Chapter 4
Analytical statistics

The most important questions of life are,
for the most part, really only questions of probability.
Pierre-Simon Laplace
(from <http://www-rohan.sdsu.edu/%7Emalouf/>)

In the previous chapter, I discussed a variety of descriptive statistics. In this chapter, I will now explain how these measures and others are used in the domain of hypothesis-testing. For example, in Section 3.1 I explained how you compute a measure of central tendency (such as a mean) or a measure of dispersion (such as a standard deviation) for a particular sample. In this chapter, you will see how you test whether such a mean or such a standard deviation deviates significantly from a known mean or standard deviation or the mean or standard deviation of a second sample. I will assume that you have downloaded the data files from the companion website. Before we get started, let me remind you once again that in your own data your nominal/categorical variables should ideally always be coded with meaningful character strings so that R recognizes them as factors when reading in the data from a file.

1. Distributions and frequencies

In this section, I will illustrate how to test whether distributions and frequencies from one sample differ significantly from a known distribution (cf. Section 4.1.1) or from another sample (cf. Section 4.1.2). In both sections, we begin with variables from the interval/ratio level of measurement and then proceed to lower levels of measurement.

1.1. Distribution fitting

1.1.1. One dep. variable (ratio-scaled)

In this section, I will discuss how you compare whether the distribution of one dependent interval-/ratio-scaled variable is significantly different from a known distribution. I will restrict my attention to one of the most frequent cases, the situation where you test whether a variable is normally distributed (because as was mentioned above in Section 1.3.4.2, many statistical require a normal distribution so you must be able to do this test).

We will deal with an example from the first language acquisition of tense and aspect in Russian. Simplifying a bit here, one general tendency that is often observed is a relatively robust correlation between past tense and perfective aspect as well as non-past tenses and imperfective aspect. Such a correlation can be quantified with Cramer’s *V* values (cf. Stoll and Gries, forthc., and Section 4.2.1 below). Let us assume you studied how this association – the Cramer’s *V* values – changes for one child over time. Let us further assume you had 117 recordings for this child, computed a Cramer’s *V* value for each one, and now you want to see whether these are normally distributed. This scenario involves

- a dependent interval/ratio-scaled variable called TenseAspect, consisting of the Cramer’s *V* values;
- no independent variable because you are not testing whether the distribution of the variable TenseAspect is influenced by, or correlated with, something else.

You can test for normality in several ways. The test we will use is the Shapiro-Wilk test, which does not really have any assumptions other than ratio-scaled data and involves the following procedure:

Procedure
Formulating the hypotheses
Inspecting a graph
Computing the test statistic *W* and the probability of error *p*

We begin with the hypotheses:

H_0: The data points are normally distributed; $W = 1$.
H_1: The data points are not normally distributed; $W \neq 1$.

First, you load the data from <C:/_sflwr/_inputfiles/04-1-1-1_tense-aspect.txt> and create a graph; the code for the left panel is shown below but you can also generate the right panel using the kind of code discussed in Section 3.1.1.5.

```
> RussianTensAps<-
      read.table(choose.files(),·header=T,·sep="\t",·
      comment.char="",·quote="")¶
> attach(RussianTensAps)¶
> hist(TENSE_ASPECT,·xlim=c(0,·1),·xlab="Tense-
      Aspect·correlation",·
      ylab="Frequency",·main="")·#·left·panel¶
```

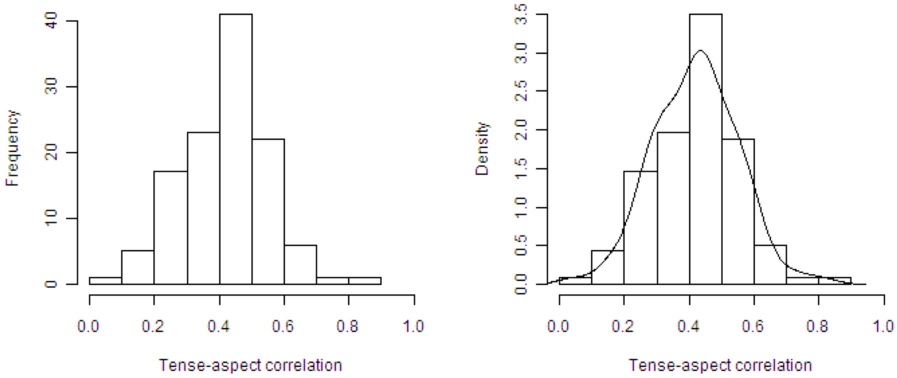

Figure 40. Histogram of the Cramer's *V* values reflecting the strengths of the tense-aspect correlations

At first glance, this looks very much like a normal distribution, but of course you must do a real test. The Shapiro-Wilk test is rather cumbersome to compute semi-manually, which is why its manual computation will not be discussed here (unlike nearly all other tests). In R, however, the computation could not be easier. The relevant function is called `shapiro.test` and it only requires one argument, the vector to be tested:

```
> shapiro.test(TENSE_ASPECT)¶
········Shapiro-Wilk·normality·test
data:··TENSE_ASPECT
W·=·0.9942,·p-value·=·0.9132
```

What does this mean? This simple output teaches an important lesson: Usually, you want to obtain a significant result, i.e., a *p*-value that is smaller than 0.05 because this allows you to accept the alternative hypothesis. Here, however, you may actually welcome an insignificant result because normally-distributed variables are often easier to handle. The reason for this is again the logic underlying the falsification paradigm. When $p < 0.05$, you reject the null hypothesis H_0 and accept the alternative hypothesis H_1. But here you 'want' H_0 to be true because H_0 states that the data are normally distributed. You obtained a *p*-value of 0.9132, which means you cannot reject H_0 and, thus, consider the data to be normally distributed. You would therefore summarize this result in the results section of your paper as follows: “According to a Shapiro-Wilk test, the distribution of this child’s Cramer’s *V* values measuring the tense-aspect correlation does not deviate significantly from normality: $W = 0.9942$; $p = 0.9132$.” (In parentheses or after a colon you usually mention all statistics that helped you decide whether or not to accept the alternative hypothesis H_1.)

Recommendation(s) for further study

- the function `shapiroTest` (from the `library(fBasics)`) as an alternative to the above function
- the function `ks.test` (in `ks.test(a.vector,·"pnorm",·mean=mean(a.vector),·sd=sd(a.vector))`) or the function `ksnormTest` (from the `library(fBasics)`) as an alternative to the above function. This test, the Kolmogorov-Smirnov test for distribution-fitting, is less conservative and more flexible than the Shapiro-Wilk-Test, since it can not only test for normality and can also be applied to vectors with more than 5000 data points. We will discuss a variant of this test below
- as alternatives to the above functions, the functions `jarqueberaTest` and `dagoTest` (both from the `library(fBasics)`)
- the function `mshapiro.test` (from the `library(mvnormtest)`) to test for multivariate normality
- Crawley (2005: 100f.), Crawley (2007: 316f.)

1.1.2. One dep. variable (nominal/categorical)

In this section, we are going to return to an example from Section 1.3, the constructional alternation of particle placement in English, which is again represented in (24).

(24) a. He picked up the book. (verb - particle - direct object)
b. He picked the book up. (verb - direct object - particle)

As you already know, usually both constructions are perfectly acceptable and native speakers can often not explain their preference for one of the two constructions. One may therefore expect that both constructions are equally frequent, and this is what you are going to test. This scenario involves

- a dependent nominal/categorical variable CONSTRUCTION: *VERB-PARTICLE-OBJECT* vs. CONSTRUCTION: *VERB-OBJECT-PARTICLE*;
- no independent variable, because you do not investigate whether the distribution of CONSTRUCTION is dependent on anything else.

Such questions are generally investigated with tests from the family of chi-square tests, which is one of the most important and widespread tests. Since there is no independent variable, you test the degree of fit between your observed and an expected distribution, which should remind you of Section 3.1.5.2. This test is referred to as the chi-square goodness-of-fit test and involves the following steps:

Procedure
Formulating the hypotheses
Tabulating the observed frequencies; inspecting a graph
Computing the frequencies you would expect given H_0
Testing the assumption(s) of the test:
all observations are independent of each other
80% of the expected frequencies are larger than or equal to 5[19]
all expected frequencies are larger than 1
Computing the contributions to chi-square for all observed frequencies
Summing the contributions to chi-square to get the test statistic χ^2
Determining the degrees of freedom *df* and the probability of error *p*

The first step is very easy here. As you know, the null hypothesis typically postulates that the data are distributed randomly/evenly, and that means that both constructions occur equally often, i.e., 50% of the time (just as tossing a fair coin many times will result in an approximately equal distribution). Thus:

H_0: The frequencies of the two variable levels of CONSTRUCTION are identical – if you find a difference in your sample, this difference is

[19] This threshold value of 5 is the one most commonly mentioned. There are a few studies that show that the chi-square test is fairly robust even if this assumption is violated – especially when, as is here the case, the null hypothesis postulates that the expected frequencies are equally high (cf. Zar 1999: 470). However, to keep things simple, I stick to the most common conservative threshold value of 5 and refer you to the literature quoted in Zar. If your data violate this assumption, then you must compute a binomial test (if, as here, you have two groups) or a multinomial test (for three or more groups); cf. the recommendations for further study.

just random variation.

H$_1$: The frequencies of the two variable levels of CONSTRUCTION are not identical.

From this, the statistical forms are obvious.

H$_0$: $n_{\text{V Part DO}} = n_{\text{V DO Part}}$

H$_1$: $n_{\text{V Part DO}} \neq n_{\text{V DO Part}}$

Note that this is a two-tailed hypothesis; no direction of the difference is provided. Next, you would collect some data and count the occurrences of both constructions, but we will abbreviate this step and use frequencies reported in Peters (2001). She conducted an experiment in which subjects described pictures and obtained the construction frequencies represented in Table 19.

Table 19. Observed construction frequencies of Peters (2001)

Verb - Particle - Direct Object	Verb - Direct Object - Particle
247	150

Obviously, there is a strong preference for the construction in which the particle follows the verb directly. At first glance, it seems very unlikely that the null hypothesis could be correct, given these data.

First, you should have a look at a graphical representation of the data. A first but again not optimal possibility would be to generate, say, a pie chart. Thus, you first enter the data and then create a pie chart or a bar plot as follows:

```
> VPCs<-c(247,·150)·#·VPCs="verb-particle·constructions"¶
> pie(VPCs,·labels=c("Verb-Particle-Direct·Object",·"Verb-
      Direct·Object-Particle"))¶
> barplot(VPCs,·names.arg=c("Verb-Particle-
      Direct·Object",·"Verb-Direct·Object-Particle"))¶
```

The question now of course is whether this preference is statistically significant or whether it could just as well have arisen by chance. According to the above procedure, you must now compute the frequencies that follow from H$_0$. In this case, this is easy: since there are altogether 247+150 = 397 constructions, which should be made up of two equally large groups, you divide 397 by 2:

```
> VPCs.exp<-rep(sum(VPCs)/length(VPCs),·length(VPCs))
[1]·198.5·198.5
```

Table 20. Expected construction frequencies for the data of Peters (2001)

Verb - Particle - Direct Object	Verb - Direct Object - Particle
198.5	198.5

You must now check whether you can actually do a chi-square test here, but the observed frequencies are obviously larger than 5 and we assume that Peters's data points are in fact independent (because we assume for now that, for instance, each construction has been provided by a different speaker). We can therefore proceed with the chi-square test, the computation of which is fairly straightforward and summarized in (25).

(25) $$\text{Pearson chi-square} = \chi^2 = \sum_{i=1}^{n} \frac{(observed - expected)^2}{expected}$$

That is to say, for every value of your frequency distribution you compute a so-called contribution to chi-square by (i) computing the difference between the observed and the expected frequency, (ii) squaring this difference, and (iii) dividing that by the expected frequency again. The sum of these contributions to chi-square is the test statistic chi-square. Here, chi-square is approximately 23.7.

(26) $$\text{Pearson } \chi^2 = \frac{(247-198.5)^2}{198.5} + \frac{(150-198.5)^2}{198.5} \approx 23.7$$

```
> sum(((VPCs-VPCs.exp)^2)/VPCs.exp)¶
[1]·23.70025
```

Obviously, this value increases as the differences between observed and expected frequencies increase (because then the numerators become larger). In addition, you can see from (26) that chi-square becomes 0 when all observed frequencies correspond to all expected frequencies because then the numerators become 0. We can therefore simplify our statistical hypotheses to the following:

H_0: $\chi^2 = 0$.
H_1: $\chi^2 > 0$.

But the chi-square value alone does not show you whether the differences are large enough to be statistically significant. So, what do you do with this value? Before computers became more widespread, a chi-square value was used to look up in a chi-square table whether the result is significant or not. Such tables typically have the three standard significance levels in the columns and different numbers of degrees of freedom (*df*) in the rows. The number of degrees of freedom here is the number of categories minus 1, i.e., *df* = 2-1 = 1, because when we have two categories, then one category frequency can vary freely but the other is fixed (so that we can get the observed number of elements, here 397). Table 21 is one such chi-square table for the three significance levels and 1 to 3 degrees of freedom.

Table 21. Critical χ^2-values for $p_{\text{two-tailed}} = 0.05, 0.01$, and 0.001 for $1 \leq df \leq 3$

	$p = 0.05$	$p = 0.01$	$p = 0.001$
$df = 1$	3.841	6.635	10.828
$df = 2$	5.991	9.21	13.816
$df = 3$	7.815	11.345	16.266

You can actually generate those values yourself with the function `qchisq`. As you saw above in Section 1.3.4.2, the function requires three arguments:

- `p`: the *p*-value(s) for which you need the critical chi-square values (for some *df*);
- `df`: the *df*-value(s) for the *p*-value for which you need the critical chi-square value;
- `lower.tail=F` : the argument to instruct R to only use the area under the chi-square distribution curve that is to the right of / larger than the observed chi-square value.

That is to say:

```
> qchisq(c(0.05,·0.01,·0.001),·1,·lower.tail=F)¶
[1]··3.841459··6.634897·10.827566
```

Or, for more advanced users:

```
> p.values<-matrix(rep(c(0.05,·0.01,·0.001),·3),·byrow=T,·
      ncol=3)¶
> df.values<-matrix(rep(1:3,·3),·byrow=F,·ncol=3)¶
> qchisq(p.values,·df.values,·lower.tail=F)¶
```

```
·········[,1]······[,2]·····[,3]
[1,]·3.841459··6.634897·10.82757
[2,]·5.991465··9.210340·13.81551
[3,]·7.814728·11.344867·16.26624
```

Once you have such a table, you can test your observed chi-square value for significance by determining whether your chi-square value is larger than the chi-square value(s) tabulated at the observed number of degrees of freedom. You begin with the smallest tabulated chi-square value and compare your observed chi-square value with it and continue to do so as long as your observed value is larger than the tabulated ones. Here, you first check whether the observed chi-square is significant at the level of 5%, which is obviously the case: 23.7 > 3.841. Thus, you can check whether it is also significant at the level of 1%, which again is the case: 23.7 > 6.635. Thus, you can finallyeven check if the observed chi-square value is maybe even highly significant, and again this is so: 23.7 > 10.827. You can therefore reject the null hypothesis and the usual way this is reported in your results section is this: "According to a chi-square goodness-of-fit test, the distribution of the two verb-particle constructions deviates highly significantly from the expected distribution ($\chi^2 = 23.7$; $df = 1$; $p_{\text{two-tailed}} < 0.001$): the construction where the particle follows the verb directly was observed 247 times although it was only expected 199 times, and the construction where the particle follows the direct objet was observed only 150 times although it was expected 199 times."

With larger and more complex amounts of data, this semi-manual way of computation becomes more cumbersome (and error-prone), which is why we will simplify all this a bit. First, you can of course compute the *p*-value directly from the chi-square value using the mirror function of `qchisq`, viz. `pchisq`, which requires the above three arguments:

```
> pchisq(23.7,·1,·lower.tail=F)¶
[1]·1.125825e-06
```

As you can see, the level of significance we obtained from our stepwise comparison using Table 21 is confirmed: *p* is indeed much smaller than 0.001, namely 0.00000125825. However, there is another even easier way: why not just do the whole test with one function? The function is called `chisq.test`, and in the present case it requires maximally three arguments:

- `x`: a vector with the observed frequencies;
- `p`: a vector with the expected percentages (not the frequencies!);

- `correct=T` or `correct=F`: when the sample size *n* is small ($15 \leq n \leq 60$), it is sometimes recommended to apply a so-called continuity correction (after Yates); `correct=T` is the default setting.[20]

In this case, this is easy: you already have a vector with the observed frequencies, the sample size *n* is much larger than 60, and the expected probabilities result from H_0. Since H_0 says the constructions are equally frequent and since there are just two constructions, the vector of the expected probabilities contains two times $^{1}/_{2} = 0.5$. Thus:

```
> chisq.test(VPCs,·p=c(0.5,·0.5))¶
········Chi-squared·test·for·given·probabilities
data:··VPCs
X-squared·=·23.7003,·df·=·1,·p-value·=·1.126e-06
```

You get the same result as from the manual computation but this time you immediately also get the exact *p*-value. What you do not also get are the expected frequencies, but these can be obtained very easily, too. The function `chisq.test` does more than it returns. It returns a data structure (a so-called list) so you can assign this list to a named data structure and then inspect the list for its contents:

```
> test<-chisq.test(VPCs,·p=c(0.5,·0.5))¶
> str(test)¶
List·of·8
·$·statistic:·Named·num·23.7
··..-·attr(*,·"names")=·chr·"X-squared"
·$·parameter:·Named·num·1
··..-·attr(*,·"names")=·chr·"df"
·$·p.value··:·num·1.13e-06
·$·method···:·chr·"Chi-squared·test·for·given·probabilities"
·$·data.name:·chr·"VPCs"
·$·observed·:·num·[1:2]·247·150
·$·expected·:·num·[1:2]·199·199
·$·residuals:·num·[1:2]··3.44·-3.44
·-·attr(*,·"class")=·chr·"htest"
```

Thus, if you require the expected frequencies, you just ask for them as follows, and of course you get the result you already know.

```
> test$expected¶
[1]·198.5·198.5
```

[20] For further options, cf. `?chisq.test¶` or `formals(chisq.test)¶`.

Let me finally mention that the above method computes a *p*-value for a two-tailed test. There are many tests in R where you can define whether you want a one-tailed or a two-tailed test. However, this does not work with the chi-square test. If you require the critical chi-square test value for $p_{\text{one-tailed}}$ = 0.05 for *df* = 1, then you must compute the critical chi-square test value for $p_{\text{two-tailed}}$ = 0.1 for *df* = 1 (with `qchisq(0.1,·1,·lower.tail=F)¶`), since your prior knowledge is rewarded such that a less extreme result in the predicted direction will be sufficient (cf. Section 1.3.4). Also, this means that when you need the $p_{\text{one-tailed}}$-value for a chi-square value, just take half of the $p_{\text{two-tailed}}$-value of the same chi-square value (with, say, `pchisq(23.7,·1,·lower.tail=F)/2¶`). But again: only with *df* = 1.

Warning/advice

Above I warned you to never change your hypotheses *after* you have obtained your results and then sell your study as successful support of the 'new' alternative hypothesis. The same logic does not allow you to change your hypothesis from a two-tailed one to a one-tailed one because your $p_{\text{two-tailed}}$ = 0.08 (i.e., non-significant) so that the corresponding $p_{\text{one-tailed}}$ = 0.04 (i.e., significant). Your choice of a one-tailed hypothesis must be motivated *conceptually*.

Recommendation(s) for further study

- the functions `binom.test` or `dbinom` to compute binomial tests
- the function `prop.test` (cf. Section 3.1.5.2) to test relative frequencies / percentages for significant deviations from expected frequencies / percentages
- the function `dmultinom` to help compute multinomial tests
- Dalgaard (2002: Ch. 7), Baayen (2008: Section 4.1.1)

1.2. Tests for differences/independence

In Section 4.1.1, we looked at goodness-of-fit tests for distributions / frequencies, but we now turn to tests for differences/independence.

1.2.1. One dep. variable (ordinal/interval/ratio scaled) and one indep. variable (nominal) (indep. samples)

Let us look at an example in which two independent samples are compared with regard to their overall distributions. You will test whether men and women differ with regard to the frequencies of hedges they use in discourse (i.e., expressions such as *kind of* or *sort of*). Again, note that we are here only concerned with the overall distributions – not just means or just variances. We could of course do such an investigation, too, but it is of course theoretically possible that the means are very similar while the variances are not and a test for different means might not uncover the overall distributional difference.

Let us assume you have recorded 60 two-minute conversations between a confederate of an experimenter, each with one of 30 men and 30 women, and then counted the numbers of hedges that the male and female subjects produced. You now want to test whether the distributions of hedge frequencies differs between men and women. This question involves

- an independent nominal/categorical variable, SEX: *MALE* and SEX: *FEMALE*;
- a dependent interval/ratio-scaled: the number of hedges produced: HEDGES.

The question of whether the two sexes differ in terms of the distributions of hedge frequencies is investigated with the two-sample Kolmogorov-Smirnov test:

Procedure
Formulating the hypotheses
Tabulating the observed frequencies; inspecting a graph
Testing the assumption(s) of the test: the data are continuous
Computing the cumulative frequency distributions for both samples
Computing the maximal absolute difference *D* of both distributions
Determining the probability of error *p*

First the hypotheses: the text form is straightforward and the statistical version is based on a test statistic called *D*.

