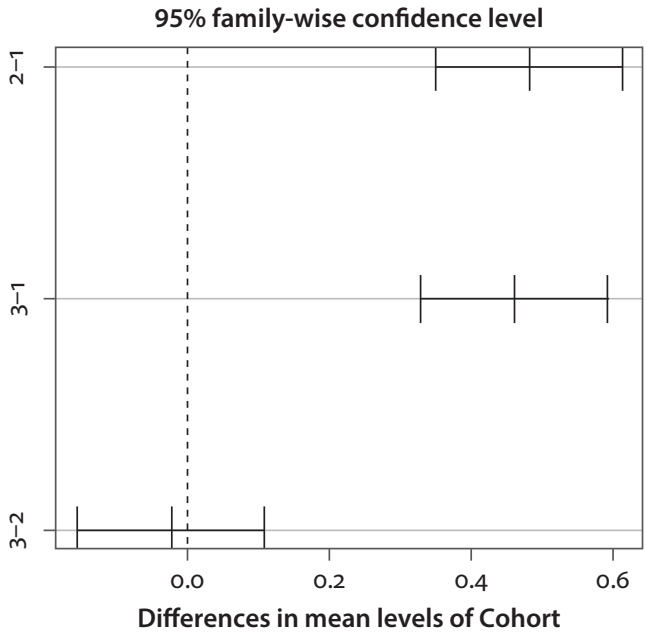


Figure 8.2. Confidence intervals of differences between group means

the first Cohort is different from the second and third ones, but the difference between the second and third Cohorts is not statistically significant.

More on post hoc tests

Other useful functions include `pairwise.t.test()`, which offers a range of corrections for inflation of surprise, such as the famous Bonferroni correction, which is, however, less widely used now in comparison with more modern corrections, such as the Holm method. See also `glht()` in the package `multcomp`. The statistical details for different correction methods can be found in Sheskin (2011: 903–924), and a more concise and practical overview is provided in Larson-Hall (2010: 282).

8.3 Development of spatial modulations in Nicaraguan Sign Language: Independent factorial (two-way) ANOVA

8.3.1 The data and hypothesis

In this case study we will use two add-on packages, which should be installed (unless you have already done so) and then loaded:

```
> install.packages("car")
> library(Rling); library(car)
```

As the previous case study, this section focuses on Nicaraguan Sign Language (NSL). See the background information about this language in Section 8.2. The research question comes from the paper by Senghas & Coppola (2001), who compared newly emerging spatial devices in two different cohorts of NSL learners: those who joined the community before and after 1983 (the median year). The cohort served thus as a between-group variable. The other factor was the age group of the signer when he or she joined the community. The dependent variable is the average number of spatial modulations per verb for every signer. This is a particular type of spatial modulation that indicates 'shared reference'. Most signs are produced neutrally in the central position in front of the signer's chest. However, sign languages also include spatial modulations, when signs are produced in a non-neutral location, e.g. on the right or left from the signer. In developed sign languages, such spatial modulations convey deictic, locative or temporal information. In particular, one can use non-neutral locations to show that different actions were related to one and the same referent, e.g. the verbs in *he came, saw and conquered* describe actions of the same agent.

In the experiment, the signers watched a short cartoon and then signed the story to a peer. Their use of spatial modulation was videotaped and later analysed by the researchers. The hypothesis of the study is that shared reference modulations have been emerging as a grammatical device in the new generation of NSL signers. This is why shared reference modulations (per verb) are expected to be higher in the output of the second cohort signers. One can also expect differences depending on the age of exposure.

The imitation dataset is called `sharedref` and is available in `Rling`. The data frame contains 48 observations, which correspond to individual subjects. The first variable, *mod*, represents the average number of shared reference modulations per verb. The second variable, *age*, indicates the age when the signer joined the community. There were three age groups: 'early' (before 6 years and a half), 'middle' (between 6 years and a half and 10 years old), and 'late' (after 10). Finally, the variable *cohort* indicates which cohort each signer belongs to. Note that the variable is a factor, although its levels are represented by the numbers 1 and 2.

```
> data(sharedref)
> head(sharedref)
    mod   age     cohort
1   0.75  early   1
2   0.85  early   1
3   0.93  early   1
4   0.80  early   1
5   1.24  early   2
6   1.38  early   2
```

Since both predictors are between-group variables and there are no within-subject variables because every person was tested once, the appropriate choice is independent factorial ANOVA.