H_0: The distribution of the dependent variable HEDGES does not differ depending on the levels of the independent variable SEX; $D = 0$.

H_1: The distribution of the dependent variable HEDGES differs depending on the levels of the independent variable SEX; $D > 0$.

Before we do the actual test, let us again inspect the data graphically. You first load the data from <C:/_sflwr/_inputfiles/04-1-2-1_hedges.txt>, make the variable names available, and check the data structure.

```
> Hedges<-read.table(choose.files(),·header=T,·sep="\t",·
      comment.char="",·quote="")¶
> attach(Hedges);·str(Hedges)¶
'data.frame':···60·obs.·of··3·variables:
·$·CASE··:·int··1·2·3·4·5·6·7·8·9·10·...
·$·HEDGES:·int··17·17·17·17·16·13·14·16·12·11·...
·$·SEX···:·Factor·w/·2·levels·"F","M":·1·1·1·1·1·1·1·1·1·1·...
```

Since you are interested in the general distribution, you create a strip-chart. In this kind of plot, the frequencies of hedges are plotted separately for each sex, but to avoid that identical frequencies are plotted directly onto each other (and can therefore not be distinguished anymore), you also use the argument `method=jitter` to add a tiny value to each data point, which in turn minimizes the proportion of overplotted data points. Then, you do not let R decide about the range of the *x*-axis but include the meaningful point at $x = 0$ yourself. Finally, with the function `rug` you add little bars to the *x*-axis (`side=1`) which also get jittered. The result is shown in Figure 41.

```
> stripchart(HEDGES~SEX,·method="jitter",·xlim=c(0,·25),·
      xlab="Number·of·hedges",·ylab="Sex")¶
> rug(jitter(HEDGES),·side=1)¶
```

It is immediately obvious that the data are distributed quite differently: the values for women appear to be a little higher on average and more homogeneous than those of the men. The data for the men also appear to fall into two groups, a suspicion that also receives some prima facie support from the following two histograms in Figure 42. (Note that all axis limits are again defined identically to make the graphs easier to compare.)

```
> par(mfrow=c(1,·2))·#·plot·into·two·adjacent·graphs¶
> hist(HEDGES[SEX=="M"],·xlim=c(0,·25),·ylim=c(0,·10),·ylab=
      "Frequency",·main="")¶
> hist(HEDGES[SEX=="F"],·xlim=c(0,·25),·ylim=c(0,·10),·ylab=
      "Frequency",·main="")¶
> par(mfrow=c(1,·1))·#·restore·the·standard·plotting·setting¶
```

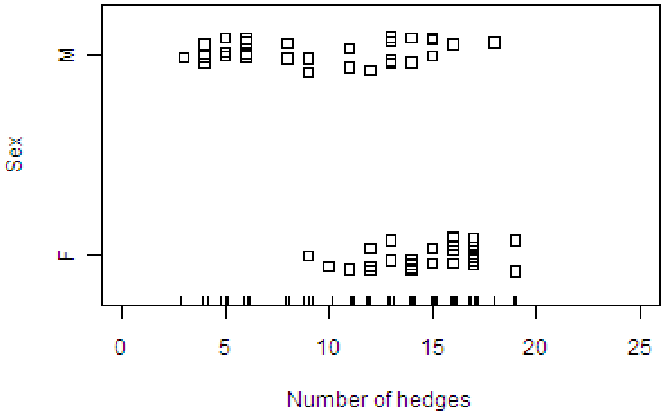

Figure 41. Stripchart for HEDGES~SEX

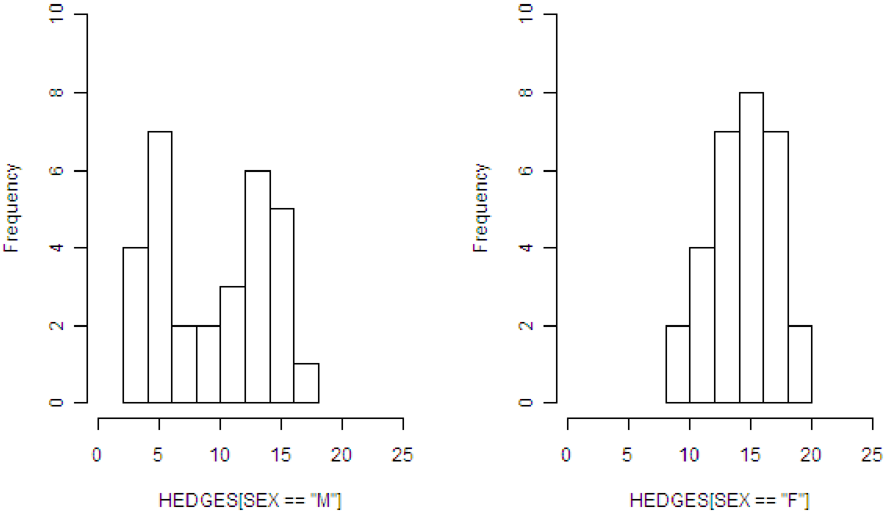

Figure 42. Histograms of the number of hedges by men and women

The assumption of continuous data points is not exactly met because frequencies are discrete – there are no frequencies 3.3, 3.4, etc. – but HEDGES is ratio-scaled and we are therefore rather safe (and could in fact jitter the values to avoid ties). To test these distributional differences with the Kolmogorov-Smirnov test, you first rank-order the data. To that end, you sort the values of SEX in the order in which you need to sort HEDGES, and then you do the same to HEDGES itself:

```
> SEX<-SEX[order(HEDGES)]¶
> HEDGES<-HEDGES[order(HEDGES)]¶
```

The next step is a little more complex. You must now compute the maximum of all differences of the two cumulative distributions of the hedges. You can do this in three steps: First, you generate a frequency table with the numbers of hedges in the rows and the sexes in the columns. This table in turn serves as input to `prop.table`, which generates a table of column percentages (hence `margin=2`; cf. Section 3.2.1):

```
> dists<-prop.table(table(HEDGES,·SEX),·margin=2);·dists¶
······SEX
HEDGES··········F··········M
····3··0.00000000·0.03333333
····4··0.00000000·0.10000000
····5··0.00000000·0.10000000
····6··0.00000000·0.13333333
····8··0.00000000·0.06666667
····9··0.03333333·0.06666667
····10·0.03333333·0.00000000
····11·0.03333333·0.06666667
····12·0.10000000·0.03333333
····13·0.06666667·0.13333333
····14·0.16666667·0.06666667
····15·0.06666667·0.13333333
····16·0.20000000·0.03333333
····17·0.23333333·0.00000000
····18·0.00000000·0.03333333
····19·0.06666667·0.00000000
```

That means that, say, 10% of all numbers of hedges of men are 4, but these are of course not cumulative percentages yet. The second step is therefore to convert these percentages into cumulative percentages. You can use `cumsum` to generate the cumulative percentages for both columns and can even compute the differences in the same line:

```
> differences<-cumsum(dists[,1])-cumsum(dists[,2])¶
```

That is, you subtract from every cumulative percentage of the first column (the values of the women) the corresponding value of the second column (the values of the men). The third and final step is then to determine the maximal absolute difference, which is the test statistic *D*:

```
> max(abs(differences))¶
[1]·0.4666667
```

You can then look up this value in a table for Kolmogorov-Smirnov tests; for a significant result, the computed value must be larger than the tabulated one. For cases in which both samples are equally large, Table 22 shows the critical *D*-values for two-tailed Kolmogorov-Smirnov tests.

Table 22. Critical *D*-values for two-sample Kolmogorov-Smirnov tests (for equal sample sizes)[21]

	$p = 0.05$	$p = 0.01$
$n_1 = n_2 = 29$	$^{10}/_{29}$	$^{12}/_{29}$
$n_1 = n_2 = 30$	$^{10}/_{30}$	$^{12}/_{30}$
$n_1 = n_2 = 31$	$^{10}/_{31}$	$^{12}/_{31}$

The observed value of $D = 0.4667$ is not only significant ($D > {}^{10}/_{30}$), but even very significant ($D > {}^{12}/_{30}$). You can therefore reject H_0 and summarize the results: “According to a two-sample Kolmogorov-Smirnov test, there is a significant difference between the distributions of hedge frequencies of men and women: on the whole, women seem to use more hedges and behave more homogeneously than the men, who use fewer hedges and whose data appear to fall into two groups ($D = 0.4667$, $p_{\text{two-tailed}} < 0.01$).”

The logic underlying this test is not always immediately clear. Since it is a very versatile test with hardly any assumptions, it is worth to briefly explore what this test is sensitive to. To that end, we again look at a graphical representation. The following lines plot the two cumulative distribution functions of men (in dark grey) and women (in black) as well as a vertical line at position $x = 8$, where the largest difference ($D = 0.4667$) was found. This graph in Figure 43 below shows what the Kolmogorov-Smirnov test reacts to: different cumulative distributions.

```
> plot(cumsum(dists[,1]),·type="b",·col="black",·
      xlab="Numbers·of·Hedges",·ylab="Cumulative·frequency·
      in·%",·xlim=c(0,·16));·grid()¶
> lines(cumsum(dists[,2]),·type="b",·col="darkgrey")¶
> text(14,·0.1,·labels="Women",·col="black")¶
> text(2.5,·0.9,·labels="Men",·col="darkgrey")¶
> abline(v=8,·lty=2)¶
```

[21] For sample sizes $n \geq 40$, the *D*-values for $p_{\text{two-tailed}} = 0.05$ are approximately $1.92/n0.5$.

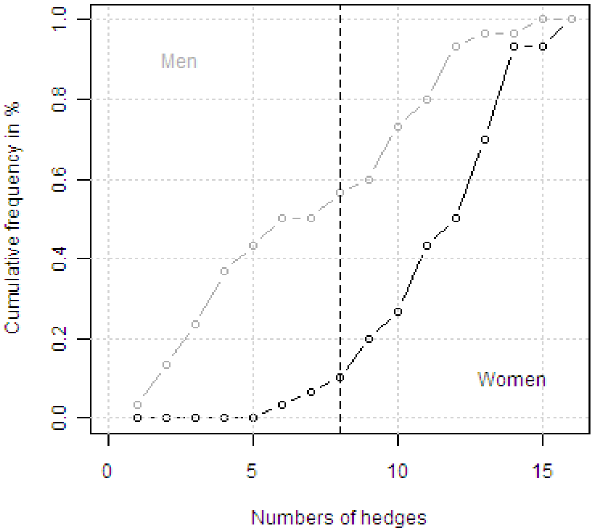

Figure 43. Cumulative distribution functions of the numbers of hedges of men and women

For example, the facts that the values of the women are higher and more homogeneous is indicated especially in the left part of the graph where the low hedge frequencies are located and where the values of the men already rise but those of the women do not. More than 40% of the values of the men are located in a range where no hedge frequencies for women were obtained at all. As a result, the largest different at position $x = 8$ is in the middle where the curve for the men has already risen considerably while the curve for the women has only just begun to rise. This graph also explains why H_0 postulates $D = 0$. If the curves are completely identical, there is no difference between them and D becomes 0.[22]

The above explanation simplified things a bit. First, you do not always have two-tailed tests and identical sample sizes. Second, identical values – so-called *ties* – can complicate the computation of the test. Fortunately, you do not really have to worry about that because the R function `ks.test` does

[22] An alternative way to produce a similar graph involves the function `ecdf` (for *empirical cumulative distribution function*):

```
> plot(ecdf(HEDGES[SEX=="M"]),·do.points=F,·verticals=T,·
         col.h="black",·col.v="black",·main="Hedges:·men·vs.·
         women")¶
> lines(ecdf(HEDGES[SEX=="F"]),·do.points=F,·verticals=T,·
         col.h="darkgrey",·col.v="darkgrey")¶
```

everything for you in just one line. You just need the following arguments:[23]

- x and y: the two vectors whose distributions you want to compare;
- alternative="two-sided" for two-tailed tests (the default) or alternative="greater" or alternative="less" for one-sided tests depending on which alternative hypothesis you want to test: the argument alternative="..." refers to the first-named vector so that alternative="greater" means that the cumulative distribution function of the first vector is above that of the second.

When you test a two-tailed H_1 as we do here, then the line to enter into R reduces to the following, and you get the same *D*-value and the *p*-value. (I omitted the warning about ties here but, again, you can jitter the vectors to get rid of it.)

```
> ks.test(HEDGES[SEX=="M"],·HEDGES[SEX=="F"])¶
········Two-sample·Kolmogorov-Smirnov·test
data:··HEDGES[SEX·==·"M"]·and·HEDGES[SEX·==·"F"]
D·=·0.4667,·p-value·=·0.002908
alternative·hypothesis:·two-sided
```

Recommendation(s) for further study

- apart from the function mentioned in note 22 (plot(ecdf(...))), you can create such graphs also with plot.stepfun or, even easier, with plot and the argument type="s"; cf. the file with the R code for this chapter
- Crawley (2005: 100f.), Crawley (2007: 316f.), Baayen (2008: Section 4.2.1)

1.2.2 One dep. variable (nominal/categorical) and one indep. variable (nominal/categorical) (indep. samples)

In Section 4.1.1.2 above, we discussed how you test whether the distribution of a dependent nominal/categorical variable is significantly different from another known distribution. A probably more frequent situation is that you test whether the distribution of one nominal/categorical variable is dependent on another nominal/categorical variable.

Above, we looked at the frequencies of the two verb-particle construc-

[23] Unfortunately, the function ks.test does not take a formula as input.

tions. We found that their distribution was not compatible with H_0. However, we also saw earlier that there are many variables that are correlated with the constructional choice. One of these is whether the referent of the direct object is given information, i.e., known from the previous discourse, or not. More specifically, previous studies found that objects referring to given referents prefer the position before the particle whereas objects referring to new referents prefer the position after the particle. In what follows, we will look at this hypothesis (as a two-tailed hypothesis, though). The question involves

- a dependent nominal/categorical variable, namely CONSTRUCTION: *VERB-PARTICLE-OBJECT* vs. CONSTRUCTION: *VERB-OBJECT-PARTICLE*;
- an independent variable nominal/categorical variable, namely the givenness of the referent of the direct object: GIVENNESS: *GIVEN* vs. GIVENNESS: *NEW*;
- independent samples because we will assume that, in the data below, the fact that a particular object is given is unrelated to whether another object is also given or not (this is often far from obvious, but I cannot discuss this issue here in more detail).

As before, such questions are investigated with chi-square tests: you test whether the levels of the independent variable result in different frequencies of the levels of the dependent variable. The overall procedure for a chi-square test for independence is very similar to that of a chi-square test for goodness of fit, but you will see below that the computation of the expected frequencies is (only superficially) a bit different from above.

Procedure

Formulating the hypotheses
Tabulating the observed frequencies; inspecting a graph
Computing the frequencies you would expect given H_0
Testing the assumption(s) of the test:
    all observations are independent of each other
    80% of the expected frequencies are larger than or equal to 5 (cf. n. 19)
    all expected frequencies are larger than 1
Computing the contributions to chi-square for all observed frequencies
Summing the contributions to chi-square to get the test statistic χ^2
Determining the degrees of freedom *df* and the probability of error *p*

The text forms of the hypotheses are simple:

H_0: The frequencies of the levels of the dependent variable CONSTRUCTION do not vary as a function of the levels of the independent variable GIVENNESS.

H_1: The frequencies of the levels of the dependent variable CONSTRUCTION vary as a function of the levels of the independent variable GIVENNESS.

Formulating the statistical hypothesis is a bit more complex here and can be seen as related to the tabulation of the observed frequencies and the computation of the expected frequencies, which is why I will discuss these three things together and only explain later how you can stick to the order of the procedure above.

In order to discuss this version of the chi-square test, we return to the data from Peters (2001). As a matter of fact, the above discussion did not utilize all of Peters's data because I omitted an independent variable, namely GIVENNESS. Peters (2001) did not just study the frequency of the two constructions – she studied what we are going to look at here, namely whether GIVENNESS is correlated with CONSTRUCTION. In the picture-description experiment described above, she manipulated the variable GIVENNESS and obtained the already familiar 397 verb-particle constructions, which patterned as represented in Table 23.

Before we discuss how to do the significance test, let us first explore the data graphically. You load the data from <C:/_sflwr/_inputfiles/04-1-2-2_vpcs.txt>, create a table of the two factors, and get a first visual impression of the distribution of the data:

Table 23. Observed construction frequencies of Peters (2001)

	GIVENNESS:	GIVENNESS:	Row
	GIVEN	*NEW*	totals
CONSTRUCTION: *V DO PART*	85	65	150
CONSTRUCTION: *V PART DO*	100	147	247
Column totals	185	212	397

```
> VPCs<-read.table(choose.files(),·header=T,·sep="\t",·
      comment.char="",·quote="")¶
> attach(VPCs);·str(VPCs)¶
'data.frame':···397·obs.·of··3·variables:
```

```
·$·CASE········:·int··1·2·3·4·5·6·7·8·9·10·...
·$·GIVENNESS···:·Factor·w/·2·levels·"given","new":·1·1·1·...
·$·CONSTRUCTION:·Factor·w/·2·levels·"V_DO_Part","V_Part_DO":·
      1·1·...
> Peters.2001<-table(CONSTRUCTION,·GIVENNESS)¶
> plot(CONSTRUCTION~GIVENNESS)¶
```

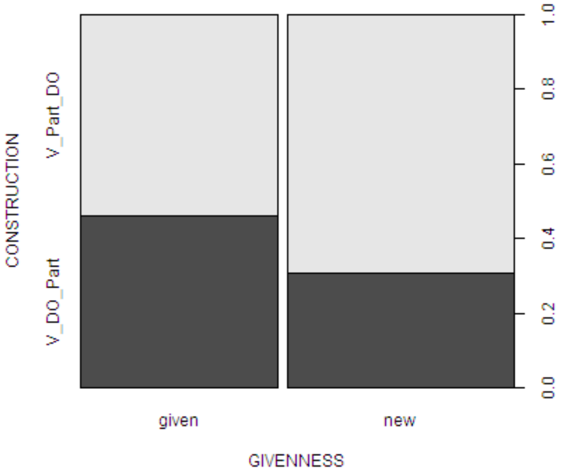

Figure 44. Mosaic plot for CONSTRUCTION~GIVENNESS

Obviously, the differently-colored areas are differently big between rows/columns. To test these differences for significance, we need the frequencies expected from H_0. But how do we formulate H_0 and compute these frequencies? Since this is a central question, we will discuss this in detail. Let us assume Peters had obtained the totals in Table 24.

Table 24. Fictitious observed construction frequencies of Peters (2001)

	GIVENNESS: *GIVEN*	GIVENNESS: *NEW*	Row totals
CONSTRUCTION: *V DO PART*			100
CONSTRUCTION: *V PART DO*			100
Column totals	100	100	200

What would the distribution following from H_0 look like? Above in Section 4.1.1.2, we said that H_0 typically postulates equal frequencies. Thus, you might assume – correctly – that the expected frequencies are those represented in Table 25. All marginal totals are 100 and every variable has two equally frequent levels so we have 50 in each cell.

Table 25. Fictitious expected construction frequencies of Peters (2001)

	GIVENNESS: *GIVEN*	GIVENNESS: *NEW*	Row totals
CONSTRUCTION: *V DO PART*	50	50	100
CONSTRUCTION: *V PART DO*	50	50	100
Column totals	100	100	200

The statistical hypotheses would then be:

H$_0$: $n_{\text{V DO Part \& Ref DO = given}} = n_{\text{V DO Part \& Ref DO} \neq \text{given}} = n_{\text{V Part DO \& Ref DO = given}} = n_{\text{V Part DO \& Ref DO} \neq \text{given}}$

H$_1$: as H$_0$, but there is at least one “≠” instead of an “=“.

However, life is usually not that simple, for example when (a) as in Peters (2001) not all subjects answer all questions or (b) naturally-observed data are counted that are not as nicely balanced. Thus, let us now return to Peters’s real data. In her case, it does not make sense to simply assume equal frequencies. Put differently, H$_1$ cannot be the above because we know from the row totals of Table 23 that the different levels of GIVENNESS are not equally frequent. If GIVENNESS had no influence on CONSTRUCTION, then you would expect that the frequencies of the two constructions for each level of GIVENNESS would exactly reflect the frequencies of the two constructions in the sample as whole. That means (i) all marginal totals must remain constant (since they reflect the numbers of the investigated elements), and (ii) the proportions of the marginal totals determine the cell frequencies in each row and column. From this, a rather complex set of hypotheses follows (which we will simplify presently):

H$_0$: $n_{\text{V DO Part \& Ref DO = given}} : n_{\text{V DO Part \& Ref DO} \neq \text{given}} \propto$
$n_{\text{V Part DO \& Ref DO = given}} : n_{\text{V Part DO \& Ref DO} \neq \text{given}} \propto$
$n_{\text{Ref DO = given}} : n_{\text{Ref DO} \neq \text{given}}$ and
$n_{\text{V DO Part \& Ref DO = given}} : n_{\text{V Part DO \& Ref DO = given}} \propto$
$n_{\text{V DO Part \& Ref DO} \neq \text{given}} : n_{\text{V Part DO \& Ref DO} \neq \text{given}} \propto$
$n_{\text{V DO Part}} : n_{\text{V Part DO}}$

H_1: as H_0, but there is at least one "≠" instead of an "=".

In other words, you cannot simply say, "there are 2·2 = 4 cells and I assume each expected frequency is 397 divided by 4, i.e., approximately 100." If you did that, the upper row total would amount to nearly 200 – but that can't be correct since there are only 150 cases of CONSTRUCTION: *VERB-OBJECT-PARTICLE* and not ca. 200. Thus, you must include this information, that there are only 150 cases of CONSTRUCTION: *VERB-OBJECT-PARTICLE* into the computation of the expected frequencies. The easiest way to do this is using percentages: there are $^{150}/_{397}$ cases of CONSTRUCTION: *VERB-OBJECT-PARTICLE* (i.e. 0.3778 = 37.78%). Then, there are $^{185}/_{397}$ cases of GIVENNESS: *GIVEN* (i.e., 0.466 = 46.6%). If the two variables are independent of each other, then the probability of their joint occurrence is 0.3778·0.466 = 0.1761. Since there are altogether 397 cases to which this probability applies, the expected frequency for this combination of variable levels is 397·0.1761 = 69.91. This logic can be reduced to the formula in (27).

(27) $$n_{\text{expected cell frequency}} = \frac{row\ sum \cdot column\ sum}{n}$$

If you apply this logic to every cell, you get Table 26.

Table 26. Expected construction frequencies of Peters (2001)

	GIVENNESS: *GIVEN*	GIVENNESS: *NEW*	Row totals
CONSTRUCTION: *V DO PART*	69.9	80.1	150
CONSTRUCTION: *V PART DO*	115.1	131.9	247
Column totals	185	212	397

You can immediately see that this table corresponds to the above null hypothesis: the ratios of the values in each row and column are exactly those of the row totals and column totals respectively. For example, the ratio of 69.9 to 80.1 to 150 is the same as that of 115.1 to 131.9 to 247 and as that of 185 to 212 to 397, and the same is true in the other dimension. Thus, the null hypothesis does not mean 'all cell frequencies are identical' – it means 'the ratios of the cell frequencies are equal (to each other and the respective marginal totals.'

This method to compute expected frequencies can be extended to arbitrarily complex frequency tables (as you will see again in Chapter 5). But how do we test whether these deviate strongly enough from the observed frequencies? And do we really need such complicated hypotheses? Fortunately, there are simple and interrelated answers to these questions. As was mentioned above, the chi-square test for independence is very similar to the chi-square goodness-of-fit test: for each cell, you compute a contribution to chi-square, you sum those up to get the chi-square test statistic, and since we have discussed above that chi-square becomes zero when the observed frequencies are the same as the expected, we can abbreviate our hypothesis considerably:

H_0: $\chi^2 = 0$.
H_1: $\chi^2 > 0$.

And since this kind of null hypothesis does not require any specific observed or expected frequencies, it allows you to stick to the order of steps in the above procedure and formulate hypotheses *before* having data.

As before, the chi-square test can only be used when its assumptions are met. The expected frequencies are large enough and for simplicity's sake we assume here that every subject only gave just one sentence so that the observations are independent of each other: for example, the fact that some subject produced a particular sentence on one occasion does then not affect any other subject's formulation. We can therefore proceed as above and compute (the sum of) the contributions to chi-square on the basis of the same formula, here repeated as (28):

(28) $$\text{Pearson } \chi^2 = \sum_{i=1}^{n} \frac{\left(observed - expected\right)^2}{expected}$$

The results are shown in Table 27 and the sum of all contributions to chi-square, chi-square itself, is 9.82. However, we again need the number of degrees of freedom. For two-dimensional tables and when the expected frequencies are computed on the basis of the observed frequencies as you did above, the number of degrees of freedom is computed as shown in (29).[24]

[24] In our example, the expected frequencies were computed from the observed frequencies

Table 27. Contributions to chi-square for the data of Peters (2001)

	GIVENNESS: *GIVEN*	GIVENNESS: *NEW*	Row totals
CONSTRUCTION: *V DO PART*	3.26	2.85	
CONSTRUCTION: *V PART DO*	1.98	1.73	
Column totals			9.82

(29) df = (no. of rows-1) · (no. of columns-1) = (2-1)·(2-1) = 1

With both the chi-square and the *df-val*ue, you can look up the result in a chi-square table. As above, if the observed chi-square value is larger than the one tabulated for $p = 0.05$ at the required *df-val*ue, then you can reject H_0. Thus, Table 28 is the same as Table 21 and can be generated with `qchisq` as explained above.

Table 28. Critical χ^2-values for $p_{\text{two-tailed}} = 0.05$, 0.01, and 0.001 for $1 \leq df \leq 3$

	$p = 0.05$	$p = 0.01$	$p = 0.001$
$df = 1$	3.841	6.635	10.828
$df = 2$	5.991	9.21	13.816
$df = 3$	7.815	11.345	16.266

Here, chi-square is not only larger than the critical value for $p = 0.05$ and $df = 1$, but also larger than the critical value for $p = 0.01$ and $df = 1$. But, since the chi-square value is not also larger than 10.827, the actual *p*-value is somewhere between 0.01 and 0.001: the result is very significant, but not highly significant.

Fortunately, all this is much easier when you use R's built-in function. Either you compute just the *p*-value as before,

```
> pchisq(9.82,·1,·lower.tail=F)¶
[1]·0.001726243
```

or you use the function `chisq.test` and do everything in a single step. The most important arguments for our purposes are

in the marginal totals. If you compute the expected frequencies not from your observed data but from some other distribution, the computation of df changes to: df = (number of rows · number of columns)-1.

- `x`: the two-dimensional table for which you want to do a chi-square test;
- `correct=T` or `correct=F`; cf. above for the continuity correction.[25]

```
> eval.Pet<-chisq.test(Peters.2001,·correct=F);·eval.Pet¶
········Pearson's·Chi-squared·test
data:··Peters.2001
X-squared·=·9.8191,·df·=·1,·p-value·=·0.001727
```

As before, you can also obtain the expected frequencies or just the chi-square value itself:

```
> eval.Pet$expected¶
············GIVENNESS
CONSTRUCTION·····given·······new
···V_DO_Part··69.89924··80.10076
···V_Part_DO·115.10076·131.89924
> eval.Pet$statistic¶
X-squared
·9.819132
```

You now know that GIVENNESS is correlated with CONSTRUCTION, but you neither know yet how strong that effect is nor which variable level combinations are responsible for this result. As for the effect size, even though you might be tempted to use the size of the chi-square value to quantify the effect, you must not do that. This is because the chi-square value is dependent on the sample size, as we can easily see:

```
> chisq.test(Peters.2001*10,·correct=F)¶
········Pearson's·Chi-squared·test
data:··Peters.2001·*·10
X-squared·=·98.1913,·df·=·1,·p-value·<·2.2e-16
```

For effect sizes, this is of course a disadvantage since just because the sample size is larger, this does not mean that the relation of the values to each other has changed, too. You can easily verify this by noticing that the ratios of percentages, for example, have stayed the same. For that reason, the effect size is often quantified with a coefficient of correlation (called ϕ in the case of $k\times2$/$m\times2$ tables or Cramer's V for $k\times m$ tables with k or $m > 2$), which falls into the range between 0 and 1 (0 = no correlation; 1 = perfect correlation) and is unaffected by the sample size. This correlation coefficient is computed according to the formula in (30):

[25] For further options, cf. again `?chisq.test¶`. Note also what happens when you enter `summary(Peters.2001)¶`.

(30) ϕ / Cramer's *V* / Cramer's index *I* =

$$\sqrt{\frac{\chi^2}{n \cdot (\min[n_{rows}, n_{columns}] - 1)}}$$

In R, you can of course do this in one line of code:

```
> sqrt(eval.Pet$statistic/
      sum(Peters.2001)*(min(dim(Peters.2001))-1))¶
X-squared
0.1572683
```

Given the theoretical range of values, this is a rather small effect size.[26] Thus, the correlation is probably not random, but practically not extremely relevant.

Another measure of effect size, which can however only be applied to 2×2 tables, is the so-called odds ratio. An odds ratio tells you how the likelihood of one variable level changes in response to a change of the other variable's level. The odds of an event *E* correspond to the fraction in (31).

(31) $$odds = \frac{p_E}{1 - p_E} \quad \text{(you get probabilities from odds with } \frac{odds}{1 + odds}\text{)}$$

The odds ratio for a 2×2 table such as Table 23 is the ratio of the two odds (or 1 divided by that ratio, depending on whether you look at the event *E* or the event ¬*E* (not *E*)):

(32) $$odds\ ratio \text{ for Table 23} = \frac{85 \div 65}{100 \div 147} = 1.9223$$

In words, the odds of CONSTRUCTION: *V DO PART* are $(^{85}/_{185})$ / $(1-^{85}/_{185})$ = $^{85}/_{100}$ = 0.85 when the referent of the direct object is given and $^{65}/_{147}$ = 0.4422 when the referent of the direct object is new. This in turn means that CONSTRUCTION: *V DO PART* is $^{0.85}/_{0.4422} \approx 1.9223$ times more likely when the referent of the direct object is given than when it is not. From this, it also

[26] The theoretical range from 0 to 1 is really only possible in particular situations, but still a good heuristic to interpret this value.

follows that the odds ratio in the absence of an interaction is ≈ 1.[27]

This is how you would summarize the above results: “New objects are strongly preferred in the construction Verb-Particle-Direct Object and are dispreferred in Verb-Direct Object-Particle. The opposite kind of constructional preference is found for given objects. According to a chi-square test for independence, this correlation is very significant ($\chi^2 = 9.82$; $df = 1$; $p_{\text{two-tailed}} < 0.002$), but the effect is not particularly strong ($\phi = 0.157$, odds ratio = 1.9223).

Table 27 also shows which variable level combinations contribute most to the significant correlation: the larger the contribution to chi-square of a cell, the more that cell contributes to the overall chi-square value; in our example, these values are all rather small – none exceeds the chi-square value for $p = 0.05$ and $df = 1$, i.e., 3.841. In R, you can get the contributions to chi-square as follows:

```
> eval.Pet$residuals^2¶
············GIVENNESS
CONSTRUCTION····given······new
···V_DO_Part·3.262307·2.846825
···V_Part_DO·1.981158·1.728841
```

That is, you compute the Pearson residuals and square them. The Pearson residuals in turn can be computed as follows; negative and positive values mean that observed values are smaller and larger than the expected values respectively.

```
> eval.Pet$residuals¶
············GIVENNESS
CONSTRUCTION·····given·······new
···V_DO_Part··1.806186·-1.687254
···V_Part_DO·-1.407536··1.314854
```

Thus, if, given the small contributions to chi-square, one wanted to draw any further conclusions, then one could only say that the variable level combination contributing most to the significant result is the combination of CONSTRUCTION: *V DO PART* and GIVENNESS: *GIVEN*, but the individual

[27] Often, you may find the logarithm of the odds ratio. When the two variables are not correlated, this log *odds ratio* is log 1 = 0, and positive/negative correlations result in positive/negative log odds ratios, which is often a little easier to interpret. For example, if you have two odds ratios such as $odds\ ratio_1 = 0.5$ and $odds\ ratio_2 = 1.5$, then you cannot immediately see, which effect is larger. The logs of the odds ratios – $\log_{10} odds\ ratio_1 = -0.301$ and $\log_{10} odds\ ratio_2 = 0.176$ – tell you immediately the former is larger because its absolute value is larger.

cells' effects here are really rather small. An interesting and revealing graphical representation is available with the function `assocplot`, whose most relevant argument is the two-dimensional table under investigation: In this plot, "the area of the box is proportional to the difference in observed and expected frequencies" (cf. R Documentation, s.v. `assocplot` for more details). The black rectangles above the dashed lines indicate observed frequencies exceeding expected frequencies; grey rectangles below the dashed lines indicate observed frequencies smaller than expected frequencies; the heights of the boxes are proportional to the above Pearson residuals and the widths are proportional to the square roots of the expected frequencies.

```
> assocplot(Peters.2001)¶
```

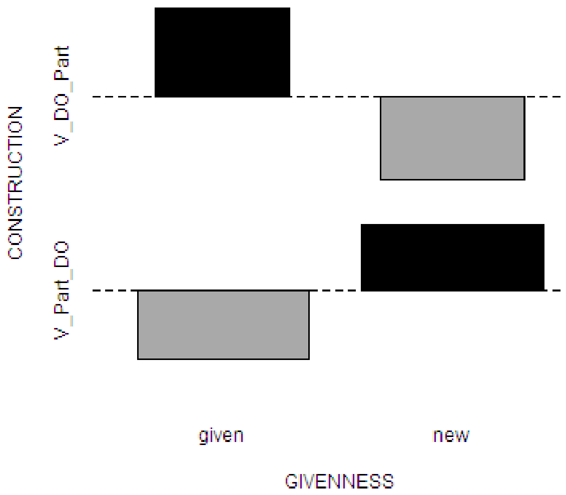

Figure 45. Association plot for CONSTRUCTION~GIVENNESS

(As a matter of fact, I usually prefer to transpose the table before I plot an association plot because then the row/column organization of the plot corresponds to that of the original table: `assocplot(t(Peters.2001))¶`) Another interesting way to look at the data is a mixture between a plot and a table. The table/graph in Figure 46 has the same structure as Table 23, but (i) the sizes in which the numbers are plotted directly reflects the size of the residuals (i.e., bigger numbers deviate more from the expected frequencies than smaller numbers, where *bigger* and *smaller* are to be understood in terms of plotting size), and (ii) the coloring indicates how the observed frequencies deviate from the expected ones: dark grey indicates positive residuals and light grey indicates negative residuals. (The function to do this is available from me upon request; for lack of a better terms, for now I refer to this as a cross-tabulation plot.)

Let me finally emphasize that the above procedure is again the one providing you with a *p*-value for a two-tailed test. In the case of 2×2 tables, you can perform a one-tailed test as discussed in Section 4.1.1.2 above, but you cannot do one-tailed tests for tables with *df* > 1. In Section 5.1, we will discuss an extension of chi-square tests to tables with more than two variables.

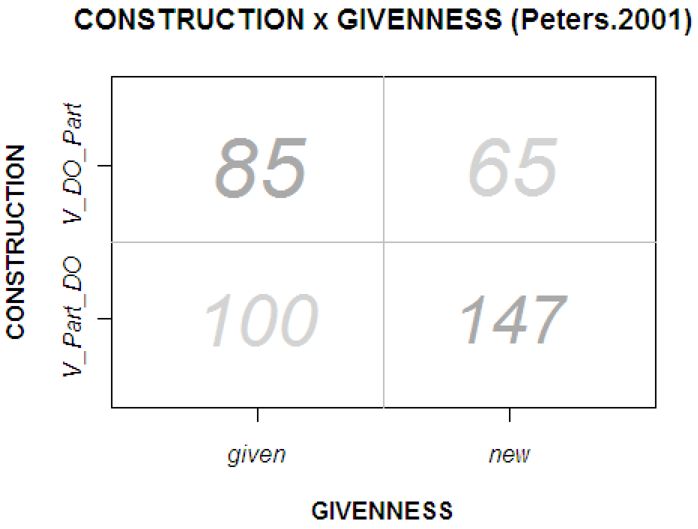

Figure 46. Cross-tabulation plot for CONSTRUCTION~GIVENNESS

Recommendation(s) for further study

- the functions `dotchart` and `sieve` (from the `library(vcd)`) as well as `table.cont` (from the `library(ade4)`) for other kinds of plots
- the function `assocstats` (from the `library(vcd)`) for a different way to compute chi-square tests and effect sizes at the same time
- the function `CrossTable` (from the `library(gmodels)`) for more comprehensive tables
- the argument `simulate.p.value=T` of the function `chisq.test` and the function `fisher.test`, which you can use when the expected frequencies are too small for a regular chi-square test

- the function `mantelhaen.test` to test for three-way interactions
- the Marascuilo procedure to test which observed row or column frequencies are different from each other in pairwise tests (cf. Gries, forthc., who also discusses how to test a subtable out of a larger table)
- Dalgaard (2002: Ch. 7), Crawley (2005: 85ff.), Crawley (2007: 301ff.)
- Good and Hardin (2006: 79f.) on problems of chi-square tests and some alternative suggestions

Let me mention one additional useful application of the chi-square test that is similar to the Mantel-Haenszel test (from Zar 1999: Section 23.4). Sometimes, you may have several isomorphic 2×2 tables on the same phenomenon, maybe because you found another source that discusses the same kind of data. You may then want to know whether or not the data are so similar that you can actually merge or amalgamate the data into one single data set. Here are the text hypotheses for that kind of question:

H_0: The trends in the data are identical: heterogeneity $\chi^2 = 0$.
H_1: The trends in the data are not identical: heterogeneity $\chi^2 > 0$.

To explore this approach, let us compare Peters's data to those of Gries (2003a). You can enter the latter into R directly:

```
> Gries.2003<-matrix(c(143,·66,·53,·141),·ncol=2,·byrow=T)¶
> rownames(Gries.2003)<-rownames(Peters.2001)¶
> colnames(Gries.2003)<-colnames(Peters.2001)¶
> Gries.2003¶
··········given·new
V_DO_Part···143··66
V_Part_DO····53·141
```

On the one hand, these data look very different from those of Peters (2001) because, here, when GIVENNESS is *GIVEN*, then CONSTRUCTION: *V_DO_PART* is nearly three times as frequent as CONSTRUCTION: *V_PART_DO* (and not in fact less frequent, as in Peters's data). On the other hand, the data are also similar because in both cases given direct objects increase the likelihood of CONSTRUCTION:*V_DO_PART*. A direct comparison of the association plots (not shown here, but you can use the following code to generate them) makes the data seem very much alike – how much more similar could two association plots be?

```
> par(mfrow=c(1,2))¶
> assocplot(Peters.2001)¶
> assocplot(Gries.2003,·xlab="CONSTRUCTION",·
      ylab="GIVENNESS")¶
> par(mfrow=c(1,1))·#·restore·the·standard·plotting·setting¶
```

However, you should not really compare the sizes of the boxes in association plots – only the overall tendencies – so we now turn to the heterogeneity chi-square test. The heterogeneity chi-square value is computed as the difference between the sum of chi-square values of the original tables and the chi-square value you get from the merged tables (that's why they have to be isomorphic), and it is evaluated with a number of degrees of freedom that is the difference between the sum of the degrees of freedom of all merged tables and the degrees of freedom of the merged table. Sounds pretty complex, but in fact it is not. The following code should make everything clear.

First, you compute the chi-square test for the data from Gries (2003):

```
> eval.Gr<-chisq.test(Gries.2003,·correct=F);·eval.Gr¶
········Pearson's·Chi-squared·test
data:··Gries.2003
X-squared·=·68.0364,·df·=·1,·p-value·<·2.2e-16
```

Then you compute the sum of chi-square values of the original tables:

```
> sum.chisq.indiv.tables<-eval.Pet$stat+eval.Gr$stat¶
```

After that, you compute the chi-square value of the combined table:

```
> eval.total<-chisq.test(Peters.2001+Gries.2003,·correct=F)¶
> sum.chisq.merged.table<-eval.total$stat¶
```

And then the heterogeneity chi-square and its degrees of freedom (you get the *df*-values with `$parameter`):

```
> het.chisq<-sum.chisq.indiv.tables-sum.chisq.merged.table¶
> het.df<-sum(eval.Pet$parameter,·eval.Gr$parameter)-
      eval.tot$parameter¶
```

How do you now get the *p*-value for these results?

THINK BREAK

```
> pchisq(het.chisq,·het.df,·lower.tail=F)¶
[1]·0.0005387754
```

As you can see, the data from the two studies are actually rather different: yes, they exhibit the same overall trends (given objects increase the likelihood of CONSTRUCTION:*V_DO_PART*, but they still differ highly significantly from each other ($\chi^2_{\text{heterogeneity}} = 11.98$; $df = 1$; $p_{\text{two-tailed}} < 0.001$). What is responsible for this difference? The different effect sizes: the odds ratio for Peters's data was 1.92, but in Gries's data it is nearly exactly three times as large:

```
> (143/66)/(53/141)¶
[1]·5.764151
```

And that is also what you would write in your results section.

1.2.3. One dep. variable (nominal/categorical) (dep. samples)

One central requirement of the chi-square test for independence is that the tabulated data points are independent of each other. There are situations, however, where this is not the case, and in this section I discuss one method you can use on one such occasion.

Let us assume you want to test whether metalinguistic knowledge influences acceptability judgments. This is relevant because many acceptability judgments used in linguistic research were produced by the investigating linguists themselves, and one may well ask oneself whether it is really sensible to rely on judgments by linguists with all their metalinguistic knowledge instead of on judgments by linguistically naïve subjects. This is especially relevant since studies have shown that judgments by linguists, who after all think a lot about sentences constructions, and other expressions, can deviate a lot from judgments by laymen, who usually don't (cf. Spencer 1973, Labov 1975, or Greenbaum 1976). In an admittedly oversimplistic case, you could ask 100 linguistically naïve native speakers to rate a sentence as 'acceptable' or 'unacceptable'. After the ratings have been made, you could tell the subjects which phenomenon the study investigated and which variable you thought influenced the sentences' acceptability. Then, you would give the sentences back to the subjects to have them rate them once more. The question would be whether the subjects' newly acquired metalinguistic knowledge would make them change their ratings and, if so, how. This question involves

- a dependent nominal/categorical variable, namely BEFORE: *ACCEPTABLE* vs. BEFORE: *UNACCEPTABLE*;
- a dependent nominal/categorical variable, namely AFTER: *ACCEPTABLE* vs. AFTER: *UNACCEPTABLE*;
- dependent samples since every subject produced two judgments.