8.3.2 Descriptive statistics for different groups and interaction plot

One can compute average means of shared reference expressions in age group by cohort with the help of `aggregate()`:

```
> ref <- aggregate(mod ~ age + cohort, data = sharedref, FUN = mean)
> ref
    age      cohort   mod
1   early    1        0.8325
2   middle   1        0.6800
3   late     1        0.3700
4   early    2        1.3550
5   middle   2        1.2200
6   late     2        0.4400
```

The group means suggest that the earlier a signer joined the community, the higher the average frequency of shared reference expressions. In addition, the means of each age group in the second cohort are higher than the corresponding means in the first cohort. These results meet our expectations. However, it is difficult to say using the numbers alone whether there is an interaction between age and cohort. To investigate that, we will create an interaction plot:

```
> interaction.plot(ref$age, ref$cohort, ref$mod)
```

The result is shown in Figure 8.3. There is an interaction. Namely, the difference between the cohorts almost disappears for those who joined the community relatively late. Are the differences between the cohorts and age groups, as well as their interaction statistically significant? This is a question for factorial ANOVA.

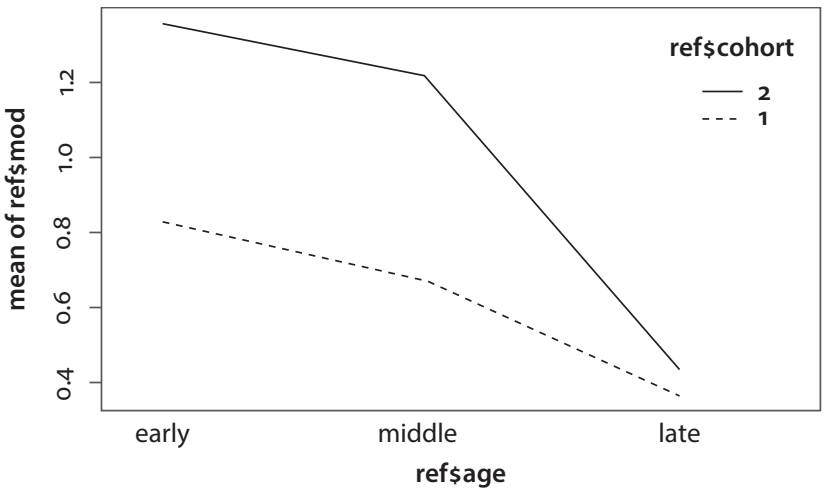


Figure 8.3. Interaction plot of use of shared reference expressions across age groups and cohorts of NSL

8.3.3 Assumptions of parametric factorial ANOVA

The assumptions of parametric factorial ANOVA are the same as those of one-way ANOVA (see Section 8.2.3). Since the design is balanced, one does not have to be very strict about the normality. However, it is useful to know how to perform the Shapiro test simultaneously on several groups created by cross-tabulated factors. Again, a convenient option is to use the function `aggregate()`:

```
> aggregate(mod ~ age + cohort, data = sharedref, function(x)
shapiro.test(x)$p.value)
   age      cohort   mod
1  early    1        0.305382517
2  middle   1        0.611790881
3  late     1        0.975408765
4  early    2        0.710437680
5  middle   2        0.424433620
6  late     2        0.002067517
```

The *p*-values in the last column are greater than 0.05, except for the last group (late learners, Cohort 2). However, ANOVA is quite robust with regard to some non-normality, especially if the design is balanced and the other assumptions are met.

To test the assumption of homogeneity of variance, one can use again the Fligner – Killeen median test, which is robust with regard to non-normal data:

```
> fligner.test(mod ~ interaction(age, cohort), data = sharedref)

     Fligner-Killeen test of homogeneity of variances

data: mod by interaction(age, cohort)
Fligner-Killeen:med chi-squared = 6.1292, df = 5, p-value =
0.2939
```

The Levene test, which was described in the previous case study, returns a similar result. Again, there are no indications that this assumption is violated, so we can move on to performing a parametric factorial ANOVA.

8.3.4 ANOVA and orthogonal contrasts

In order to evaluate the contribution of each variable and interaction term(s), one has to compute sums of squares, which are required for computing the *F*-score (see Section 8.2.4). There are different types of sums of squares: Type I, Type II, Type III and Type IV. The most important thing for this discussion is that Type III sums of squares evaluate all effects in the model taking into account all other effects in the model. This means that the effect of *age* will be evaluated after the effects of *cohort* and the interaction between *cohort* and *age*, and so on. Type III sums of squares are also the most reliable when the sample sizes are not equal.

Type III sums of squares can only be used when the predictors are coded with orthogonal contrasts. As the reader may remember from Chapter 7, there exist different contrasts: treatment, sum, Helmert, etc. For example, the default treatment contrasts involve comparison of all levels of a categorical variable with the reference level. What was not mentioned explicitly in the previous chapter, however, is the fact that all contrasts can be represented as numerical values, or weights. For example, if a binary variable has the treatment coding, the reference level by default has the weight of 0, and the other level has the weight of 1. The variable is then represented as the contrast, or dummy variable with the weights (0, 1). If there are three levels, then we have two dummy variables, or contrasts, with the weights (0, 1, 0) for the first contrast and (0, 0, 1) for the second contrast , and so on.

Contrasts are considered **orthogonal** if the products of their coefficients sum up to zero. For two levels, this can be done by coding one level as –1 and the other one as 1. The sum is –1 + 1 = 0. If a variable has three levels, the orthogonal contrasts can be (–2, 1, 1) and (0, –1, 1). The sum of products is $-2 \times 0 + 1 \times (-1) + 1 \times 1 = 0 - 1 + 1 = 0$. For four levels, there are different options, e.g. (3, –1, –1, –1), (0, 2, –1, –1) and (0, 0, 1, –1). The sum of products is $3 \times 0 \times 0 + (-1) \times 2 \times 0 + (-1) \times (-1) \times 1 + (-1) \times (-1) \times (-1) = 0 + 0 + 1 - 1 = 0$. See a summary in Table 8.1.

Now let us assign orthogonal contrasts to the factors. First, *age* with three levels can be represented by two contrasts with three weights in each:

```
> contrasts(sharedref$age) <- cbind(c(-2, 1, 1), c(0, -1, 1))
> sharedref$age
[output omitted]
attr(,"contrasts")
        [,1]  [,2]
early   -2    0
middle  1     -1
late    1     1
Levels: early middle late
```

Table 8.1 Orthogonal contrasts for factors with two, three and four levels

A factor with two levels

Group	Contrast weights	Product
1	–1	–1
2	1	1
Total	**0**	**0**

A factor with three levels

Group	Weights of Contrast 1	Weights of Contrast 2	Product
1	–2	0	0
2	1	–1	–1
3	1	1	1
Total	**0**	**0**	**0**

A factor with four levels

Group	Weights of Contrast1	Weights of Contrast 2	Weights of Contrast 3	Product
A	3	0	0	0
B	–1	2	0	0
C	–1	–1	1	1
D	–1	–1	–1	–1
Total	**0**	**0**	**0**	**0**

Orthogonal contrasts for *cohort*, which has two levels, look as follows (note that in this case the coding is identical with sum contrasts, except for the order of weights):

```
> contrasts(sharedref$cohort) <- c(-1, 1)
> sharedref$cohort
[output omitted]
attr(,"contrasts")
    [,1]
1   -1
2   1
Levels: 1 2
```

After these preparations, we can perform ANOVA with the help of `aov()`. The formula below specifies two main effects *age* and *cohort*, and their interaction. Alternatively, you can write `mod ~ age + cohort + age:cohort`.

```
> sharedref.aov <- aov(mod ~ age*cohort, data = sharedref)
```

Next, the ANOVA table based on Type III sums of squares can be obtained by using `Anova()` – mind the capital letter! – from the package `car`:

```
> Anova(sharedref.aov, type = "III")
Anova Table (Type III tests)

Response: mod
             Sum Sq   Df   F value    Pr(>F)
(Intercept)  31.981   1    7447.676   < 2.2e-16 ***
age          4.224    2    491.884    < 2.2e-16 ***
cohort       1.710    1    398.243    < 2.2e-16 ***
age:cohort   0.568    2    66.132     1.054e-13 ***
Residuals    0.180    42
--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *p*-values in the rightmost column show that both independent variables and the interaction term have highly significant effects.

8.3.5 Alternative tests

If some of the assumptions are not met, there are a few options:

- Similar to `oneway.test()` in one-way ANOVA, which provides a correction for heterogeneous variance, one can use White's adjustment for factorial ANOVA when the assumption of homogeneity of variance is violated. This option is available in `Anova()` in the package `car`, for example:

```
> Anova(sharedref.aov, type = "III", white.adjust = TRUE)
Analysis of Deviance Table (Type III tests)

Response: mod
             Df   F          Pr(>F)
(Intercept)  1    6516.716   < 2.2e-16 ***
age          2    401.274    < 2.2e-16 ***
cohort       1    348.463    < 2.2e-16 ***
age:cohort   2    60.725     4.047e-13 ***
Residuals    42
--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- In all situations when the normality and homogeneity assumptions are violated, one can fit a linear regression model and perform bootstrap validation of the confidence intervals around the estimated coefficients (see Section 7.2.8 in Chapter 7).
- Again, if the observations are not independent, one should use repeated-measures or mixed ANOVA (see Section 8.4).