For such scenarios, you use the McNemar test (or Bowker test, cf. below). This test is related to the chi-square tests discussed above in Sections 4.1.1.2 and 4.1.2.2 and involves the following procedure:

Procedure

Formulating the hypotheses
Testing the assumption(s) of the test:
    the observed variable levels are related in a pairwise manner
    the expected frequencies are larger than 5
Computing the frequencies you would expect given H_0
Computing the contributions to chi-square for all observed frequencies
Summing the contributions to chi-square to get the test statistic χ^2
Determining the degrees of freedom *df* and the probability of error *p*

First, the hypotheses:

H_0: The frequencies of the two possible ways in which subjects produce a judgment in the second rating task that differs from that in the first rating task are equal (or shorter $\chi^2 = 0$).

H_1: The frequencies of the two possible ways in which subjects produce a judgment in the second rating task that differs from that in the first rating task are not equal (or shorter $\chi^2 > 0$).

To get to know this test, we use the fictitious data summarized in Table 29, which you first read in from the file <C:/_sflwr/_inputfiles/04-1-2-3_accjudg.txt>.

```
> AccBeforeAfter<-read.table(choose.files(),·header=T,·
       sep="\t",·comment.char="",·quote="")¶
> attach(AccBeforeAfter);·str(AccBeforeAfter)¶
`data.frame':···100·obs.·of··3·variables:
·$·SENTENCE:·int··1·2·3·4·5·6·7·8·9·10·...
·$·BEFORE··:·Factor·w/·2·levels·"acceptable","inacceptable":·
       1·...
·$·AFTER···:·Factor·w/·2·levels·"acceptable","inacceptable":·
       1·...
```

Table 29. Observed frequencies in a fictitious study on acceptability judgments

		AFTER		
		ACCEPTABLE	*INACCEPTABLE*	Row totals
BEFORE	*ACCEPTABLE*	31	39	70
	INACCEPTABLE	13	17	30
	Column totals	44	56	100

Table 29 already suggests that there has been a major change of judgments: Of the 100 rated sentences, only 31+17 = 48 sentences – not even half! – were judged identically in both ratings. But now you want to know whether the way in which the 52 judgments changed is significantly different from chance. But what does the chance expectation look like?

The McNemar test only involves those cases where the subjects changed their opinion. If these are distributed equally, then the expected distribution of the 52 cases in which subjects change their opinion is that in Table 30.

Table 30. Expected frequencies in a fictitious study on acceptability judgments

		AFTER		
		ACCEPTABLE	*INACCEPTABLE*	Row totals
BEFORE	*ACCEPTABLE*		26	
	INACCEPTABLE	26		
	Column totals			

From this, you can see that both expected frequencies are larger than 5 so you can indeed do the McNemar test. As before, you compute a chi-square value (using the by now familiar formula in (33)) and a *df-val*ue according to the formula in (34) (where *k* is the number of rows/columns):

(33) $$\chi^2 = \sum_{i=1}^{n} \frac{(observed - expected)^2}{expected} = 13$$

(34) $$df = \frac{k \cdot (k-1)}{2} = 1$$

As before, you can look up this chi-square value in the kind of chi-square table and, again as before, if the computed chi-square value is larger than the tabulated one for the relevant *df-val*ue for $p = 0.05$, you may reject H_0. As you can see, the number of changes is too large to be compatible with H_0 and we accept H_1. As usual, you can of course compute the exact *p*-value with `pchisq(13,·1,·lower.tail=F)¶`.

This is how you summarize this finding in the results section: "According to a McNemar test, the way 52 out of 100 subjects changed their judgments after they were informed of the purpose of the experiment is significantly different from chance: in the second rating task, the number of 'acceptable' judgments is much smaller ($\chi^2 = 13$; $df = 1$; $p_{\text{two-tailed}} < 0.001$)."

Table 31. Critical chi-square values for $p_{\text{two-tailed}} = 0.05$, 0.01, and 0.001 for $1 \leq df \leq 3$

	$p = 0.05$	$p = 0.01$	$p = 0.001$
$df = 1$	3.841	6.635	10.828
$df = 2$	5.991	9.21	13.816
$df = 3$	7.815	11.345	16.266

In R, this is again much easier. You need the function `mcnemar.test` and it typically requires two arguments:

- `x`: a two-dimensional table for which you want to compute the McNemar test;
- `correct=F` or `correct=T` (the default): when the number of changes is smaller than 30, then sometimes the continuity correction is recommended.

```
> mcnemar.test(table(BEFORE,·AFTER),·correct=F)¶
········McNemar's·Chi-squared·test
data:··table(BEFORE,·AFTER)
McNemar's·chi-squared·=·13,·df·=·1,·p-value·=·0.0003115
```

The summary and conclusions are of course the same. When you do this test for *k*×*k* tables (with $k > 2$), this test is sometimes called Bowker test.

Recommendation(s) for further study

- for Cochran's extension of the McNemar test to test three or more measurements of a dichotomous variable, cf. Bortz (2005: 161f.)
- when a McNemar test yields a significant result, this result may theoretically be (in part) attributable to the order of the tests, which you can check with the Gart test
- the function `runs.test` (from the `library(tseries)`) to test the randomness of a binary sequence

2. Dispersions

Sometimes, it is necessary and/or interesting to not just look at the general characteristics of a distribution but with more narrowly defined distributional characteristics. The two most obvious characteristics are the dispersion and the central tendency of a distribution. This section is concerned with the dispersion – more specifically, the variance or standard deviation – of a variable; Section 4.3 discusses measures of central tendency.

For some research questions, it is useful to know, for example, whether two distributions have the same or a similar dispersion. Put differently, do two distributions spread around their means in a similar or in a different way. We touched upon this topic a little earlier in Section 3.1.3.6, but to illustrate the point once more, consider Figure 47.

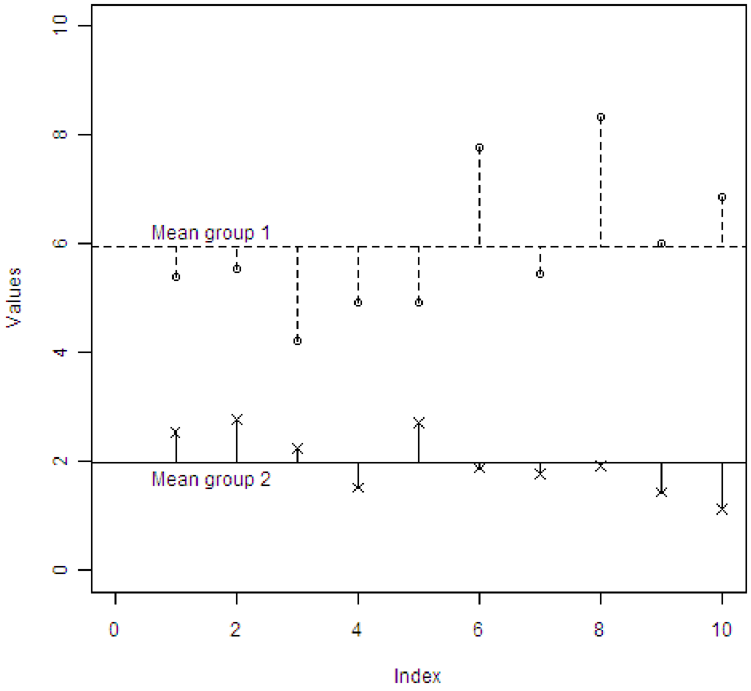

Figure 47. Two fictitious distributions

Figure 47 shows two distributions, one group of 10 values (represented by round points) and another group of 10 values (represented by crosses). The means of these groups are shown with the two horizontal lines (dashed for the first group), and the deviations of each point from its group mean are shown with the vertical lines. As you can easily see, the groups do not just differ in terms of their means ($mean_{\text{group 2}}$ = 1.99; $mean_{\text{group 1}}$ = 5.94), but also in terms of their dispersion: the deviations of the points of group 1 from their mean are much larger than their counterparts in group 2. While this difference is obvious in Figure 47, it can be much harder to discern in other cases, which is why we need a statistical test. In Section 4.2.1, we discuss how you test whether the dispersion of one dependent interval/ratio-scaled variable is significantly difference from a known dispersion value. In Section 4.2.2, we discuss how you test whether the dispersion of one dependent ratio-scaled variable differs significantly in two groups.

2.1. Goodness-of-fit test for one dep. variable (ratio-scaled)

As an example for this test, we return to the above example on first language acquisition of Russian tense-aspect patterning. In Section 4.1.1.1 above, we looked at how the correlation between the use of tense and aspect of one child developed over time. Let us assume, you now want to test whether the overall variability of the values for this child is significantly different from that of other children for which you already have data. Let as further assume that for these other children you found a variance of 0.025.

This question involves the following variables and is investigated with a chi-square test as described below:

- a dependent ratio-scaled variable, namely the variable TENSEASPECT, consisting of the Cramer's *V* values;
- no independent variable because you are not testing whether the distribution of the variable TENSEASPECT is influenced by, or correlated with, something else.

Procedure
Formulating the hypotheses
Computing the observed sample variance
Testing the assumption(s) of the test: the population from which the sample has been drawn or at least the sample from which the sample variance is computed is normally distributed
Computing the test statistic χ^2, the degrees of freedom *df*, and the probability of error *p*

As usual, you begin with the hypotheses:

H_0: The variance of the data for the newly investigated child does not differ from the variance of the children investigated earlier; *sd*^{2} TENSEASPECT of the new child = *sd*^{2} TENSEASPECT of the already investigated children, or *sd*^{2} of the new child = 0.025, or the quotient of the two variances is 1.

H_1: The variance of the data for the newly investigated child differs from the variance of the children investigated earlier; *sd*^{2} TENSEASPECT of the new child ≠ *sd*^{2} TENSEASPECT of the already investigated children, or *sd*^{2} of the new child ≠ 0.025, or the quotient of the two variances is not 1.

You load the data from <C:/_sflwr/_inputfiles/04-2-1_tense-aspect.txt>.

```
> RussTensAsp<-read.table(choose.files(),·header=T,·
      sep="\t",·comment.char="",·quote="")¶
> attach(RussTensAsp)¶
```

As a next step, you must test whether the assumption of this chi-square test is met and whether the data are in fact normally distributed. We have discussed this in detail above so we run the test here without further ado.

```
> shapiro.test(TENSE_ASPECT)¶
········Shapiro-Wilk·normality·test
data:···TENSE_ASPECT
W·=·0.9942,·p-value·=·0.9132
```

Just like in Section 4.1.1.1 above, you get a *p*-value of 0.9132, which means you must not reject H_0, you can consider the data to be normally distributed, and you can compute the chi-square test. You first compute the sample variance that you want to compare to the previous results:

```
> var(TENSE_ASPECT)¶
[1]·0.01687119
```

To test whether this value is significantly different from the known variance of 0.025, you compute a chi-square statistic as in formula (35).

(35) $$\chi^2 = \frac{(n-1)\cdot \textit{sample variance}}{\textit{population variance}}$$

This chi-square value has n-1 = 116 degrees of freedom. In R:

```
> chi.square<-((length(TENSE_ASPECT)-
      1)*var(TENSE_ASPECT))/0.025¶
> chi.square¶
[1]·78.28232
```

As usual, you can create those critical values yourself or you then look up this chi-square value in the familiar kind of table.

```
> qchisq(c(0.05,·0.01,·0.001),·116,·lower.tail=F)¶
```

Table 32. Critical chi-square values for $p_{\text{two-tailed}}$ = 0.05, 0.01, and 0.001 for 115 ≤ *df* ≤ 117

	p = 0.05	p = 0.01	p = 0.001
df = 115	141.03	153.191	167.61
df = 116	142.138	154.344	168.813
df = 117	143.246	155.496	170.016

Since the obtained value of 78.28 is much smaller than the relevant critical value of 142.138, the difference between the two variances is not significant. You can compute the exact p-value as follows:

```
> pchisq(chi.square,·(length(TENSE_ASPECT)-1),·lower.tail=F)¶
[1]·0.9971612¶
```

This is how you would summarize the result: “According to a chi-square test, the variance of the newly investigated child (0.017) does not differ significantly from the variance of the children investigated earlier (0.025): χ^2 = 78.28; *df* = 116; $p_{\text{two-tailed}}$ > 0.05.”

2.2. One dep. variable (ratio-scaled) and one indep. variable (nominal)

The probably more frequent scenario in the domain ‘testing dispersions’ is the case where you test whether two samples or two variables exhibit the same dispersion (or at last two dispersions that do not differ significantly. Since the difference of dispersions or variances is probably not a concept you spent much time thinking about so far, let us look at one illustrative example from the domain of sociophonetics). Gaudio (1994) studied the pitch range of heterosexual and homosexual men. At issue was therefore not the average pitch, but its variability, a good example for how variability as such can be interesting, In that study, four heterosexual and four homosexual men were asked to read aloud two text passages and the resulting recordings were played to 14 subjects who were asked to guess which speakers were heterosexual and which were homosexual. Interestingly, the subjects were able to distinguish the sexual orientation nearly perfectly. The only (insignificant) correlation which suggested itself as a possible explanation was that the homosexual men exhibited a wider pitch range in one of the text types, i.e., a result that has to do with variability and dispersion.

To get to know the statistical procedure needed for such cases we look at an example from the domain of second language acquisition. Let us assume you want to study how native speakers of a language and very advanced learners of that language differed in a synonym-finding task in which both native speakers and learners are presented with words for which they are asked to name synonyms. You may now be not be interested in the exact numbers of synonyms – maybe, the learners are so advanced that these are actually fairly similar in both groups – but in whether the learners exhibit more diversity in the amounts of time they needed to come up with all the synonyms they can name. This question involves

- a dependent ratio-scaled variable, namely SYNTIMES, the time subjects needed to name the synonyms;
- a nominal/categorical independent variable, namely SPEAKER: *LEARNER* and SPEAKER: *NATIVE*.

This kind of questions is investigated with the so-called *F*-test for homogeneity of variances, which involves the following steps.

Procedure
Formulating the hypotheses
Computing the sample variance; inspecting a graph
Testing the assumption(s) of the test:
the population from which the sample mean has been drawn or at least the sample itself is normally distributed
the samples are independent of each other
Computing the test statistic *t*, the degrees of freedom *df*, and the probability of error *p*

First, you formulate the hypotheses. Note that the alternative hypothesis is non-directional / two-tailed.

H_0: The times the learners need to name the synonyms they can think of are not differently variable from the times the native speakers need to name the synonyms they can think of; $sd^2_{\text{learner}} = sd^2_{\text{native}}$.

H_1: The times the learners need to name the synonyms they can think of are differently variable from the times the native speakers need to name the synonyms they can think of; $sd^2_{\text{learner}} \neq sd^2_{\text{native}}$.

As an example, we use the (fictitious) data in the file <C:/_sflwr/_inputfiles/04-2-2_synonymtimes.txt>:

```
> SynonymTimes<-read.table(choose.files(),·header=T,·
       sep="\t",·comment.char="",·quote="")¶
> attach(SynonymTimes);·str(SynonymTimes)¶
`data.frame':···80·obs.·of··3·variables:
·$·CASE····:·int··1·2·3·4·5·6·7·8·9·10·...
·$·SPEAKER·:·Factor·w/·2·levels·"Learner","Native":·1·1·1·...
·$·SYNTIMES:·int··11·7·11·11·8·4·7·10·12·7·...
```

You compute the variances for both subject groups and plot the data:

```
> tapply(SYNTIMES,·SPEAKER,·var)¶
·Learner···Native
10.31731·15.75321
> boxplot(SYNTIMES~SPEAKER,·notch=T)¶
> rug(jitter(SYNTIMES),·side=2)¶
```

At first sight, the data are very similar to each other: the medians are very close to each other, each median is within the notch of the other, the boxes have similar sizes, only the ranges of the whiskers differ.

The *F*-test requires a normal distribution of the population or at least the sample. We again use the Shapiro-Wilk test from Section 4.1.1.1.

```
> shapiro.test(SYNTIMES[SPEAKER=="Learner"])¶
········Shapiro-Wilk·normality·test
data:··SYNTIMES[SPEAKER·==·"Learner"]
W·=·0.9666,·p-value·=·0.279
> shapiro.test(SYNTIMES[SPEAKER=="Native"])¶
········Shapiro-Wilk·normality·test
data:···SYNTIMES[SPEAKER·==·"Native"]
W·=·0.9774,·p-value·=·0.5943
```

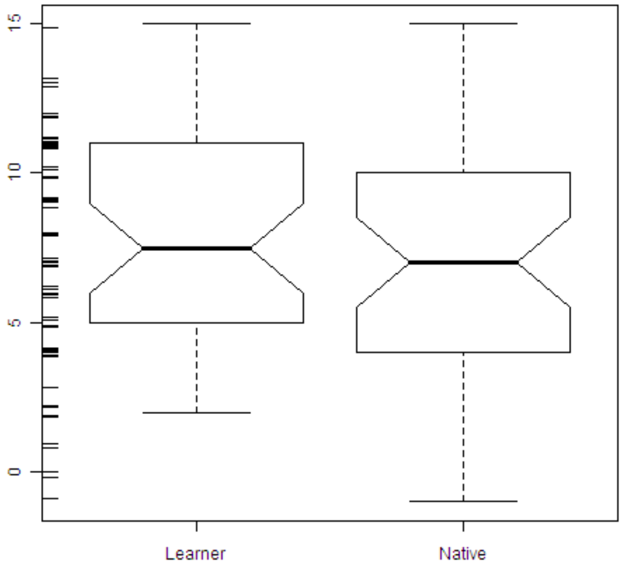

Figure 48. Boxplot for SYNTIMES~SPEAKER

By the way, this way of doing the Shapiro-Wilk test is not particularly elegant – can you think of a better one?

THINK BREAK

In Section 3.2.2 above, we used the function `tapply`, which allows you to apply a function to elements of a vector that are grouped according to another vector/factor. You can therefore write:

```
>tapply(SYNTIMES,·SPEAKER,·shapiro.test)¶
$Learner
········Shapiro-Wilk·normality·test
data:··X[[1L]]
W·=·0.9666,·p-value·=·0.2791
$Native
········Shapiro-Wilk·normality·test
data:··X[[2L]]
W·=·0.9774,·p-value·=·0.5943
```

Nothing to worry about: both samples do not deviate significantly from normality and you can do an *F*-test. This test requires you to compute the quotient of the two variances (traditionally, the larger variance is used as the numerator). You can therefore adapt your statistical hypotheses:

H_0: $^{variance1}/_{variance2} = F = 1$
H_1: $^{variance1}/_{variance2} = F \neq 1$

If the result is significant, you must reject the null hypothesis and consider the variances as heterogeneous – if the result is not significant, you must not accept the alternative hypothesis: the variances are homogeneous.

```
> F.val<-var(SYNONYME[SPEAKER=="Native"])/
      var(SYNONYME[SPEAKER=="Learner"]);·F.val¶
[1]·1.526872
```

You again need to consider degrees of freedom, this time even two: one for the numerator, one for the denominator. Both can be computed very easily by just subtracting one from the sample sizes (of the samples for the variances); cf. the formulae in (36).

(36) $df_{\text{numerator}} = n_{\text{numerator sample}}\text{-}1;\ df_{\text{denominator}} = n_{\text{denominator sample}}\text{-}1$

You get 39 in both cases and can look up the result in an *F*-table.

Table 33. Critical *F*-values for $p_{\text{two-tailed}} = 0.05$ and $38 \leq df_{1,\,2} \leq 40$

	$df_2 = 38$	$df_2 = 39$	$df_2 = 40$
$df_1 = 38$	1.907	1.8963	1.8862
$df_1 = 39$	1.9014	1.8907	1.8806
$df_1 = 40$	1.8961	1.8854	1.8752

Obviously, the result is not significant: the computed *F*-value is smaller than the tabulated one for $p = 0.05$ (which is 1.8907). As usual, you can compute the critical *F*-values yourself, and you would have to use the function `qf` for that. We need four arguments:

- `p`: the *p*-value for which you want to determine the critical *F*-value (for some *df*-values);
- `df1` and `df2`: the two *df*-values for the *p*-value for which you want to determine the critical *F*-value;
- the argument `lower.tail=F`, to instruct R to only consider the area under the curve above / to the right of the relevant *F*-value.