8.3.6 Post hoc tests

As in one-way ANOVA, one can use post hoc tests for factors with more than two levels to find out which groups are different and which are not. Although it is does not make much sense to perform post hoc tests on the main effects of a model in the presence of a significant interaction, we will demonstrate how one can compare the group means of *age*. For example, one could perform the Tukey Honest Significant Differences test of this predictor as follows:

```
> TukeyHSD(sharedref.aov, "age")
Tukey multiple comparisons of means
     95% family-wise confidence level

Fit: aov(formula = mod ~ age *cohort, data = sharedref)

$age
               diff       lwr          upr          p adj
middle-early  -0.14375  -0.2000365  -0.0874635   6e-07
late-early    -0.68875  -0.7450365  -0.6324635   0e+00
late-middle   -0.54500  -0.6012865  -0.4887135   0e+00
```

The confidence intervals and the adjusted *p*-values tell us that all differences between the age groups are significant. Of course, since the independent variable interacts with the other one, this example serves only as an illustration.

When reporting the results of a factorial ANOVA, it is convenient to summarize the data with the help of `model.tables()`. If you add `type = "means"`, the function will return the means of the response for all groups and their intersections, including the grand mean for the entire dataset. By default, the function returns 'effects', or coefficients, which are more difficult to interpret. You can also compute standard errors for differences of means by adding `se = TRUE`.

```
> model.tables(sharedref.aov, type = "means")
Tables of means
Grand mean

0.81625

age
age
early middle  late
1.0938 0.9500 0.4050

cohort
cohort
1       2
0.6275  1.0050

age:cohort
```

```
      cohort
         age      1   2
early    0.8325   1.3550
middle   0.6800   1.2200
late     0.3700   0.4400
```

To summarize, the case study has demonstrated that language can be created by sequences of child learners who are not exposed to a developed language. The second cohort of NSL signers did not reproduce the language; rather, it changed the language in the process, making it more grammaticized. The difference between the cohorts is present, however, only for children who joined the community relatively early (before the age of 10). These findings are important for our understanding of language dynamics in general, from transformation of pidgins into creoles, to different processes of slow incremental language change.

8.4 Do native English and native Mandarin Chinese speakers conceptualize time differently? Repeated-measures and mixed ANOVA (mixed GLM method)

8.4.1 The data and hypothesis

To reproduce the code from this case study, you will need several add-on packages, which should be installed and loaded.

```
> install.packages(c("gplots", "ggplot2", "nlme"))
> library(Rling); library(gplots); library(ggplot2); library(nlme)
```

This case study deals with metaphoric conceptualization of time. In English, as in many other languages, time is construed as FUTURE IS IN FRONT OF EGO and PAST IS IN BACK OF EGO, as in the examples *The exam is three weeks ahead* and *The worst is behind us*. In Aymara language, the pattern is reverse: the past is in front, and the future is behind. Consider the following expressions (examples from Núñez & Sweetser 2006):

(1) *nayra* *mara*
eye/sight/front year
'last year'

(2) *qhipa* *mara* *-na*
back/behind year at
'in the next year'

Aymara speakers also produce frontward gestures when they speak about the past and backward gestures when they speak about the future. Núñez & Sweetser (2006) argue that Aymara speakers conceptualize time in a static way. When one stands, the objects in front are visible and therefore known, and the objects behind are invisible and therefore unknown. The future is unknown and unseen, as if it were located behind the speaker.

The past is known and 'observable', so it is 'located' in front. In contrast, English and other Western languages use a dynamic mapping. When a person walks, the past is what is left behind and known, and the future is the unknown in front. However, in spite of these differences, both Aymara and English share the underlying metaphor KNOWLEDGE IS SEEING, which represents the future as unseen, and the past as seen.

Such observations rekindle the old debate about the hypothesis of linguistic relativity that goes back to Benjamin Whorf. Does language shape the way we think and, if yes, to what extent? Although only few linguists nowadays would support the original (strong) version of the hypothesis, according to which language fully determines thought, many researchers would agree with a weak version: language determines thinking because we get used to pay more attention to the distinctions coded in language, and less attention to other possible distinctions. Such differences are sociocultural, and are acquired gradually. Consider English and Korean. The languages exhibit substantial differences in the spatial domain. For instance, in Korean, containment events are subdivided into putting things into containers that fit tightly (*kkita*), e.g. putting a book in a book cover, and putting objects into containers that fit loosely (*nehta*), e.g. putting a book in a large box. Bowerman & Choi (2003) studied spatial distinctions made by English and Korean infants. In the prelinguistic stage, both English and Korean infants were sensitive to this distinction. However, this sensitivity diminishes over time, as children master their mother tongue. Trying to fit the expressions they hear to the referential situations they observe, they gradually acquire the relevant distinctions that are specific to their language, like the tight vs. loose fit distinction. Potentially, we are capable to make the distinctions that are not salient in our language, but we normally do not have to worry about them.