There is one last thing, though. When we discussed one- and two-tailed tests in Section 1.3.4 above, I mentioned that in the graphical representation of one-tailed tests (cf. Figure 6 and Figure 8) you add the probabilities of the events you see when you move away from the expectation of the null hypothesis in one direction while in the graphical representation of two-tailed tests (cf. Figure 7 and Figure 9) you add the probabilities of the events you see when you move away from the expectation of the null hypothesis in both directions. The consequence of that was that prior knowledge that allowed you to formulate a directional alternative hypothesis was rewarded such that you needed a less extreme findings to get a significant result. This also means, however, that when you want to compute a two-tailed *p*-value using `lower.tail=F`, then you need the *p*-value for $^{0.05}/_{2}$ = 0.025. This value tells you which *F*-value cuts off 0.025 on the right side of the graph, but since a two-tailed test requires that you cut off the same area on the left side, too, this means that this is also the desired critical *F*-value for $p_{\text{two-tailed}} = 0.05$. Figure 49 illustrates this logic:

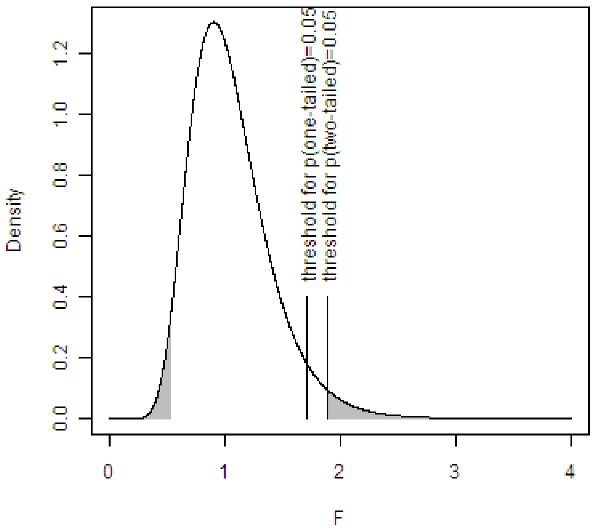

Figure 49. Density function for an *F*-distribution with $df_1 = df_2 = 39$, two-tailed test

The right vertical line indicates the *F*-value you need to obtain for a significant two-tailed test with $df_{1,\,2} = 39$; this *F*-value is the one you already know from Table 33 – 1.8907 – which means you get a significant two-tailed result if one variance is 1.8907 times larger than the other. The left vertical line indicates the *F*-value you need to obtain for a significant one-tailed test with $df_{1,\ 2}$ = 39; this *F*-value is 1.7045, which means you get a significant one-tailed result if the variance you predict to be larger (!) is 1.7045 times larger than the other. To compute the *F*-values for the two-tailed tests yourself, as a beginner you may want to enter just these lines and proceed in a similar way for all other cells in Table 33.

```
> qf(0.025,·39,·39,·lower.tail=T)·#·the·value·at·the·right
      margin·of·the·left·grey·area¶
[1]·0.5289
> qf(0.025,·39,·39,·lower.tail=F)·#·the·value·at·the·left
      margin·of·the·right·grey·area
[1]·1.890719
```

Alternatively, if you are more advanced already, you can generate all of Table 33 right away:

```
> p.values<-matrix(rep(0.025,·9),·byrow=T,·ncol=3)¶
> df1.values<-matrix(rep(c(38,·39,·40),·3),·byrow=F,·ncol=3)¶
> df2.values<-matrix(rep(c(38,·39,·40),·3),·byrow=T,·ncol=3)¶
> qf(p.values,·df1.values,·df2.values,·lower.tail=F)¶
·········[,1]·····[,2]·····[,3]
[1,]·1.907004·1.896313·1.886174
[2,]·1.901431·1.890719·1.880559
[3,]·1.896109·1.885377·1.875197
```

The observed *F*-value is obviously too small for a significant result: 1.53 < 1.89. It is more useful, however, to immediately compute the *p*-value for your *F*-value. Since you now use the reverse of `qf`, `pf`, you must now not divide by 2 but multiply by 2:

```
> 2*pf(F.val,·39,·39,·lower.tail=F)¶
[1]·0.1907904
```

As we've seen, with a *p*-value of *p* = 0.1908, the *F*-value of about 1.53 for $df_{1,\ 2}$ = 39 is obviously not significant. The function for the *F*-test in R that easily takes care of all of the above is called `var.test` and it requires at least two arguments, the two samples. Just like many other functions, you can approach this in two ways: you can provide R with a formula,

```
> var.test(SYNTIMES~SPEAKER)¶
········F·test·to·compare·two·variances
data:··SYNTIMES·by·SPEAKER
F·=·0.6549,·num·df·=·39,·denom·df·=·39,·p-value·=·0.1908
```

```
alternative·hypothesis:·true·ratio·of·variances·is·not·
      equal·to·1
95·percent·confidence·interval:
·0.3463941·1.2382959
sample·estimates:
ratio·of·variances
·········0.6549339
```

or you can use a vector-based alternative:

```
> var.test(SYNTIMES[SPEAKER=="Learner"],·
      SYNTIMES[SPEAKER=="Native"])¶
```

Do not be confused if the *F*-value you get from R is not the same as the one you computed yourself. Barring mistakes, the value outputted by R is then $^{1}/_{F}$-value – R does not automatically put the larger variance into the numerator, but the variance whose name comes first in the alphabet, which here is “Learner” (before “Native”). The *p*-value then shows you that R’s result is the same as yours. You can now sum this up as follows: “The learners synonym-finding times exhibit a variance that is approximately 50% larger than that of the native speakers (15.75 vs. 10.32), but according to an *F*-test, this different is not significant: $F = 1.53$; $df_{\text{learner}} = 39$; $df_{\text{native}} = 39$; $p_{\text{two-tailed}} = 0.1908$.”

Recommendation(s) for further study

- Dalgaard (2002: 89), Crawley (2007: 289ff.), Baayen (2008: Section 4.2.3)
- the function `fligner.test` to test the homogeneity of variance when the data violate the assumption of normality
- Good and Hardin (2006: 61f., 67–70) for other possibilities to compare variances, to compensate for unequal variances, and for discussion of the fact that unequal variances can actually be more interesting than unequal means

3. Means

The probably most frequent use of simple significance tests apart from chi-square tests are tests of differences between means. In Section 4.3.1, we will be concerned with goodness-of-fit tests, i.e., scenarios where you test whether an observed measure of central tendency is significantly different from another already known mean (recall this kind of question from Section 3.1.5.1); in Section 4.3.2, we then turn to tests where measures of central tendencies from two samples are compared to each other.

3.1. Goodness-of-fit tests

3.1.1. One dep. variable (ratio-scaled)

Let us assume you are again interested in the use of hedges. Early studies suggested that men and women exhibit different communicative styles with regard to the frequency of hedges (and otherwise). Let us also assume you knew from the literature that female subjects in experiments used on average 12 hedges in a two-minute conversation with a female confederate of the experimenter. You also knew that the frequencies of hedges are normally distributed. You now did an experiment in which you recorded 30 two-minute conversations of female subjects with a male confederate and counted the same kinds of hedges as were counted in the previous studies (and of course we assume that with regard to all other parameters, your experiment was an exact replication of the earlier standards of comparison). The average number of hedges you obtain in this experiment is 14.83 (with a standard deviation of 2.51). You now want to test whether this average number of hedges of yours is significantly different from the value of 12 from the literature. This question involves

- a dependent ratio-scaled variable, namely NUMBERHEDGES, which will be compared to the value from the literature;
- no independent variable since you do not test whether NUMBERHEDGES is influenced by something else.

For such cases, you use a one-sample *t*-test, which involves these steps:

Procedure
Formulating the hypotheses
Testing the assumption(s) of the test: the population from which the sample mean has been drawn or at least the sample itself is normally distributed
Computing the test statistic *t*, the degrees of freedom *df*, and the probability of error *p*

As always, you begin with the hypotheses.

H_0: The average of NUMBERHEDGES in the conversations of the subjects with the male confederate does not differ significantly from the already known average; hedges in your experiment = 12, or hedges in your experiment-12 = 0.

H_1: The average of NUMBERHEDGES in the conversations of the subjects with the male confederate differs significantly from the previously reported average; hedges in your experiment ≠ 12, or hedges in your experiment-12 ≠ 0.

Then you load the data from <C:/_sflwr/_inputfiles/04-3-1-1_hedges.txt>:

```
> Hedges<-read.table(choose.files(),·header=T,·sep="\t",·
      comment.char="",·quote="")¶
> attach(Hedges)¶
```

While the literature mentioned that the numbers of hedges are normally distributed, you test whether this holds for your data, too:

```
> shapiro.test(HEDGES)¶
········Shapiro-Wilk·normality·test
data:··HEDGES
W·=·0.946,·p-value·=·0.1319
```

It does. You can therefore immediately proceed with computing the *t*-value using the formula in (37).

(37) $$t = \frac{\bar{x}_{sample} - \bar{x}_{population}}{sd_{sample} \Big/ \sqrt{n_{sample}}}$$

```
> numerator<-mean(HEDGES)-12¶
> denominator<-sd(HEDGES)/sqrt(length(HEDGES))¶
> abs(numerator/denominator)¶
[1]·6.191884
```

To see what this value means, we need degrees of freedom again. Again, this is easy here since *df* = *n*-1, i.e., *df* = 29. When you look up the *t*-value for *df* = 29 in the usual kind of table, the *t*-value you computed must again be larger than the one tabulated for your *df* at *p* = 0.05.

Table 34. Critical *t*-values for $p_{\text{two-tailed}} = 0.05, 0.01$, and 0.001 for $28 \leq df \leq 30$

	p = 0.05	*p* = 0.01	*p* = 0.001
df = 28	2.0484	2.7633	3.6739
df = 29	2.0452	2.7564	3.6594
df = 30	2.0423	2.75	3.646

To compute the exact *p*-value, you can use `qt` with the *p*-value and the required *df*-value. Since you do a two-tailed test, you must cut off $^{0.05}/_{2}$ = 2.5% on both sides of the distribution, which is illustrated in Figure 50.

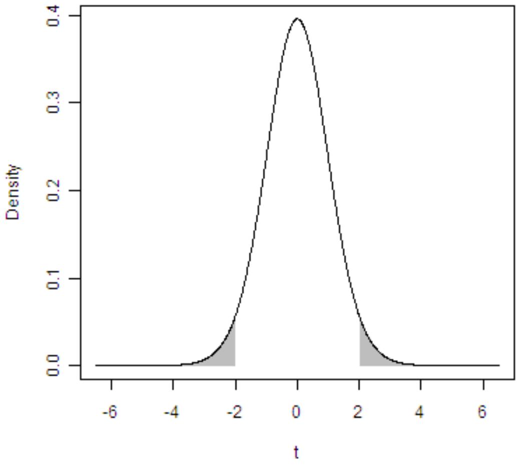

Figure 50. Density function for a *t*-distribution for *df* = 29, two-tailed test

The critical *t*-value for *p* = 0.025 and *df* = 29 is therefore:

```
> qt(c(0.025,·0.0975),·29,·lower.tail=F)·#·note·that·0.05·is·
       again·divided·by·2!¶
[1]··2.045230·-2.045230
```

The exact *p*-value can be computed with `pt` and the obtained *t*-value is highly significant: 6.1919 is not just larger than 2.0452, but even larger than the *t*-value for *p* = 0.001 and *df* = 29. You could also have guessed that because the *t*-value of approx. 6.2 is very far in the right grey margin in Figure 50.

To sum up: “On average, female subjects that spoke to a male confederate of the experimenter for two minutes used 14.83 hedges (standard deviation: 2.51). According to a one-sample *t*-test, this average is highly significantly larger than the value previously noted in the literature (for female subjects speaking to a female confederate of the experimenter): $t = 6.1919$; $df = 29$; $p_{\text{two-tailed}} < 0.001$."

```
> 2*pt(6.191884,·29,·lower.tail=F)·#·note·that·the·t-value·
      is·multiplied·with·2!¶
[1]·9.42153e-07
```

With the right function in R, you need just one line. The relevant function is called `t.test` and requires the following arguments:

- `x`: a vector with the sample data;
- `mu=`..., i.e., the population mean to which the sample mean computed from `x` is to be compared;
- `alternative="two-sided"` for two-tailed tests (the default) or one of `alternative="greater"` or `alternative="less"`, depending on which alternative hypothesis you want to test: `alternative="less"` states the sample mean is smaller than the population mean, and `alternative="greater"` states that the sample mean is larger than the population mean respectively.

```
> t.test(HEDGES,·mu=12)¶
········One·Sample·t-test
data:··HEDGES
t·=·6.1919,·df·=·29,·p-value·=·9.422e-07
alternative·hypothesis:·true·mean·is·not·equal·to·12
95·percent·confidence·interval:
·13.89746·15.76921
sample·estimates:
mean·of·x
·14.83333
```

You get the already known mean of 14.83 as well as the *df*- and *t*-value we computed semi-manually. In addition, we get the exact *p*-value and the confidence interval of the mean which, and that is why the result is significant, does not include the tested value of 12.

Recommendation(s) for further study
Dalgaard (2002: 81ff.), Baayen (2008: Section 4.1.2)

3.1.2. One dep. variable (ordinal)

In the previous section, we discussed a test that allows you to test whether the mean of a sample from a normally-distributed population is different from an already known population mean. This section deals with a test you can use when the data violate the assumption of normality or when they are not ratio-scaled to begin with. We will explore this test by looking at an interesting little morphological phenomenon, namely subtractive word-formation processes in which parts of usually two source words are merged into a new word. Two such processes are blends and complex clippings; some well-known examples of the former are shown in (38a), while (38b) provides a few examples of the latter; in all examples, the letters of the source words that enter into the new word are underlined.

(38) a. *brunch* (*breakfast* × *lunch*), *motel* (*motor* × *hotel*), *smog* (*smoke* × *fog*), *foolosopher* (*fool* × *philosopher*)
b. *scifi* (*science* × *fiction*), *fedex* (*federal* × *express*), *sysadmin* (*system* × *administrator*)

One question that may arise upon looking at these coinages is to what degree the formation of such words is supported by some degree of similarity of the source words. There are many different ways to measure the similarity of words, and the one we are going to use here is the so-called Dice coefficient (cf. Brew and McKelvie 1996). You can compute a Dice coefficient for two words in two simple steps. First, you split the words up into letter (or phoneme or …) bigrams. For *motel* (*motor* × *hotel*) you get:

- *motor*: *mo*, *ot*, *to*, *or*;
- *hotel*: *ho*, *ot*, *te*, *el*.

Then you count how many of the bigrams of each word occur in the other word, too. In this case, these are two: the *ot* of *motor* also occurs in *hotel*, and thus the *ot* of *hotel* also occurs in *motor*.[28] This number, 2, is divided by the number of bigrams to yield the Dice coefficient:

(39) $Dice_{motor\ \&\ hotel} = {}^{2}/_{8} = 0.25$

[28] In R, such computations can be easily automated and done for hundreds of thousands of words. For example, for any one word contained in a vector `a`, this line returns all its bigrams: `substr(rep(a,·nchar(a)-1),·1:(nchar(a)-1),·2:(nchar(a)))¶`; for many such applications, cf. Gries (2009).

In other words, the Dice coefficient is the percentage of shared bigrams out of all bigrams (and hence ratio-scaled). We will now investigate the question of whether source words that entered into subtractive word-formation processes are more similar to each other than words in general are similar to each other. Let us assume, you know that the average Dice coefficient of randomly chosen words is 0.225 (with a standard deviation of 0.0809; the median is 0.151 with an interquartile range of 0.125). These figures already suggest that the data may not be normally distributed.[29]

This study involves

- a dependent ratio-scaled variable, namely the SIMILARITY of the source words, which will be compared with the already known mean/median;
- no independent variable since you do not test whether SIMILARITY is influenced by something else.

The hypotheses should be straightforward:

H_0: The average of SIMILARITY for the source words that entered into subtractive word-formation processes is not significantly different from the known average of randomly chosen word pairs; Dice coefficients of source words = 0.225, or Dice coefficients of source words-0.225 = 0.

H_1: The average of SIMILARITY for the source words that entered into subtractive word-formation processes is different from the known average of randomly chosen word pairs; Dice coefficients of source words ≠ 0.225, or Dice coefficients of source words-0.225 ≠ 0.

The data to be investigated here are in <C:/_sflwr/_inputfiles/04-3-1-2_dices.txt> ; they are data of the kind studied in Gries (2006).

```
> Dices<-read.table(choose.files(),·header=T,·sep="\t",·
      comment.char="",·quote="")¶
> attach(Dices);·str(Dices)¶
'data.frame':···100·obs.·of··2·variables:
·$·CASE:·int··1·2·3·4·5·6·7·8·9·10·...
·$·DICE:·num··0.19·0.062·0.06·0.064·0.101·0.147·0.062·...
```

From the summary statistics, you could already infer that the similarities of randomly chosen words are not normally distributed. We can therefore

[29] For authentic data, cf. Gries (2006), where I computed Dice coefficients for all 499,500 possible pairs of 1,000 randomly chosen words.

assume that this is also true of the sample of source words, but of course you also test this assumption:

```
> shapiro.test(DICE)¶
········Shapiro-Wilk·normality·test
data:··DICE
W·=·0.9615,·p-value·=·0.005117
```

The Dice coefficients are not normally, but symmetrically distributed (as you could also clearly see in, say, a histogram by entering `hist(DICE)¶`). Thus, even though Dice coefficients are ratio-scaled and although the sample size is larger than 30 (cf. also Section 4.3.2 below), you may want to be careful/conservative and not use the one-sample *t*-test but, for example, the so-called one-sample sign test, which involves the following steps:

Procedure

Formulating the hypotheses
Computing the frequencies of the signs of the differences between the observed values and the expected average
Computing the probability of error *p*

You first rephrase the hypotheses; I only provide the new statistical versions:

H$_0$: $median_{\text{Dice coefficients of source words}} = 0.151$.
H$_1$: $median_{\text{Dice coefficients of source words}} \neq 0.151$.

Then, you compute the median and its interquartile range:

```
> median(DICE);·IQR(DICE)¶
[1]·0.1775
[1]·0.10875
```

Obviously, the observed median Dice coefficient is a bit higher than 0.151, the median Dice coefficient of the randomly chosen word pairs, but it is hard to guess whether the difference is going to be significant or not. Hence, we do the required test. For the one-sample sign test, you first determine how many observations are above and below the expected median, because if the expected median was a good characterization of the observed data, then 50% of the observed data should be above the expected median and 50% should be below it. (NB: you must realize that this means that the exact sizes of the deviations from the expected median are not considered here – you only look at whether the observed values are larger or smaller than the expected median, but not how much larger or smaller.)

```
> sum(DICE>0.151)¶
[1]·63
```

63 of the 100 observed values are larger than the expected median – since you expected 50, it seems as if the Dice coefficients observed in the source words are significantly larger than those of randomly chosen words. As before, this issue can also be approached graphically, using the logic and the function `dbinom` from Section 1.3.4.1, Figure 6 and Figure 8. Figure 51 shows the probabilities of all possible results you can get in 100 trials – because you look at the Dice coefficients of 100 subtractive word formations, but consider the left panel of Figure 51 first. According to H_0, you would expect 50 Dice coefficients to be larger than the expected median, but you found 63. Thus, you add the probability of the observed result (the black bar for 63 out of 100) to the probabilities of all those that deviate from H_0 even more extremely, i.e., the chances to find 64, 65, …, 99, 100 Dice coefficients out of 100 that are larger than the expected median. These probabilities from the left panel sum up to approximately 0.006:

```
> sum(dbinom(63:100,·100,·0.5))¶
[1]·0.006016488
```

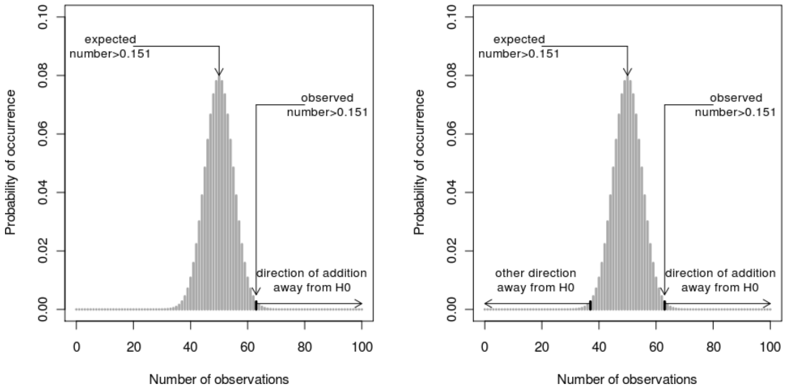

Figure 51. Probability distributions for 100 binomial trials test

But you are not finished yet … As you can see in the left panel of Figure 51, so far you only include the deviations from H_0 in one direction – the right – but your alternative hypothesis is non-directional, i.e., two-tailed. To do a two-tailed test, you must therefore also include the probabilities of the events that deviate just as much and more from H_0 in the other direction: 37, 36, …, 1, 0 Dice coefficients out of 100 that are smaller than the expected median, as represented in the right panel of Figure 51. The probabilities sum up to the same value (because the distribution of binomial probabilities around $p = 0.5$ is symmetric).

```
> sum(dbinom(37:0,·100,·0.5))¶
[1]·0.006016488
```

Again: if you expect 50 out of 100, but observe 63 out of 100, and want to do a two-tailed test, then you must add the summed probability of finding 63 to 100 larger Dice coefficients (the upper/right 38 probabilities) to the summed probability of finding 0 to 37 smaller Dice coefficients (the lower/left 38 probabilities). As a result, you get a $p_{\text{two-tailed}}$-value of 0.01203298, which is obviously significant. You can sum up: “The investigation of 100 subtractive word formations resulted in an average source-word similarity of 0.1775 (median, *IQR* = 0.10875). 63 of the 100 source words were more similar to each other than expected, which, according to a two-tailed sign test is a significant deviation from the average similarity of random word pairs (median =0.151, *IQR* range = 0.125): $p_{\text{binomial}} = 0.012$.”

Recall that this one-sample sign test only uses nominal information, whether each data point is larger or smaller than the expected reference median. If the distribution of the data is rather symmetrical – as it is here – then there is an alternative test that also takes the sizes of the deviations into account, i.e. uses at least ordinal information. This so-called one-sample signed-rank test can be computed using the function `wilcox.test`. Apart from the vector to be tested, the following arguments are relevant:

- `alternative`: a character string saying which alternative hypothesis you want to test: the default is `"two.sided"`, other possible values for one-tailed tests are `"less"` or `"greater"`, which specify how the first-named vector relates to the specified reference median;
- `mu=…`: the reference median expected according to H_0;
- `exact=T`, if you want to compute an exact test (rather than an estimation; only when your sample size is smaller than 50) or `exact=F`, if an asymptotic test is sufficient; the latter is the default;
- `correct=T` for a continuity correction or `correct=F` for none;

- `conf.level`: a value between 0 and 1 specifying the size of the confidence interval; the default is 0.95.