In this case study we will use simulation data to reproduce Boroditsky's (2001) findings with regard to conceptualization of time by native English and native Chinese speakers. Boroditsky used a priming experiment to test if conceptualization of time depends on mother tongue. As was mentioned above, English has predominantly horizontal conceptualization of time, where the past is behind, and the future is in front. In Mandarin Chinese, however, the dominant metaphor seems to be vertical: future events are *xià* 'down' and past events are *shàng* 'up'. An experiment showed that native English speakers processed time expressions faster when they were primed by a horizontal array of objects, whereas native Chinese speakers processed time expressions in English faster when the primes contained vertically arranged objects.

The dataset, which contains simulation data of Boroditsky's experiment, is called `time_exper`. It can be found in the package `Rling`. There are four variables: the subject ID, language (English or Chinese), type of prime (horizontal or vertical) and reaction times in milliseconds:

```
> data(time_exper)
> str(time_exper)
'data.frame': 200 obs. of 4 variables:
```

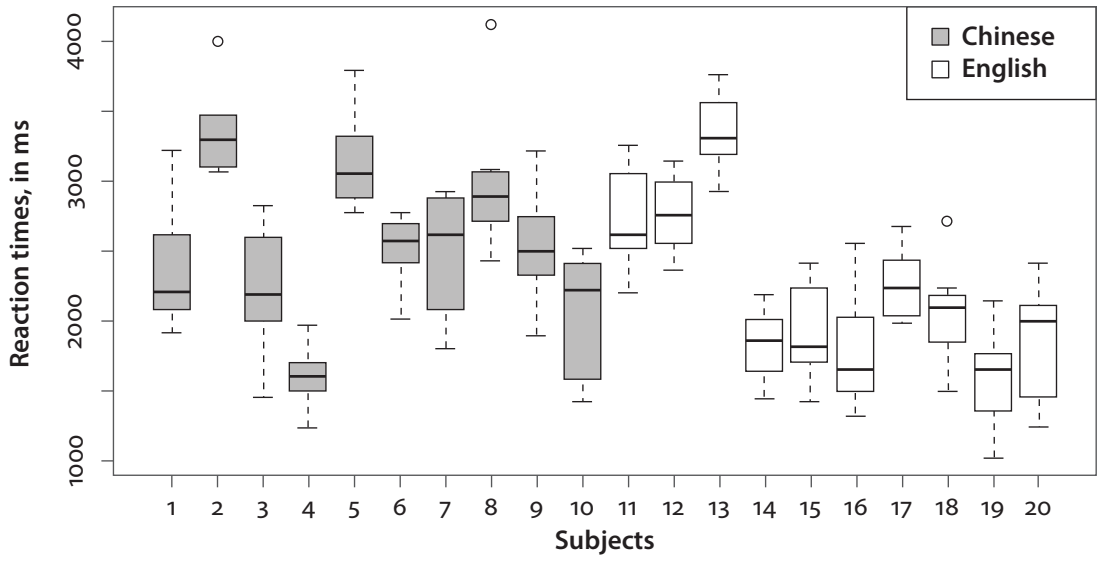
```
$ Subj: Factor w/ 20 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 …
$ Lang: Factor w/ 2 levels "CH","EN": 1 1 1 1 1 1 1 1 1 1 …
$ Prime: Factor w/ 2 levels "Horiz","Vert": 1 1 2 1 2 1 2 1 2 2 …
$ rt: num 3221 2079 1940 2655 1913 …
```

The question is whether there is difference between native English and native Mandarin speakers in their reaction times depending on the type of prime. One can expect native English speakers to react to a stimulus (an English non-metaphoric time expression) faster after horizontal primes, and native Mandarin speakers to react faster after vertical primes. However, we should also take into consideration that the observations in the dataset are not independent. Each subject participated in ten trials and gave ten responses. This situation is common in experimental studies because it would be impractical to collect only one observation per subject. As a result, reaction times may be influenced by speakers' individual characteristics. Some subjects may be faster to respond to the stimuli, and some may be slower. This study is thus an example of mixed design, with *Lang* as a between-group variable, and *Prime* as a within-subject variable nested within the *Subj* factor. In other words, scores for the type of prime can be found within each subject.

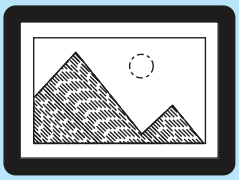
Individual variation between the subjects can be examined with the help of a box plot with different colours. Since the first ten subjects are Mandarin Chinese speakers, and the last ten are English speakers, we can use the following code:

```
> boxplot(rt ~ Subj, data = time_exper, xlab = "Subjects", ylab =
"Reaction times, in ms", col = c(rep("grey", 10), rep("white", 10)))
> legend("topright", c("Chinese", "English"), fill = c("grey", "white"))
```

The resulting plot is displayed in Figure 8.4. One can see that there is substantial individual variation among the native speakers of each language. Will the independent variables and their interaction be still significant after this individual variation has been taken into account? To answer this question, one needs to perform a mixed ANOVA.



Fugure 8.4. Box plot with different colours representing different groups



How to make a box plot with different colours for different groups with `ggplot2`

The code below allows one to reproduce the plot in Figure 8.4a.

```
> ggplot(time_exper, aes(x = Subj, y = rt, fill = Lang)) + geom_
boxplot() + labs(x = "Subjects", y = "Reaction times, in ms") +
scale_fill_grey(start = 0.5, end = 1)
```

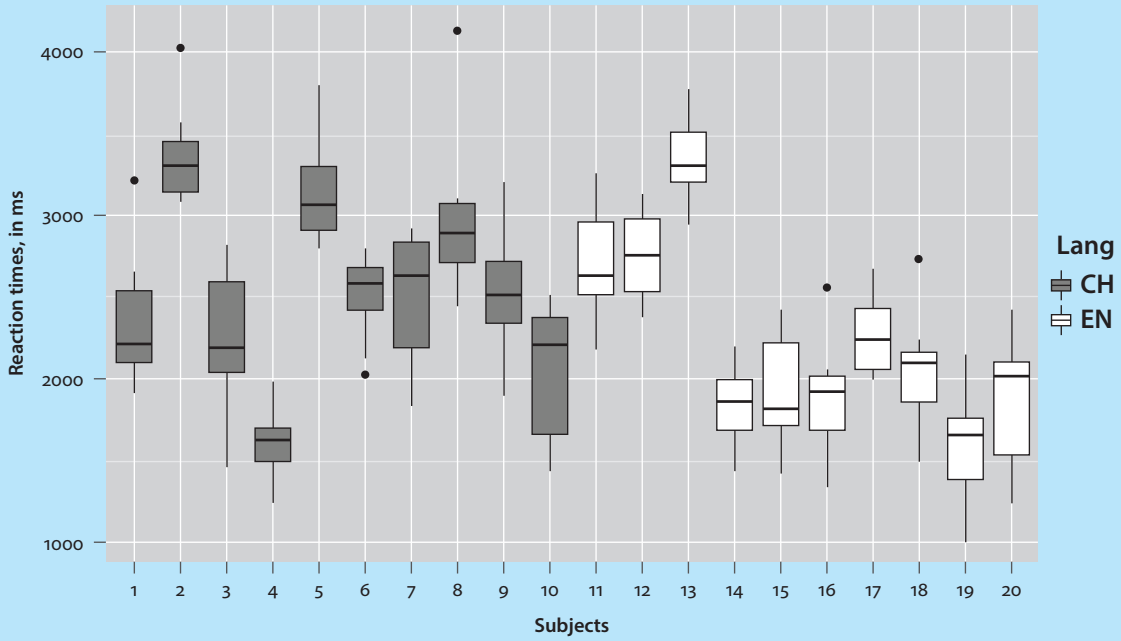


Figure 8.4a. A `ggplot2` version of box plot in Figure 8.4

8.4.2 Fitting a mixed ANOVA with the help of mixed GLM

The method that will be introduced below is essentially the same for repeated-measures and mixed designs. It involves mixed Generalized Linear Models (GLM), which are often called mixed models. The main characteristic of these models is that they contain both **fixed** effects and **random** effects. Fixed effects are 'normal' explanatory variables, whose effect should be measured. Random effects are sampled randomly from the population, for example, individuals in an experiment, lexical stimuli in a lexical decision task, etc. In a way, this is 'noise' that should be filtered out.

Mixed GLM models are superior to the traditional repeated-measures and mixed ANOVA approaches. One of the main advantages is the fact that mixed models, unlike ANOVA, do not require that all measurements should be present for all subjects. One does not have to discard all measurements related to one subject if one value is missing. In experimental research, this represents a serious advantage. Second, mixed models are part of a larger framework, which provides numerous other extensions based on the same principles, such as non-linear mixed models, complex multilevel models with several hierarchical levels, etc. Finally, mixed models, unlike repeated-measures and mixed ANOVA, do not require the assumption of sphericity to be met, which may be a challenging task (cf. Field et al. 2012: 551–554).

In this case study we will use the function `lme()` from the package `nlme`. To add a random effect, one should specify `random = ~1|YourRandomEffect`. Such effects are called **random intercepts**. This means that individual adjustments are made to the intercept for each individual (subject, stimulus, place where the experiment was performed, and other randomly selected items). Such individual adjustments are also called Best Linear Unbiased Predictors (BLUPs). In mixed ANOVA, one has to specify both the subjects and the nested variable: `random = ~1|Subj/Prime`. Another type of random effects is **random slopes**, which are added when the effect of one or more predictors varies across the individuals. Random slopes are not discussed here, but see numerous examples in Baayen (2008: Ch. 7).