Since you have a non-directional alternative hypothesis, you do a two-tailed test by simply adopting the default setting for alternative:

```
> wilcox.test(DICE,·mu=0.151,·correct=F)¶
········Wilcoxon·signed·rank·test·with·continuity·correction
data:··DICE
V·=·3454.5,·p-value·=·0.001393
alternative·hypothesis:·true·location·is·not·equal·to·0.151
```

The test confirms the previous result: both the one-sample sign test, which is only concerned with the directions of deviations, and the one-sample signed rank test, which also considers the sizes of these deviations, indicate that the source words of the subtractive word-formations are more similar to each other. This should however, encourage you to make sure you formulate exactly the hypothesis you are interested in (and then use the required test).

Recommendation(s) for further study

- for the sake of completeness, there is a slightly better function for the Wilcoxon-test, `wilcox.exact` (from the `library(exactRankTests)`, which is not under development anymore, but the successor package, `coin`, doesn't have `wilcox.ecact` (yet)). Although `wilcox.test` can take the argument `exact=T`, this function still has problems with ties – `wilcox.exact` does not and is thus sometimes preferable
- Dalgaard (2002: 85f.), Baayen (2008: Section 4.1.2)
- for the one-sample sign test, you may also want to read up on what to do in cases when one or more of the observed values is exactly as large as the expected median (e.g. in Marascuilo and McSweeney 1977)

3.2. Tests for differences/independence

A particularly frequent scenario requires you to test two groups of elements with regard to whether they differ in their central tendency. There are again several factors that determine which test to choose:

- the kind of samples: dependent or independent (cf. Section 1.3.4.1);

- the level of measurement of the dependent variable: interval/ratio-scaled vs. ordinal;
- the distribution of (interval/ratio-scaled) dependent variable: normal vs. non-normal;
- the sample sizes.

The first factor can be dealt with in isolation, but the others are interrelated. Simplifying a bit: is the dependent variable ratio-scaled as well as normally-distributed or both sample sizes are larger than 30 or are the differences between variables normally distributed, then you can usually do a *t*-test (for independent or dependent samples, as required) – otherwise you must do a *U*-test (for independent samples) or a Wilcoxon test (for dependent samples). The reason for this decision procedure is that while the *t*-test for independent samples requires, among other things, normally distributed samples, one can also show that means of samples of 30+ elements are usually normally distributed even if the samples as such are not, which was why we Section 4.3.1.2 at least considered the option of a one-sample *t*-test (and then chose the more conservative sign test or one-sample signed-rank test). Therefore, it is sufficient if the data meet one of the two conditions. Strictly speaking, the *t*-test for independent samples also requires homogenous variances, which we will also test for, but we will discuss a version of the *t*-test that can handle heterogeneous variances, the *t*-test after Welch.

3.2.1. One dep. variable (ratio-scaled) and one indep. variable (nominal) (indep. samples)

The *t*-test for independent samples is one of the most widely used tests. To explore it, we use an example from the domain of phonetics. Let us assume you wanted to study the (rather trivial) non-directional alternative hypothesis that the first formants' frequencies of men and women differed. You plan an experiment in which you record men's and women's pronunciation of a relevant set of words and/or syllables, which you then analyze with a computer (using Audacity or SpeechAnalyzer or …). This study involves

- one dependent ratio-scaled variable, namely F1-FREQUENCIES, whose averages you are interested in;
- one independent nominal variable, namely SEX: *MALE* vs. SEX: *FEMALE*;
- independent samples since, if every subject provides just one data point, the data points are not related to each other.

The test to be used for such scenarios is the *t*-test for independent samples and it involves the following steps:

Procedure

Formulating the hypotheses
Computing the relevant means; inspecting a graph
Testing the assumption(s) of the test:

- the population from which the sample has been drawn or at least the sample is normally distributed (esp. with samples of *n* < 30)
- the variances of the populations from which the samples have been drawn or at least the variances of the samples are homogeneous

Computing the test statistic *t*, the degrees of freedom *df*, and the probability of error *p*

You begin with the hypotheses.

H$_0$: The average F1 frequency of men is the same as the average F1 frequency of women: $mean_{\text{F1 frequency of men}} = mean_{\text{F1 frequency of women}}$, or $mean_{\text{F1 frequency of men}}\text{-}mean_{\text{F1 frequency of men}} = 0$.

H$_1$: The average F1 frequency of men is not the same as the average F1 frequency of women: $mean_{\text{F1 frequency of men}} \neq mean_{\text{F1 frequency of women}}$, or $mean_{\text{F1 frequency of men}}\text{-}mean_{\text{F1 frequency of men}} \neq 0$.

The data you will investigate here are part of the data borrowed from a similar experiment on vowels in Apache. First, you load the data from <C:/_sflwr/_inputfiles/04-3-2-1_f1-freq.txt> into R:

```
> Vowels<-read.table(choose.files(),·header=T,·sep="\t",·
      comment.char="",·quote="")¶
> attach(Vowels);·str(Vowels)¶
'data.frame':···120·obs.·of··3·variables:
·$·CASE·····:·int··1·2·3·4·5·6·7·8·9·10·...
·$·HZ_F1····:·num··489·558·425·626·531·...
·$·SEX:·Factor·w/·2·levels·"F","M":·2·2·2·2·2·2·2·2·2·2·...
```

Then, you compute the relevant means of the frequencies. As usual, the less elegant way to proceed is this,

```
> mean(HZ_F1[SEX=="F"])¶
> mean(HZ_F1[SEX=="M"])¶
```

… and, as usual, we use the more elegant variant with `tapply`.

```
> tapply(HZ_F1,·SEX,·mean)¶
·······F·········M
528.8548··484.2740
```

To get a better impression of what the data look like, you also immediately generate a boxplot. You set the limits of the *y*-axis such that it ranges from 0 to 1,000 so that all values are included and the representation is maximally unbiased; in addition, you use `rug` to plot the values of the women and the men onto the left and right *y*-axis respectively; cf. Figure 52 and the code file for an alternative that includes a stripchart.

```
> boxplot(HZ_F1~SEX,·notch=T,·ylim=(c(0,·1000)));·grid()¶
> rug(HZ_F1[SEX=="F"],·side=2)¶
> rug(HZ_F1[SEX=="M"],·side=4)¶
```

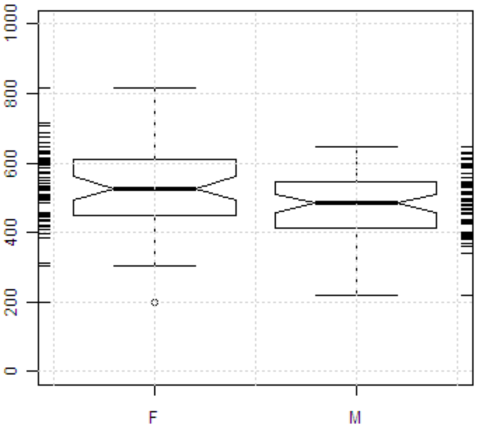

Figure 52. Boxplot for HZ_F1~SEX

The next step consists of testing the assumptions of the *t*-test. Figure 52 suggests that these data meet the assumptions. First, the boxplots for the men and the women appear as if the data are normally distributed: the medians are in the middle of the boxes and the whiskers extend nearly equally long in both directions. Second, the variances seem to be very similar since the sizes of the boxes and notches are very similar. However, of course you need to test this and you use the familiar Shapiro-Wilk test:

```
> tapply(HZ_F1,·SEX,·shapiro.test)¶
$F
```

```
········Shapiro-Wilk·normality·test
data:··X[[1L]]
W·=·0.987,·p-value·=·0.7723
$M
········Shapiro-Wilk·normality·test
data:··X[[2L]]
W·=·0.9724,·p-value·=·0.1907
```

The data do not differ significantly from normality. Now you test for variance homogeneity with the *F*-test from Section 4.2.2 (whose assumption of normality we now already tested). This test's hypotheses are:

H_0: The variance of the first sample equals that of the second; $F = 1$.
H_1: The variance of one sample is bigger than that of the second; $F \neq 1$.

The *F*-test with R yields the following result:

```
> var.test(HZ_F1~SEX)·#·with·a·formula¶
········F·test·to·compare·two·variances
data:··HZ_F1·by·SEX
F·=·1.5889,·num·df·=·59,·denom·df·=·59,·p-value·=·0.07789
alternative·hypothesis:·true·ratio·of·variances·is·not·
       equal·to·1
95·percent·confidence·interval:
·0.949093·2.660040
sample·estimates:
ratio·of·variances
·········1.588907
```

The second assumption is also met if only just about: since the confidence interval includes 1 and the *p*-value points to a non-significant result, the variances are not significantly different from each other and you can compute the *t*-test for independent samples. This test involves three different statistics: the test statistic *t*, the number of degrees of freedom *df*, and of course the *p*-value. In the case of the *t*-test we discuss here, the *t*-test after Welch, the *t*-value is computed according to formula (40), where sd^2 is the variance, *n* is the sample size, and the subscripts 1 and 2 refer to the two samples of men and women.

(40) $$t = \left| \left(\bar{x}_1 - \bar{x}_2\right) \div \sqrt{{sd_1}^2/_{n_1} + {sd_2}^2/_{n_2}} \right|$$

In R:

```
> t.numerator<-mean(HZ_F1[SEX=="M"])-mean(HZ_F1[SEX=="W"])¶
> t.denominator<-sqrt((var(HZ_F1[SEX=="M"])/
      length((HZ_F1[SEX=="M"])))+(var(HZ_F1[SEX=="W"])/
      length((HZ_F1[SEX=="W"]))))¶
> t<-abs(t.numerator/t.denominator)¶
```

You get t = 2.441581. The formula for the degrees of freedom is somewhat more complex. First, you need to compute a value called c, and with c, you can then compute df. The formula to compute c is shown in (41), and the result of 41 gets inserted into (42).

$$(41)\qquad c = \frac{{sd_1}^2 / n_1}{{sd_1}^2 / n_1 + {sd_2}^2 / n_2}$$

$$(42)\qquad df = \left(\frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1} \right)^{-1}$$

```
> c.numerator<-
      var(HZ_F1[SEX=="M"])/length((HZ_F1[SEX=="M"]))¶
> c.denominator<-t.denominator^2¶
> c<-c.numerator/c.denominator¶
> df.summand1<-c^2/(length(HZ_F1[SEX=="M"])-1)¶
> df.summand2<-((1-c)^2)/(length(HZ_F1[SEX=="F"])-1)¶
> df<-(df.summand1+df.summand2)^-1¶
```

You get c = 0.3862634 and df = 112.1946 ≈ 112. You can then look up the t-value in the usual kind of t-table (cf. Table 35) or you can compute the critical t-value in R (with `qt(c(0.025,·0.975),·112,·lower.tail=F)¶`, as before, for a two-tailed test you compute the t-value for the p-value of 0.025).

Table 35. Critical t-values for $p_{\text{two-tailed}}$ = 0.05, 0.01, and 0.001 for $111 \leq df \leq 113$

	p = 0.05	p = 0.01	p = 0.001
df = 111	1.9816	2.6208	3.3803
df = 112	1.9814	2.6204	3.3795
df = 113	1.9812	2.62	3.3787

As you can see, the observed t-value is larger than the one tabulated for p = 0.05, but smaller than the one tabulated for p = 0.01: the difference between the means is significant. The exact *p*-value can be computed with `qt` and for the present two-tailed case you simply enter this:

```
> 2*pt(t,·112,·lower.tail=F)¶
[1]·0.01618811
```

In R, you can do all this with the function `t.test`. This function takes several arguments, the first two of which – the relevant samples – can be given by means of a formula or with two vectors. These are the other relevant arguments:

- `alternative`: a character string that specifies which alternative hypothesis is tested: the default value, which can therefore be omitted, is `"two.sided"`, other values for one-tailed hypotheses are again `"less"` or `"greater"`; as before, R considers the alphabetically first variable level (i.e., here “F”) as the reference category so that the one-tailed hypothesis that the values of the men are smaller than those of the women would be tested with `alternative="greater"`;
- `paired=F` for the *t*-test for independent samples (the default) or `paired=T` for the *t*-test for dependent samples (cf. the following section);
- `var.equal=T`, when the variances of the two samples are equal, or `var.equal=F` if they are not; the latter is the default, which should hardly be changed;
- `conf.level`: a value between 0 and 1, which specifies the confidence interval of the difference between the means; the default is 0.95.

Thus, to do the *t*-test for independent samples, you can enter either variant listed below. You get the following result:

```
> t.test(HZ_F1~SEX,·paired=F)·#·with·a·formula¶
········Welch·Two·Sample·t-test
data:··HZ_F1·by·SEX
t·=·2.4416,·df·=·112.195,·p-value·=·0.01619
alternative·hypothesis:·true·difference·in·means·is·
       not·equal·to·0
95·percent·confidence·interval:
·8.403651·80.758016
sample·estimates:
mean·in·group·F·mean·in·group·M
·······528.8548········484.2740
> t.test(HZ_F1[SEX=="F"],·HZ_F1[SEX=="M"],·paired=F)·#·
       with·vectors¶
```

The first two lines of the output provide the name of the test and the data to which the test was applied. Line 3 lists the test statistic *t* (the sign is irrelevant because it only depends on which mean is subtracted from which, but it must of course be considered for the manual computation), the *df*-value, and the *p*-value. Line 4 states the alternative hypothesis tested. Then, you get the confidence interval for the differences between means (and our test is significant because this confidence interval does not include 0). At the bottom, you get the means you already know.

To be able to compare our results with those of other studies while at the same time avoiding the risk that the scale of the measurements distorts our assessment of the observed difference, we also need an effect size. There are two possibilities. One is an effect size correlation, the correlation between the values of the dependent variable and the values you get if the levels of the independent variable are recoded as 0 and 1.

```
> SEX2<-ifelse(SEX=="M",·0,·1)¶
> cor.test(SEX2,·HZ_F1)¶
```

The result contains the same *t*-value and nearly the same *p*-value as before (only nearly the same because of the different *df*), but you now also get a correlation coefficient, which is, however, not particularly high: 0.219. Another widely used effect size is Cohen's *d*, which is computed as in (43):

(43) $$\text{Cohen's } d = \left| \frac{2t}{\sqrt{n_1 + n_2}} \right|$$

```
> d<-abs(2*t.test(HZ_F1~SEX,·paired=F)$stat/
       sqrt(length(SEX)))¶
```

Since Cohen's *d* can take on values between 0 and 1, the value of 0.446 reflects an only intermediately strong effect. You can sum up you results as follows: "In the experiment, the average F1 frequency of the vowels produced by men was 484.3 Hz (95% confidence interval 461.6; 507 Hz), the average F1 frequency of the vowels produced by the women was 528.9 Hz (95% confidence interval: 500.2; 557.5 Hz). According to a *t*-test for independent samples, the difference of 44.6 Hz between the means is statistically significant, but not particularly strong: $t_{\text{Welch}} = 2.4416$; $df = 112.2$; $p_{\text{two-tailed}} = 0.0162$; Cohen's $d = 0.446$."

In Section 5.3, we will discuss the extension of this test to cases where you have more than one independent variable and/or where the independent variable has more than two levels.

Recommendation(s) for further study

- Crawley (2007: 289ff.), Baayen (2008: Section 4.2.2)
- an exact variant of the *t*-test for independent samples, which does not make any distributional assumptions, can be programmed relatively easily in R using the function `combn` (from the `library(combinat)`) and is available from me upon request

3.2.2. One dep. variable (ratio-scaled) and one indep. variable (nominal) (dep. samples)

The previous section illustrated a test for means from two independent samples. The name of that test suggests that there is a similar test for dependent samples, which is what we will discuss in this section on the basis of an example from translation studies. Let us assume you want to compare the lengths of English and Portuguese texts and their respective translations into Portuguese and English. Let us also assume you suspect that the translations are on average longer than the originals. This question involves

- one dependent ratio-scaled variable, namely the LENGTH of the texts, the average of which we are interested in;
- one independent nominal/categorical variable, namely TEXTSOURCE: *ORIGINAL* vs. TEXTSOURCE: *TRANSLATION*;
- dependent samples since there is one translation for every original text.

Performing a *t*-test for dependent samples requires the following steps:

Procedure

Formulating the hypotheses
Computing the relevant means; inspecting a graph
Testing the assumption(s) of the test: the differences of the paired values are distributed normally
Computing the test statistic *t*, the degrees of freedom *df*, and the probability of error *p*

As usual, you formulate the hypotheses, but note that this time the alternative hypothesis is directional: you suspect that the average length of the originals is *shorter* than those of their translations, not just different (i.e., shorter or longer). Therefore, the statistical form of the alternative hypothesis does not just contain a “≠”, but something more specific, “<“:

H_0: The average of the pairwise differences between the lengths of the originals and the lengths of the translations is 0; $mean_{\text{pairwise differerences}} = 0$.

H_1: The average of the pairwise differences between the lengths of the originals and the lengths of the translations is smaller than 0; $mean_{\text{pairwise differerences}} < 0$.

Note in particular (i) that the hypotheses do not involve the values of the two samples but the pairwise differences between the samples and (ii) how these difference are computed: original minus translation, not the other way round (and hence we use “< 0”). To illustrate this test, we will look at data from Frankenberg-Garcia (2004). She compared the lengths of eight English and eight Portuguese texts, which were chosen and edited such that their lengths were approximately 1,500 words, and then determined the lengths of their translations. You can load the data from <C:/_sflwr/_inputfiles/04-3-2-2_textlengths.txt>:

```
> Texts<-read.table(choose.files(),·header=T,·sep="\t",·
      comment.char="",·quote="")¶
> attach(Texts);·str(Texts)¶
`data.frame':···32·obs.·of··5·variables:
·$·CASE······:·int··1·2·3·4·5·6·7·8·9·10·...
·$·LENGTH·:·int··1501·1499·1501·1498·1499·1499·1498·1500·...
·$·TEXT·······:·int··1·2·3·4·5·6·7·8·9·...
·$·TEXTSOURCE···:·Factor·w/·2·levels·"Original","Translation"
      :·1·...
·$·LANGUAGE:·Factor·w/·2·levels·"English","Portuguese":·1·1·1
      ·...
```

Note that the data are organized so that the order of the texts and their translations is identical: case 1 is an English original (hence, TEXT is 1, TEXTSOURCE is *ORIGINAL*, LANGUAGE is *ENGLISH*), and case 17 is its translation (hence, TEXT is again 1, TEXTSOURCE is now *TRANSLATION*, and LANGUAGE is *PORTUGUESE*), etc. First, you compute the means and generate a plot (note, this boxplot does not show the dependency of the samples).

```
> tapply(LENGTH,·TEXTSOURCE,·mean)¶
····Original··Translation
····1500.062·····1579.938
```

```
> boxplot(LENGTH~TEXTSOURCE,·notch=T,·ylim=c(0,·2000))¶
> rug(LENGTH,·side=2)¶
```

(Cf. the code file for alternative plots.) The median translation length is a little higher than that of the originals. Also, the two samples have *very* different dispersions because the lengths of the originals were set to approximately 1,500 words and thus exhibit very little variation while the lengths of the translations are much more variable by comparison.

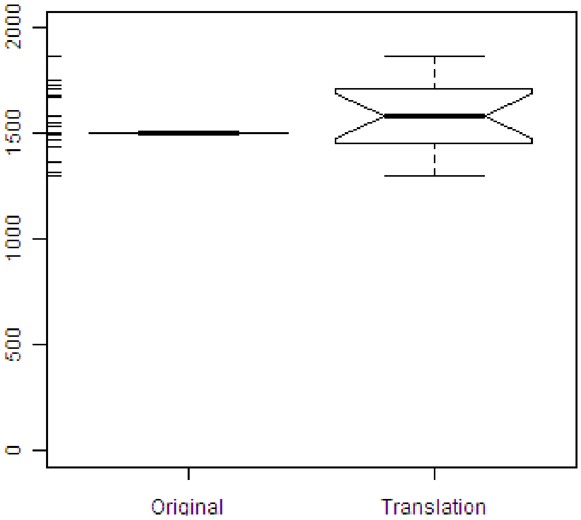

Figure 53. Boxplot for LENGTH~TEXTSOURCE

Unlike the *t*-test for independent samples, the *t*-test for dependent samples does not presuppose a normal distribution or variance homogeneity of the sample values, but a normal distribution of the differences between the pairs of sample values. You can create a vector with these differences and apply the Shapiro-Wilk test to it:

```
> differences<-LENGTH[1:16]-LENGTH[17:32]¶
> shapiro.test(differences)¶
········Shapiro-Wilk·normality·test
data:··differences
W·=·0.9569,·p-value·=·0.6057
```

The differences do not differ significantly from normality so you can in fact do the *t*-test for dependent samples. First, you compute the *t*-value according to the formula in (44), where *n* is the number of value pairs.