This section introduces the procedure which was proposed in Field et al. (2012) for both repeated-measures and mixed designs. First, one fits a baseline model with the intercept only, which contains no predictors. The baseline model can be fit as follows:

```
> m0 <- lme(rt ~ 1, random = ~ 1|Subj/Prime, data = time_exper,
method = "ML")
```

It is important to remember is that there are two available methods of fitting a mixed model, maximum likelihood (ML) and restricted maximum likelihood (REML). When one wants to compare different models with different number of parameters, as will be demonstrated below, it is recommended to use maximum likelihood estimation. The default method is REML, so it is necessary to add `method = "ML"`.[1]

Next, we add more and more parameters and compare the models with the help of ANOVA. For example, the following code can be used to add *Lang*:

```
> m1 <- lme(rt ~ Lang, random = ~ 1|Subj/Prime, data = time_exper,
method = "ML")
```

1. The ML method of estimation is also considered preferable when one is more interested in obtaining accurate estimates of fixed effect parameters than in estimation of random variation (Field et al. 2012: 879).

However, it is more convenient to use the function `update()` for adding new parameters to the model, including the interaction term:

```
> m1 <- update(m0,.~. + Lang)
> m2 <- update(m1,.~. + Prime)
> m3 <- update(m2,.~. + Lang:Prime)
```

Finally, ANOVA can be used to establish which terms have been useful when added sequentially:

```
> anova(m0, m1, m2, m3)
     Model df  AIC      BIC      logLik    Test   L.Ratio    p-value
m0   1     4  2967.152 2980.346 -1479.576
m1   2     5  2967.434 2983.926 -1478.717 1 vs 2 1.717997   0.1900
m2   3     6  2968.936 2988.726 -1478.468 2 vs 3 0.498424   0.4802
m3   4     7  2945.424 2968.512 -1465.712 3 vs 4 25.512348  <.0001
```

The *p*-values are based on the difference in the log-likelihoods (`L.Ratio`) in the adjacent models. The log-likelihoods are shown in the column `logLik` (more exactly, each log-likelihood value is multiplied by –2). Other useful statistics are AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). As the reader may remember from Chapter 7, AIC is a goodness-of-fit measure for comparison of models with different number of parameters. It penalizes a model for having too many predictors. BIC is similar to AIC, but it is considered to be more efficient when the sample is large and the number of parameters is small (Field et al. 2012: 868). The smaller AIC and BIC, the better. Interestingly, neither the language (`m1`) nor the type of prime (`m2`) brought a significant improvement individually. However, their interaction in the final model is highly significant, as can be seen from a very low *p*-value.

The final model with all predictors and the interactions looks as follows:

```
> summary(m3)
Linear mixed-effects model fit by maximum likelihood
Data: time_exper
AIC       BIC       logLik
2945.424  2968.512  -1465.712

Random effects:
Formula: ~1 | Subj
     (Intercept)
StdDev: 496.187

Formula: ~1 | Prime %in% Subj
     (Intercept) Residual
StdDev: 0.0297535 313.0908
```

```
Fixed effects: rt ~ Lang *Prime
                  Value    Std.Error  DF   t-value    p-value
(Intercept)      2674.48  164.69104  160 16.239378  0.0000
LangEN           -568.20  232.90830  18   -2.439587  0.0253
PrimeVert        -313.54  63.25390   18   -4.956849  0.0001
LangEN:PrimeVert 530.26   89.45452   18   5.927705   0.0000
Correlation:
            (Intr)   LangEN  PrmVrt
LangEN     -0.707
PrimeVert  -0.192   0.136
LangEN:PrimeVert 0.136 -0.192 -0.707

Standardized Within-Group Residuals:
Min          Q1           Med         Q3          Max
-2.06200493  -0.67236516  0.01431141 0.62678708  3.28816200

Number of Observations: 200
Number of Groups:
     Subj Prime   %in% Subj
     20           40
```

The summary shows that all coefficients in the model are significant, although we have just seen that the individual contributions of *Lang* and *Prime* added sequentially were not significant. There is a simple explanation. As you may remember from Chapter 7, in the presence of an interaction, the interacting terms shown in the table are no longer main effects. They represent the estimates for the combinations of the specified level with the reference level of the interacting variable (note that we use the default treatment contrasts). The coefficient of `LangEN` thus shows the difference between the reaction times of the English native speakers and those of the Chinese native speakers after a horizontal prime (the reference level). This difference is significant. Similarly, the coefficient of `PrimeVertical` represents the significant difference between the reaction times after the vertical and horizontal primes, but only for the native speakers of Chinese.

To obtain the 95% confidence intervals of the coefficients, one can use `intervals()` from the package `nlme`. It also returns the confidence intervals of the standard deviations of the random effects provided in the summary above.

```
> intervals(m3) # equivalent to intervals(…, level = 0.95)
Approximate 95% confidence intervals

Fixed effects:
                  lower       est.     upper
(Intercept)      2352.5003   2674.48  2996.45968
LangEN           -1052.6042  -568.20  -83.79575
PrimeVert        -445.0959   -313.54  -181.98412
LangEN:PrimeVert  344.2119    530.26   716.30811
```

```
attr(,"label")
[1] "Fixed effects:"

Random Effects:
Level: Subj
                  lower    est.     upper
sd((Intercept))  359.492  496.187  684.8597
Level: Prime
                   lower          est.        upper
sd((Intercept))  2.997266e-21  0.0297535  2.953595e+17
Within-group standard error:
lower      est.      upper
282.3632   313.0908  347.1623
```

8.4.3 Post hoc tests

One normally does not perform traditional post hoc tests when fitting a GLM. Instead, one can use the estimates in the model and their confidence intervals to see which group means are different. Of course, if a factor has more than two levels, one cannot estimate the differences between all pairs of values. In that case, one can use planned contrasts, as shown in Field et al. (2012: 617–618) to focus only on the comparisons that are theoretically relevant. Needless to say, it does not make sense to report the global differences between the group means in the presence of a significant interaction, as in this case.

8.5 Summary

This chapter has introduced one-way, two-way (factorial), as well as repeated-measures and mixed ANOVAs. One-way ANOVA is used to compare the means of three and more groups. It can be regarded as an extension of the *t*-test. Factorial ANOVA measures the effect of two and more categorical variables on the response, as well as their possible interactions. Finally, repeated-measures and mixed ANOVAs are used for the same purposes as one-way and factorial ANOVA in the situations when the assumption of independence is not met and the model contains within-subject variables. Note that you can use linear regression, which was described in Chapter 7, to carry out independent one-way and factorial ANOVA, as well.

Writing up the results of ANOVA

When reporting the results of a one-way ANOVA with independent observations, one has to report the *F*-ratio, the degrees of freedom and the corresponding *p*-value. For example, you could write: "We observe a significant effect of cohort in our model, $F(2, 24) = 53.47$, $p < 0.001$". Note that the second number of the degrees of freedom (24) comes from the residual component of variance (see the last line of the `aov()` or `Anova()` output). In case of independent factorial ANOVA, you should describe all effects in the model, including the interaction, e.g. for *age* in the factorial ANOVA, the numbers would look as follows: $F(2, 18) = 214.353$, $p < 0.001$.

The results of a post hoc test should be reported, as well. It is crucial to be very specific about the type of the post hoc test and report the *p*-value and the difference *d* (if available), as well as the mean and the standard deviation (or standard error) for each group compared. For example, one can write about the one-way ANOVA case study: "the second cohort of the NSL users ($M = 0.756$, $SD = 0.09$) produced significantly more separate expressions of manner and path than the first cohort ($M = 0.273$, $SD = 0.137$), $d = 0.482$, $p < 0.001$, according to the post hoc Tukey Honest Significant Differences test".

Reporting the results of repeated-measurements or mixed ANOVA, which was fitted with the help of `lme()`, one should mention the statistic from the `L.ratio` column, which is distributed like the χ^2 statistic, the degrees of freedom and the *p*-value. For example, one could write, "We found a significant interaction of language and the type of prime, $\chi^2(1) = 25.5$, $p < 0.001$". The number of degrees of freedom is 1, which can be computed as a product of the degrees of freedom of each interacting factor minus one: $(2 - 1) \times (2 - 1) = 1$.

More on ANOVA

There exist many other varieties of ANOVA, e.g. analysis of covariance (ANCOVA), when a model contains one or more quantitative variables, which are called covariates. Another member of the ANOVA family is MANOVA, multivariate analysis of variance, which can deal with several response variables, for example, in a situation when experimental subjects are administered different tests of language proficiency after experimental training. A combination of ANCOVA and MANOVA is called MANCOVA. A well-known method of performing one-way repeated-measures ANOVA is Friedman's test (see `?friedman.test` for more details). A discussion of all these methods is far beyond the scope of this introductory textbook. If you would like to learn more about mixed models, see numerous linguistic examples in Chapter 7 of Baayen (2008). To incorporate crossed (non-nested) random effects (e.g. subjects and items) in repeated-measures and mixed ANOVA, check the package `lme4`.