(44) $$t = {\left|\bar{x}_{diff}\right| \cdot \sqrt{n}} \Big/ {sd_{diff}}$$

```
> t<-(abs(mean(differences))*sqrt(length(differences)))/
       sd(differences);·t¶
[1]·1.927869
```

Second, you compute the degrees of freedom *df*, which is the number of differences *n* minus 1:

```
> df<-length(differences)-1;·df¶
[1]·15
```

First, you can now compute the critical values for $p = 0.05$ – this time *not* for ${}^{0.05}/_{2} = 0.025$ – at $df = 15$ or, in a more sophisticated way, create the whole *t*-table.

```
> qt(c(0.05,·0.95),·15,·lower.tail=F)¶
[1]··1.753050·-1.753050

> p.values<-matrix(rep(c(0.05,·0.01,·0.001),·3),·
       byrow=T,·ncol=3)¶
> df.values<-matrix(rep(14:16,·each=3),·byrow=T,·ncol=3)¶
> qt(p.values,·df.values,·lower.tail=F)¶
·········[,1]·····[,2]·····[,3]
[1,]·1.761310·2.624494·3.787390
[2,]·1.753050·2.602480·3.732834
[3,]·1.745884·2.583487·3.686155
```

Second, you can look up the your *t*-value in such a *t*-table, repeated here as Table 36. Since such tables usually only list the positive values, you use the absolute value of your *t*-value. As you can see, the differences between the originals and their translations is significant, but not very or highly significant: 1.927869 > 1.7531, but 1.927869 < 2.6025.

Table 36. Critical *t*-values for $p_{\text{one-tailed}} = 0.05$, 0.01, and 0.001 (for $14 \le df \le 16$)

	$p = 0.05$	$p = 0.01$	$p = 0.001$
$df = 14$	1.7613	2.6245	3.7874
$df = 15$	1.7531	2.6025	3.7328
$df = 16$	1.7459	2.5835	3.6862

Alternatively, you can compute the exact *p*-value. Since you have a directional alternative hypothesis, you only need to cut off 5% of the area under the curve on one side of the distribution. The *t*-value following from the null hypothesis is 0 and the *t*-value you computed is approximately -1.93 so you must compute the area under the curve from to 1.93 to +∞; cf. Figure 54. Since you are doing a one-tailed test, you need not multiply the *p*-value with 2 as you did above in Sections 4.2.2, 4.3.1.1, and 4.3.2.1.

```
> pt(t,·15,·lower.tail=F)¶
[1]·0.03651145
```

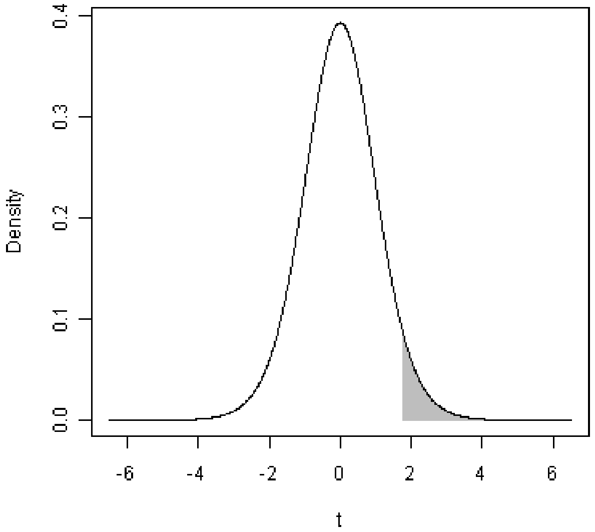

Figure 54. Density function for a *t*-distribution for *df* = 15, one-tailed test

Note that this also means that the difference is only significant because you did a one-tailed test – because of the multiplication with 2, a two-tailed test would not have yielded a significant result but $p = 0.07302292$.

Now the same test with R. Since you already know the arguments of the function `t.test`, we can focus on the only major differences to before, the facts that you now have a directional alternative hypothesis and need to do a one-tailed test and that you now do a paired test. To do that properly, you must first understand how R computes the difference. As mentioned before above, R proceeds alphabetically and computes the difference 'alphabetically first level minus alphabetically second level' (which is why the alternative hypothesis was formulated this way above). Since "Original" comes before "Translation" and we saw that the mean of the former is smaller than that of the latter, the difference is smaller than 0. You therefore tell R that the difference is “less” than zero.

Of course you can use the formula or the vector-based notation. I show the output of the formula notation, where the setting of `alternative` pertains, as usual, to the first named vector. Both ways result in the same output. You get the *t*-value (which is negative here, because R subtracts the other way round), the *df*-value, a *p*-value, and a confidence interval which, since it does not include 0, also reflects the significant result.

```
> t.test(LENGTH~TEXTSOURCE,·paired=T,·alternative="less")¶
········Paired·t-test
data:··LENGTH·by·TEXTSOURCE
t·=·-1.9279,·df·=·15,·p-value·=·0.03651
alterna-
       tive·hypothesis:·true·difference·in·means·is·less·than
       ·0
95·percent·confidence·interval:
·······-Inf·-7.243041
sample·estimates:
mean·of·the·differences
················-79.875
> t.test(LENGTH[TEXTSOURCE=="Original"],·LENGTH[TEXTSOURCE==
       "Translation"],·paired=T,·alternative="less")¶
```

Finally, let us compute an effect size. The formula for Cohen’s *d* for this *t*-test is represented in (45):

$$\text{(45)} \qquad \text{Cohen's } d = t \cdot \sqrt{\frac{2 \cdot \left(1 - r_{group1, group2}\right)}{n_{pairs}}}$$

```
> d<-abs(t.test(LENGTH~TEXTSOURCE,·paired=T,·alternative=
       "less")$stat*sqrt((2*(1-cor(LENGTH[TEXTSOURCE==
       "Original"],·LENGTH[TEXTSOURCE=="Translation"])))/
       (length(LENGTH)/2)))¶
```

Again, you get only an intermediately high value of 0.405. To sum up: “On average, the originals are approximately 80 words shorter than their translations (the 95% confidence interval of this difference is -Inf, -7.24). According to a *t*-test for dependent samples, this difference is significant: $t = -1.93$; $df = 15$; $p_{\text{one-tailed}} = 0.0365$. However, the effect is relatively small: the difference of 80 words corresponds to only about 5% of the length of the texts; Cohen’s $d = 0.405$.”

Recommendation(s) for further study

- Crawley (2007: 298ff.), Baayen (2008: Section 4.3.1)
- an exact variant of the *t*-test for dependent samples, which does not make any distributional assumptions, can be programmed relatively easily in R and is available from me upon request

3.2.3. One dep. variable (ordinal) and one indep. variable (nominal) (indep. samples)

In this section, we discuss a non-parametric test for two independent samples of ordinal data, the *U*-test. Since I mentioned at the beginning of Section 4.3.2 that the *U*-test is not only used when the samples to be compared consist of ordinal data, but also when they violate distributional assumptions, this section will again involve an example where only a test of these distributional assumptions allows you to decide which test to use.

In Section 4.3.1.2 above, you looked at the similarities of source words entering into subtractive word formations and you tested whether these similarities were on average different from the known average similarity of random words to each other. The data you used were of the kind studied in Gries (2006) but in the above example no distinction was made between source words entering into different kinds of subtractive word formations. This is what we will do here by comparing similarities of source words entering into blends to similarities of complex clippings. If both kinds of word-formation processes differed according to this parameter, this would provide empirical motivation for distinguishing these processes in the first place. This example, thus, involves

- one dependent ratio-scaled variable, namely the SIMILARITY of the source words whose average you are interested in;
- one independent nominal variable, namely PROCESS: *BLEND* vs. PROCESS: *COMPLCLIP*;
- independent samples since the Dice coefficient of any one pair of source words has nothing to do with any one other pair of source words.

This kind of question would typically be investigated with the *t*-test for independent samples we discussed above. According to the above procedure, you first formulate the hypotheses (non-directionally since we may have no a priori reason to assume a particular difference):

H_0: The mean of the Dice coefficients of the source words of blends is as large as the mean of the Dice coefficients of the source words of complex clippings; $mean_{\text{Dice coefficients of blends}} = mean_{\text{Dice coefficients of complex clippings}}$, or $mean_{\text{Dice coefficients of blends}} - mean_{\text{Dice coefficients of complex clippings}} = 0$.

H_1: The mean of the Dice coefficients of the source words of blends is not as large as the mean of the Dice coefficients of the source words of complex clippings; $mean_{\text{Dice coefficients of blends}} \neq mean_{\text{Dice coefficients of complex clippings}}$, or $mean_{\text{Dice coefficients of blends}} - mean_{\text{Dice coefficients of complex clippings}} \neq 0$.

You can load the data from the file <C:/_sflwr/_inputfiles/04-3-2-3_dices.txt>. As before, this file contains the Dice coefficients, but now also in an additional column the word formation process for each Dice coefficient.

```
> Dices<-read.table(choose.files(),·header=T,·sep="\t",·
      comment.char="",·quote="")¶
> attach(Dices);·str(Dices)¶
'data.frame':···100·obs.·of··3·variables:
·$·CASE···:·int··1·2·3·4·5·6·7·8·9·10·...
·$·PROCESS:·Factor·w/·2·levels·"Blend","ComplClip":·2·2·2·2·2
      ·2·...
·$·DICE:·num··0.19·0.062·0.06·0.064·0.101·0.147·0.062·0.184·.
      ..
```

As usual, you should begin by exploring the data graphically:

```
> boxplot(DICE~PROCESS,·notch=T,·ylim=c(0,·1))¶
> rug(jitter(DICE[PROCESS=="Blend"]),·side=2)¶
> rug(jitter(DICE[PROCESS=="ComplClip"]),·side=4)¶
> text(1:2,·tapply(DICE,·PROCESS,·mean),·"+")¶
```

As usual, this graph already gives away enough information to nearly obviate the need for statistical analysis. The probably most obvious aspect is the difference between the two medians, but since the data are ratio-scaled you also need to explore the means. These are already plotted into the graph and here is the usual line of code to compute them directly; note how much the central tendency of the complex clippings differs from that of the blends.

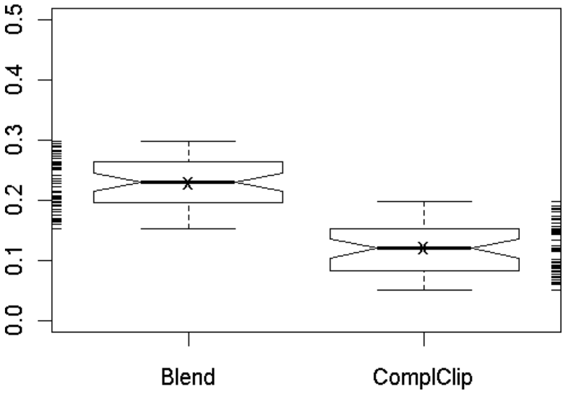

Figure 55. Boxplot for SIMILARITY~PROCESS

```
> tapply(DICE,·PROCESS,·mean)¶
····Blend·ComplClip
··0.22996···0.12152
```

In order to test whether the *t*-test for independent samples can be used here, we need to test both of its assumptions, normality in the groups and variance homogeneity. Since the *F*-test for homogeneity of variances presupposes normality, you begin by testing whether the data are normally distributed. As a first step, you generate histograms for both samples. The argument `main=""` suppresses an otherwise very wide headline and, more importantly, the arguments `xlim=c(0,·0.5)` and `ylim=c(0,·15)` force R to plot the histograms into identical coordinate systems so that we cannot be mislead by automatically chosen ranges of plots; cf. Figure 56. You can immediately see that the data are not normally distributed, which is supported by the Shapiro-Wilk test.

```
> par(mfrow=c(1,·2))¶
> hist(DICE[PROCESS=="Blend"],·main="",·xlab="Blends,·
     "ylab="Frequency",·xlim=c(0,·0.5),·ylim=c(0,15))¶
> hist(DICE[PROCESS=="ComplClip"],·main="",·xlab="Complex·
     clippings",·ylab="Frequency",·xlim=c(0,·0.5),·
     ylim=c(0,15))¶
> par(mfrow=c(1,·1))·#·restore·the·standard·plotting·setting¶
```

Given these violations of normality, you can actually not do the regular *F*-test to test the second assumption of the *t*-test for independent samples. You therefore do the Fligner-Killeen test of homogeneity of variances, which does not require the data to be normally distributed and which I mentioned in Section 4.2.2 above.

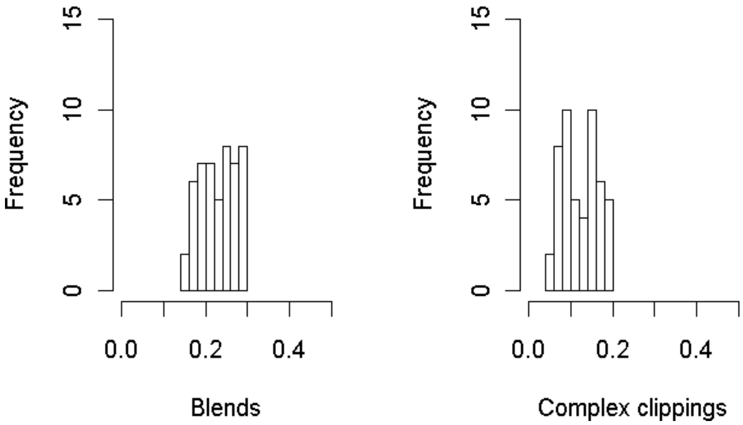

Figure 56. Histograms of Dice coefficients for both word-formation processes

```
> tapply(DICE,·PROCESS,·shapiro.test)¶
$Blend
········Shapiro-Wilk·normality·test
data:··X[[1L]]
W·=·0.9455,·p-value·=·0.02231
$ComplClip
········Shapiro-Wilk·normality·test
data:··X[[2L]]
W·=·0.943,·p-value·=·0.01771
```

```
> fligner.test(DICE~PROCESS)¶
········Fligner-Killeen·test·of·homogeneity·of·variances
data:··DICE·by·PROCESS
Fligner-Killeen:med·chi-squared·=·3e-04,·df·=·1,·p-
       value·=·0.9863
```

The variances are homogeneous, but normality is still violated. It follows that even though the data are ratio-scaled and even though the sample sizes are larger than 30, it is probably safer to compute a test that does not make these assumptions, the *U*-test.

Procedure
Formulating the hypotheses
Computing the observed medians, inspecting a graph
Testing the assumption(s) of the test:
    the values are independent of each other
    the populations from which the values were sampled are identically

distributed[30]
Computing the test statistics U and z as well as the probability of error p

While the two histograms do not seem to be from samples that are identically distributed, they are at least a bit similar, the variances of the two groups are not significantly different, and the *U*-test is relatively robust so we use it here. Since the *U*-test assumes only ordinal data, you now compute medians, not just means. You therefore adjust your hypotheses:

H_0: The median of the Dice coefficients of the source words of blends is as large as the median of the Dice coefficients of the source words of complex clippings; $median_{\text{Dice coefficients of blends}} = median_{\text{Dice coefficients of complex clippings}}$, or $median_{\text{Dice coefficients of blends}} - median_{\text{Dice coefficients of complex clippings}} = 0$.

H_1: The median of the Dice coefficients of the source words of blends is not as large as the median of the Dice coefficients of the source words of complex clippings; $median_{\text{Dice coefficients of blends}} \neq median_{\text{Dice coefficients of complex clippings}}$, or $median_{\text{Dice coefficients of blends}} - median_{\text{Dice coefficients of complex clippings}} \neq 0$.

Correspondingly, you compute the medians and interquartile ranges:

```
> tapply(DICE,·PROCESS,·median)¶
····Blend·ComplClip
····0.2300·····0.1195
> tapply(DICE,·PROCESS,·IQR)¶
····Blend·ComplClip
···0.0675····0.0675
```

Here, the assumptions can be tested fairly unproblematically: The values are independent of each other since no word-formation influences another one and the distributions of the data in Figures 55 and 56 appear to be rather similar.

Unfortunately, computing the *U*-test is somewhat more cumbersome than many other tests. First, you transform all Dice coefficients into ranks, and then you compute the sum of all ranks for each word-formation process. In R:

```
> Ts<-tapply(rank(DICE),·PROCESS,·sum)¶
```

[30] According to Bortz, Lienert, and Boehnke (1990:211), the *U*-test can discover differences of measures of central tendency well even if this assumption is violated.

Then, both of these *T*-values and the two sample sizes are inserted into the formulae in (46) and (47) to compute two *U*-values, the smaller one of which is the required test statistic:

(46) $$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - T_1$$

(47) $$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - T_2$$

```
> n1<-length(DICE[PROCESS=="Blend"])¶
> n2<-length(DICE[PROCESS=="ComplClip"])¶
> U1<-n1*n2+((n1*(n1+1))/2)-Ts[1]¶
> U2<-n1*n2+((n2*(n2+1))/2)-Ts[2]¶
> U<-min(U1,·U2)¶
```

The resulting *U*-value, 84, can be looked up in a *U*-table or, because there are few *U*-tables for larger samples,[31] converted into a normally-distributed *z*-score. This *z*-score is computed in several steps. First, you use the formulae in (48) and (49) to compute an expected *U*-value as well as its dispersion.

(48) $$U_{\text{expected}} = 0.5 \cdot n_1 \cdot n_2$$

(49) $$\textit{Dispersion } U_{\text{expected}} = \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$$

Second, you insert these values together with the observed *U*-value into the formula in (50).

(50) $$z = \frac{U - U_{expected}}{Dispersion\ U_{expected}}$$

```
> expU<-n1*n2/2¶
> dispersion.expU<-sqrt(n1*n2*(n1+n2+1)/12)¶
> z<-abs((U-expU)/dispersion.expU)¶
```

To decide whether the null hypothesis can be rejected, you look up this

[31] Bortz, Lienert and Boehnke (1990:202 and Table 6) provide critical *U*-values for $n \leq 20$ and mention references for tables with critical values for $n \leq 40$ – I at least know of no *U*-tables for larger samples.

value, 8.038194, in a *z*-table such as Table 37 or you compute a critical *z*-score for $p_{\text{two-tailed}}$ = 0.05 with `qnorm` (as was mentioned in Section 1.3.4.2 above). Since you have a non-directional alternative hypothesis, you apply the same logic as above and compute *z*-scores for half of the $p_{\text{two-tailed}}$-values you are interested in:

```
> qnorm(c(0.0005,·0.005,·0.025,·0.975,·0.995,·0.995),·
      lower.tail=F)¶
[1]··3.290527··2.575829··1.959964·-1.959964·-2.575829·-
      2.575829
```

Table 37. Critical *z*-scores for $p_{\text{two-tailed}}$ = 0.05, 0.01, and 0.001

z-score	*p*-value
1.96	0.05
2.575	0.01
3.291	0.001

It is obvious that the observed *z*-score is not only much larger than the one tabulated for $p_{\text{two-tailed}}$ = 0.001 but also very distinctly in the grey-shaded area in Figure 57: the difference between the medians is highly significant, as the non-overlapping notches already anticipated. Obviously, you can now also compute exact *p*-value with the usual ‘mirror function’ of `qnorm`:

```
> pnorm(z,·lower.tail=F)¶
[1]·4.558611e-16
```

In R, you compute the *U*-test with the same function as the Wilcoxon test, `wilcox.test`, and again you can either use a formula or two vectors. Apart from these arguments, the following ones are useful, too:

- `alternative`: a character string specifying which alternative hypothesis you want to test: the default is `"two.sided"`, other possible values for one-tailed tests are again `"less"` or `"greater"`, which specify how the first-named vector or factor level relates to the second;
- `paired=F` for the *U*-Test for independent samples or `paired=T` for the Wilcoxon test for dependent samples (cf. the following section);

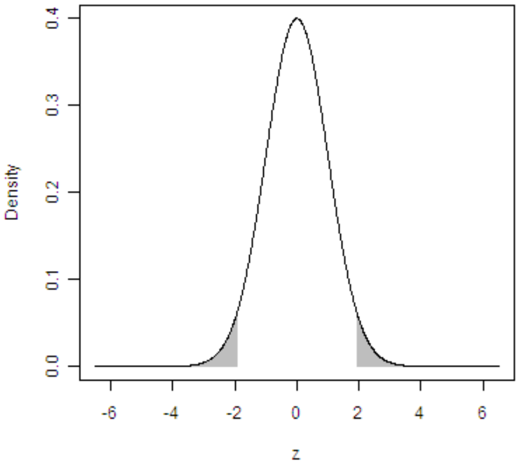

Figure 57. Density function of the standard normal distribution; two-tailed test

- `exact=T`, if you want to compute an exact test, or `exact=F` if you don't (if you don't change `exact`'s default setting of `NULL` and your data set has fewer than 50 data points and no ties, an exact *p*-value is computed automatically);
- `correct=T` for a continuity correction (the default) and `correct=F` for none;
- `conf.level`: a value between 0 and 1 specifying the size of the confidence interval; the default is 0.95.

The standard version to be used here is this:

```
> wilcox.test(DICE~PROCESS,·paired=F)¶
········Wilcoxon·rank·sum·test·with·continuity·correction
data:··DICE·by·PROCESS
W·=·2416,·p-value·=·8.882e-16
alternative·hypothesis:·true·location·shift·is·not·equal·to·0
```

You get a *U*-value (here referred to as *W*) and a *p*-value; *W* is not the minimum of U_1 and U_2, but the maximum here, which value you get depends on which vector or factor level comes first in the alphabet. The *p*-value here is a bit different from yours since R uses a slightly different algorithm and the continuity correction. I am not aware of a widely used effect size for median differences other than the observed difference itself, so you can now sum up: "According to a *U*-test, the median Dice coefficient of the source words of blends (0.23, *IQR* = 0.0675) and the median of the Dice coefficients for complex clippings (0.1195, *IQR* = 0.0675) are very significantly different: *U* = 84 (or *W* = 2416), $p_{\text{two-tailed}} < 0.0001$. The creators of blends appear to be more concerned with selecting source words that are similar to each other than the creators of complex clippings."

Recommendation(s) for further study:
Dalgaard (2002: 89f.), Crawley (2007: 297f.), Baayen (2008: Section 4.3.1) and recall the above comments regarding `wilcox.exact`

3.2.4. One dep. variable (ordinal) and one indep. variable (nominal) (dep. samples)

Just like the *U*-test, the test in this section has two major applications. First, you really may have two dependent samples of ordinal data such as when you have a group of subjects perform two rating tasks to test whether each subject's first rating differs from the second. Second, the probably more frequent application arises when you have two dependent samples of ratio-scaled data but cannot do the *t*-test for dependent samples because its distributional assumptions are not met. We will discuss an example of the latter kind in this section.

In a replication of Bencini and Goldberg, Gries and Wulff (2005) studied the question which verbs or sentence structures are more relevant for how German foreign language learners of English categorize sentences. They crossed four syntactic constructions and four verbs to get 16 sentences, each verb in each construction. Each sentence was printed onto a card and 20 advanced German learners of English were given the cards and asked to sort them into four pile of four cards each. The question was whether the subjects' sortings would be based on the verbs or the constructions. To determine the sorting preferences, each subject's four stacks were inspected with regard to how many cards one would minimally have to move to create either four completely verb-based or four completely construction-based sortings. The investigation of this question involves

- one dependent ratio-scaled variable, namely SHIFTS, the number of times a card had to be shifted from one stack to another to create the perfectly clean sortings, and we are interested in the average of these numbers;
- one independent nominal variable, namely CRITERION: *CONSTRUCTION* vs. CRITERION: *VERB*;

– dependent samples since each subject 'generated' two numbers of changes, one to create the verb-based sorting, one to create the construction-based sorting.

To test some such result for significance, you should first consider a *t*-test for dependent samples since you have two samples of ratio-scaled values. As usual, you begin by formulating the relevant hypotheses:

H_0: The average of the pairwise differences between the numbers of rearrangements towards perfectly verb-based stacks and the numbers of rearrangements towards perfectly construction-based stacks is 0; $mean_{\text{pairwise differerences}} = 0$.
H_1: The average of the pairwise differences between the numbers of rearrangements towards perfectly verb-based stacks and the numbers of rearrangements towards perfectly construction-based stacks is not 0; $mean_{\text{pairwise differerences}} \neq 0$.

Then, you load the data that Gries and Wulff (2005) obtained in their experiment from the file <C:/_sflwr/_inputfiles/04-3-2-4_sortingstyles.txt>:

```
> SortingStyles<-read.table(),·header=T,·sep="\t",·
     comment.char="",·quote="")¶
> attach(SortingStyles)¶
> head(SortingStyles,·3)¶
··CASE·SHIFTS····CRITERION
1····1······0·Construction
2····2······0·Construction
3····3······4·Construction
```

As usual, you compute the means and generate a graph of the results.

```
> tapply(SHIFTS,·CRITERION,·mean)¶
Construction·········Verb
·········3.45·········8.85
> boxplot(SHIFTS~CRITERION,·notch=T)¶
> rug(jitter(SHIFTS[CRITERION=="Construction"]),·side=2)¶
> rug(jitter(SHIFTS[CRITERION=="Verb"]),·side=4)¶
```

(Note that the boxplot does not represent the 'pairwise-ness' of the differences.) Both medians and notches indicate that the average numbers of card rearrangements are very different. You then test the assumption of the *t*-test for dependent samples, the normality of the pairwise differences:

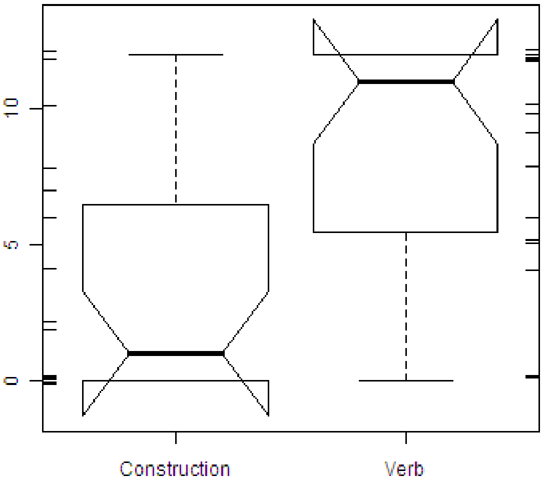

Figure 58. Boxplot for SHIFTS~CRITERION

```
> differences<-SHIFTS[CRITERION=="Construction"]-
       SHIFTS[CRITERION!="Construction"]¶
> shapiro.test(differences)¶
········Shapiro-Wilk·normality·test
data:··differences
W·=·0.7825,·p-value·=·0.0004797
```

The distribution of the differences deviates highly significantly from normality: you cannot use the *t*-test. Instead, you compute a test for two dependent samples of ordinal variables, the Wilcoxon test.

As a first step, you adjust your hypotheses:

H_0: $median_{\text{pairwise differerences}} = 0$
H_1: $median_{\text{pairwise differerences}} \neq 0$

Procedure
Formulating the hypotheses
Computing the observed medians, inspecting a graph
Testing the assumption(s) of the test:
the pairs of values are independent of each other
the populations from which the samples were obtained are distributed identically
Computing the test statistic *T* and the probability of error *p*

Since the data are now analyzed on an ordinal level of measurement, you compute the medians and their interquartile ranges:

```
> tapply(SHIFTS,·CRITERION,·median)¶
Construction·········Verb
···········1···········11
> tapply(SHIFTS,·CRITERION,·IQR)¶
Construction·········Verb
········6.25·········6.25
```

These are the medians that you could already infer from the above boxplot. The assumptions appear to be met because the pairs of values are independent of each other (since the sorting of any one subject does not affect any other subject's sorting) and, somewhat informally, there is little reason to assume that the populations are distributed differently especially since most of the values to achieve a perfect verb-based sorting are the exact reverse of the values to get a perfect construction-based sorting. Thus, you compute the Wilcoxon test; for reasons of space we only consider the standard variant. First, you transform the vector of pairwise differences, which you already computed for the Shapiro-Wilk test, into ranks:

```
> ranks<-rank(abs(differences))¶
```

Second, all ranks whose difference was negative are summed to a value *T*-, and all ranks whose difference was positive are summed to *T*+; the smaller of the two values is the required test statistic *T*:[32]

```
> T.minus<-sum(ranks[differences<0])¶
> T.plus<-sum(ranks[differences>0])¶
> T<-min(T.minus,·T.plus)¶
```

This *T*-value of 41.5 can be looked up in a *T*-table, but note that here, for a significant result, the observed test statistic must be *smaller* than the tabulated one. The observed *T*-value of 41.5 is smaller than the one tabulated for n = 20 and p = 0.05 (but larger than the one tabulated for n= 20 and p = 0.01): the result is significant.

[32] The way of computation discussed here is the one described in Bortz (2005). It disregards ties and cases where the differences are zero.

Table 38. Critical T-values for $p_{\text{two-tailed}} = 0.05$, 0.01, and 0.001 for $14 \leq df \leq 16$

	$p = 0.05$	$p = 0.01$	$p = 0.001$
$n = 19$	46	32	18
$n = 20$	52	37	21
$n = 21$	58	42	25

Let us now do this test with R: You already know the function for the Wilcoxon test so we need not discuss it again in detail. The relevant difference is that you now instruct R to treat the samples as dependent/paired. As nearly always, you can use the vector-based function call or the formula.

```
> wilcox.test(SHIFTS[CRITERION=="Verb"],·SHIFTS[CRITERION==
      "Construction"],·paired=T,·exact=F)¶
> wilcox.test(SHIFTS~CRITERION,·paired=T,·exact=F)¶
········Wilcoxon·signed·rank·test·with·continuity·correction
data:··SHIFTS·by·CRITERION
V·=·36.5,·p-value·=·0.01616
alternative·hypothesis:·true·location·shift·is·not·equal·to·0
```

R computes the test statistic differently but arrives at the same kind of decision: the result is significant, but not very significant.

To sum up: “On the whole, the 20 subjects exhibited a strong preference for a construction-based sorting style: the median number of card rearrangements to arrive at a perfectly construction-based sorting was 1 while the median number of card rearrangements to arrive at a perfectly verb-based sorting was 11 (both *IQR*s = 6.25). According to a Wilcoxon test, this difference is significant: $V = 36.5$, $p_{\text{two-tailed}} = 0.0162$. In this experiment, the syntactic patterns were a more salient characteristic than the verbs (when it comes to what triggered the sorting preferences).”

Recommendation(s) for further study:
recall `wilcox.exact`

4. Coefficients of correlation and linear regression

In this section, we discuss the significance tests for the coefficients of correlation discussed in Section 3.2.3.

4.1. The significance of the product-moment correlation

While the manual computation of the product-moment correlation above was a bit complex, its significance test is not. It involves these steps:

Procedure

Formulating the hypotheses

Computing the observed correlation; inspecting a graph

Testing the assumption(s) of the test: the population from which the sample was drawn is bivariately normally distributed. Since this criterion *can* be hard to test (cf. Bortz 2005: 213f.), we simply require both samples to be distributed normally

Computing the test statistic *t*, the degrees of freedom *df*, and the probability of error *p*

Let us return to the example in Section 3.2.3, where you computed a correlation coefficient of 0.9337 for the correlation of the lengths of 20 words and their reaction times. You formulate the hypotheses and we assume for now your alternative hypothesis is non-directional.

H_0: The length of a word in letters does not correlate with the word's reaction time in a lexical decision task; $r = 0$.

H_1: The length of a word in letters correlates with the word's reaction time in a lexical decision task; $r \neq 0$.

You load the already familiar data from <C:/_sflwr/_inputfiles/03-2-3_reactiontimes.txt>:

```
> ReactTime<-read.table(choose.files(),·header=T,·sep="\t")¶
> attach(ReactTime);·str(ReactTime)¶
'data.frame':···20·obs.·of··3·variables:
·$·CASE······:·int··1·2·3·4·5·6·7·8·9·10·...
·$·LENGTH····:·int··14·12·11·12·5·9·8·11·9·11·...
·$·MS_LEARNER:·int··233·213·221·206·123·176·195·207·172·...
```

Since we already generated a scatterplot above (cf. Figure 36 and Figure 37), we do not do that again. We do, however, have to test the assumption of normality of both vectors. You can either proceed in a stepwise fashion and enter `shapiro.test(LENGTH)¶` and `shapiro.test(MS_LEARNER)¶` or use a shorter variant:

```
> apply(ReactTime[,2:3],·2,·shapiro.test)¶
$LENGTH
········Shapiro-Wilk·normality·test
data:··newX[,·i]
W·=·0.9748,·p-value·=·0.8502
$MS_LEARNER
········Shapiro-Wilk·normality·test
data:··newX[,·i]
W·=·0.9577,·p-value·=·0.4991
```

This line of code means 'take the data mentioned in the first argument of `apply` (the second and third column of the data frame `ReactTime`), look at them column by column (the 2 in the second argument slot – a 1 would mean look at them row-wise; recall this notation from `prop.table` in Section 3.2.1), and apply the function `shapiro.test` to each column. Clearly, both variables do not differ significantly from a normal distribution.

To compute the test statistic *t*, you insert the correlation coefficient *r* and the number of correlated value pairs *n* into the formula in (51):

(51) $$t = \left| \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \right|$$

```
> r<-cor(LENGTH,·MS_LEARNER,·method="pearson")¶
> numerator<-r*sqrt(length(LENGTH)-2)¶
> denominator<-sqrt(1-r^2)¶
> t<-numerator/denominator¶
```

This *t*-value, 11.06507, has $df = n$-2 = 18 degrees of freedom.

```
> df<-length(LENGTH)-2¶
```

Just as with the *t*-tests before, you can now look this *t*-value up in a *t*-table, or you can compute a critical value: if the observed *t*-value is higher than the tabulated/critical one, then *r* is significantly different from 0. Since your *t*-value is much larger than even the one for $p = 0.001$, the correlation is highly significant.

```
> qt(c(0.025,·0.975),·18,·lower.tail=F)·#·division·by·2!¶
[1]··2.100922·-2.100922
```

The exact *p*-value can be computed as follows, and do not forget to again double the *p*-value.

```
> 2*pt(t,·18,·lower.tail=F)·#·multiplication·with·2!¶
[1]·1.841060e-09
```

Table 39. Critical *t*-values for $p_{\text{two-tailed}}$ = 0.05, 0.01, and 0.001 for $17 \leq df \leq 19$

	p = 0.05	*p* = 0.01	*p* = 0.001
df = 17	2.1098	2.8982	3.9561
df = 18	2.1009	2.8784	3.9216
df = 19	2.093	2.8609	3.8834

This *p*-value is obviously much smaller than 0.001. However, you will already suspect that there is an easier way to get all this done. Instead of the function `cor`, which we used in Section 3.2.3 above, you simply use `cor.test` with the two vectors whose correlation you are interested (and, if you have a directional alternative hypothesis, you specify whether you expect the correlation to be less than 0 (i.e., negative) or greater than 0 (i.e., positive) using `alternative=…`:

```
> cor.test(LENGTH,·MS_LEARNER,·method="pearson")¶
········Pearson's·product-moment·correlation
data:··LENGTH·and·MS_LEARNER
t·=·11.0651,·df·=·18,·p-value·=·1.841e-09
alternative·hypothesis:·true·correlation·is·not·equal·to·0
95·percent·confidence·interval:
·0.8370608·0.9738525
sample·estimates:
······cor
0.9337171
```

You can also look at the results of the corresponding linear regression:

```
> model<-lm(MS_LEARNER~LENGTH)¶
> summary(model)¶
Call:
lm(formula·=·MS_LEARNER·~·LENGTH)

Residuals:
·····Min·······1Q···Median·······3Q······Max
-22.1368··-7.8109···0.8413···7.9499··18.9501
Coefficients:
············Estimate·Std.·Error·t·value·Pr(>|t|)
(Intercept)··93.6149·····9.9169····9.44·2.15e-08·***
LENGTH·······10.3044·····0.9313···11.06·1.84e-09·***
---
Signif.·codes:··0·'***'·0.001·'**'·0.01·'*'·0.05·
       '.'·0.1·'·'·1
Residual·standard·error:·11.26·on·18·degrees·of·freedom
```

```
Multiple·R-Squared:·0.8718,·····Adjusted·R-squared:·0.8647
F-statistic:·122.4·on·1·and·18·DF,··p-value:·1.841e-09
```

We begin at the bottom: the last row contains information we already know. The *F*-value is our *t*-value squared; we find the 18 degrees of freedom and the *p*-value we computed. In the line above that you find the coefficient of determination you know plus an adjusted version we will only talk about later (cf. Section 5.2). We ignore the line about the residual standard error (for now) and the legend for the *p*-values. The table above that shows the intercept and the slope we computed in Section 3.2.3 (in the column labeled “Estimate”), their standard errors, *t*-values – do you recognize the *t*-value from above? – and *p*-values. The *p*-value for LENGTH says whether the slope of the regression line is significantly different from 0; the *p*-value for the intercept says whether the intercept of 93.6149 is significantly different from 0. We skip the info on the residuals because we discussed above how you can investigate those yourself (with `residuals(model)¶`).

To sum up: “The lengths of the words in letters and the reaction times in the experiment correlate highly positively with each other: $r = 0.9337$; $r^2 = 0.8718$. This correlation is highly significant: $t = 11.07$; $df = 18$; $p < 0.001$. A linear regression shows that every additional letter increases the reaction time by approximately 10.3 ms.”

In Section 5.2, we deal with the extensions of linear regression to cases where we include more than one independent variable, and we will also discuss more comprehensive tests of the regression’s assumptions (using `plot(model)¶`).

4.2. The significance of Kendall’s Tau

If you need a *p*-value for Kendall’s tau *τ*, you follow the following procedure:

Procedure
Formulating the hypotheses
Computing the observed correlation; inspecting a graph
Testing the assumption(s) of the test: the data are at least ordinal
Computing the test statistic *z* and the probability of error *p*

Again, we simply use the example from Section 3.2.3 above (even though we know we can actually use the product-moment correlation; we use this example again just for simplicity's sake). How to formulate the hypotheses should be obvious by now:

H_0: The length of a word in letters does not correlate with the word's reaction time in a lexical decision task; $\tau = 0$.

H_1: The length of a word in letters correlates with the word's reaction time in a lexical decision task; $\tau \neq 0$.

As for the assumption: we already know the data are ordinal – after all, we know they are even interval/ratio-scaled. You load the data again from <C:/_sflwr/_inputfiles/03-2-3_reactiontimes.txt> and compute Kendall's τ:

```
> ReactTime<-read.table(choose.files(),·header=T,·sep="\t")¶
> attach(ReactTime)¶
> tau<-cor(LENGTH,·MS_LEARNER,·method="kendall")·#·0.8189904¶
```

To test Kendall's tau τ for significance, you compute a *z*-score of the kind that is by now familiar. You insert τ and the number of value pairs n into the formula in (52).

(52) $$z = \left|\tau\right| \div \sqrt{\frac{2 \cdot (2 \cdot n + 5)}{9 \cdot n \cdot (n - 1)}}$$

In R:

```
> numerator.root<-2*(2*length(LENGTH)+5)¶
> denominator.root<-9*length(LENGTH)*(length(LENGTH)-1)¶
> z<-abs(tau)/sqrt(numerator.root/denominator.root);·z¶
[1]·5.048596
```

This value can then be looked up in a *z*-table such as Table 40 or you generate these values. The *z*-score for a significant two-tailed test must cut off at least 2.5% of the area under the standard normal distribution:

```
> qnorm(c(0.0005,·0.005,·0.025,·0.975,·0.995,·0.9995),·
      lower.tail=T)¶
[1]·-3.290527·-2.575829·-1.959964··1.959964··2.575829··
      3.290527
```

Table 40. Critical *z*-scores for $p_{\text{two-tailed}}$ = 0.05, 0.01, and 0.001

z-score	*p*
1.96	0.05
2.576	0.01
3.291	0.001

For a result to be significant, the *z*-score must be larger than 1.96. Since the observed *z*-score is actually larger than 5, this result is highly significant:

```
> 2*pnorm(z,·lower.tail=F)¶
[1]·4.450685e-07
```

The function to get this result much faster is again `cor.test`. Since R uses a slightly different method of calculation, you get a slightly different *z*-score and *p*-value, but the results are for all intents and purposes identical.

```
> cor.test(LENGTH,·MS_LEARNER,·method="kendall")¶
········Kendall's·rank·correlation·tau
data:··LENGTH·and·MS_LEARNER
z·=·4.8836,·p-value·=·1.042e-06
alternative·hypothesis:·true·tau·is·not·equal·to·0
sample·estimates:
······tau
0.8189904
Warning·message:
In·cor.test.default(LENGTH,·MS_LEARNER,·method·=·"kendall")·:
··Cannot·compute·exact·p-value·with·ties
```

(The warning refers to ties such as the fact that the length value 11 occurs more than once). To sum up: “The lengths of the words in letters and the reaction times in the experiment correlate highly positively with each other: $\tau = 0.819$, $z = 5.05$; $p < 0.001$.”

4.3 Correlation and causality

Especially in the area of correlations, but also more generally, you need to bear in mind a few things even if the null hypothesis is rejected: First, one can often hear a person A making a statement about a correlation (maybe even a significant one) by saying “The more X, the more Y” and then hear a person B objecting on the grounds that B knows of an exception. This argument is flawed. The exception quoted by B would only invalidate A’s statement if A considered the correlation to be perfect (*r* = 1 or *r* = -1) – but if A did not mean that (and A never does!), then there may be strong and significant correlation although there is (at least) one exception. The exception or exceptions are the reason why the correlation is not 1 or -1 but 'only', say, 0.9337. Second, a correlation as such does not necessarily imply causality. As is sometimes said, a correlation between X and Y is a *necessary* condition for a causal relation between X and Y, but not a *sufficient* one, as you can see from many examples:

- There is a positive correlation between the number of firefighters trying to extinguish a fire and the amount of damage that is caused at the site where the fire was fought. This does of course not mean that the firefighters arrive at the site and destroy as much as they can – the correlation results from a third variable, the size of the fire: the larger the fire, the more firefighters are called to help extinguish it *and* the more damage the fire causes.
- There is a negative correlation between the amount of hair men have and their income which is unfortunately only due to the effect of a third variable: the men's age.
- There is a positive correlation such that the more likely a drug addict was to go to therapy to get off his addiction, the more likely he was to die. This is not because the therapy leads to death – the variable in the background correlated with both is the severity of the addiction: the more addicted addicts were, the more likely they were to go to therapy, but also the more likely they were to die.

Thus, beware of jumping to conclusions …

Now you should do the exercise(s) for Chapter 4 …

Recommendation(s) for further study

- the functions `ckappa` and `lkappa` (from the `library(psy)`) to compute the kappa coefficient and test how well two or more raters conform in their judgments of stimuli
- the function `cronbach` (from the `library(psy)`) to compute Cronbach's alpha and test how consistently several variables measure a construct the variables are supposed to reflect
- Crawley (2007: Ch. 10), Baayen (2008: Section 4.3.2), Johnson (2008: Section 2.4)