15 Basic significance testing

Stefan Th. Gries

1 Introduction

This chapter introduces the fundamentals of inferential statistics – that is, methods that help you make inferences or predictions based on your sample data. More specifically, in most empirical studies, researchers cannot study the complete *population* of a phenomenon of interest – that is, the complete set of objects or speakers of interest – but only a small *sample* of the phenomenon under investigation. For example, instead of investigating all relative clauses, you investigate a (hopefully carefully constructed) sample of relative clauses in a (part of a) corpus; instead of testing all non-native speakers of a language, you test a (hopefully randomly selected) sample of speakers, and so on. Obviously, you hope that whatever results – percentages, means, correlation coefficients – you obtain from a sample (which you studied) will generalize to the population (which you did not study). However, if researchers draw different samples from the same population and compute point estimates of percentages, means, correlation coefficients, they will just as obviously also get different point estimates; they will encounter variability. The most important application of inferential statistics is to assist researchers in quantifying and studying this variability to (i) arrive at better estimates of population parameters, and (ii) test hypotheses and separate random/accidental from systematic/meaningful variation.

Section 2 will introduce several basic concepts that underlie most inferential statistics. Section 3 presents a set of questions based on Chapter 14 and Section 2 of this chapter that are necessary to identify which statistical test is applicable in a particular research scenario. Sections 4.1 and 4.2 then discuss a small selection of statistical tests involving frequency data of discrete/categorical data and central tendencies (means and medians) respectively.

2 The logic of significance tests

To put the notion of statistical testing into perspective, an introduction to the framework of *null hypothesis significance testing* (NHST) is required. As the term NHST suggests, the notion of *hypothesis* plays a central role in this framework. A hypothesis is a statement that makes a prediction about the distribution of one variable (or about the relation between two or more variables) in a

population and that has the implicit structure of a conditional sentence (*if . . ., then . . .* or *the more/less . . ., the more/less . . .*). Two different ways of characterizing hypotheses must be distinguished:

- *alternative hypotheses (H_1)* vs *null hypotheses (H_0)*: the former is a statement about an effect, a difference, a correlation regarding one or more variables; the latter is the logical counterpart of the former (i.e., a statement that predicts the absence of an effect, a difference, a correlation). Most of the time, the research hypothesis that is explored in an empirical study is an alternative hypothesis, predicting, say, a difference between percentages, a difference between group averages, a correlation between two or more variables, and so on.
- *text hypotheses* vs *statistical hypotheses*: each of the two above hypotheses comes in two forms. The former is a prediction in natural, "normal" language, such as in the H_1, in English ditransitives, recipients are shorter than patients. The latter is the former's translation into something that can be counted or measured – that is, its *operationalization*. This is an important step, not only because a proper operationalization is required to ensure the study's validity, but also because one text hypothesis can be translated into different statistical hypotheses. For instance, one statistical hypothesis for the above text hypothesis involves central tendencies such as means: in English ditransitives, the mean syllabic length of recipients is smaller than the mean syllabic length of patients. However, an operationalization based on counts/frequencies would also be possible: in English ditransitives, the number of recipients that are shorter than the average of all recipients and patients is larger than the number of patients shorter than that average. Needless to say, it is possible that the first statistical hypothesis is supported whereas the second is not, which is why a careful operationalization is essential and, obviously, will determine which statistical test you need to perform.

Significance tests are based on the following logic and steps: (i) you compute the effect you observe in your data (e.g., a frequency distribution, a difference in means, a correlation), (ii) you compute the so-called *probability of error p* to obtain the (summed/combined) probability of the observed effect and every other result that deviates from H_0 even more when H_0 is true, and (iii) you compare *p* to a significance level (usually 5 percent, i.e., 0.05) and, if *p* is smaller than the significance level, you reject H_0 (because it is not compatible enough with the data to stick to it) and accept H_1. Note that this does not mean you have *proven* H_1: after all, there is still a probability *p* that your observed effect/result arises under H_0 – *p* is just too small (by convention) to stick to H_0 (see Cohen 1994 for a critical discussion of NHST).

The above immediately leads to the question of how that probability *p* is computed. One way is to write up all results possible and their probabilities

Table 15.1 *All possible results from asking three subjects to classify* walk *as a noun or a verb*

Subject 1	Subject 2	Subject 3	# noun	# verb	Probability
noun	noun	noun	3	0	0.125
noun	noun	verb	2	1	0.125
noun	verb	noun	2	1	0.125
noun	verb	verb	1	2	0.125
verb	noun	noun	2	1	0.125
verb	noun	verb	1	2	0.125
verb	verb	noun	1	2	0.125
verb	verb	verb	0	3	0.125

under H_0 and then check how likely the observed result and everything more extremely deviating from H_0 is. Imagine a linguist interested in conversion/zero-derivation in English. He presents the word *walk* independently to three subjects and asks them which word class it is: *noun* or *verb*. Imagine further that all three subjects responded *verb*. How likely is this result, assuming that, under H_0, *noun* and *verb* are equally likely? To answer this question, Table 15.1 summarizes the whole result space: the three left columns represent the subjects and their possible answers, columns four and five summarize the numbers of noun and verb responses for each possible outcome, and the rightmost column provides the probability for each of the eight results, which are equally likely under the H_0 and, thus, all $^1/_8$.

The linguist can now determine how probable the observed result – three times verb – and all other results deviating from H_0 even more – none, three times verb is the most extreme verb-favoring result you can get from three subjects – are. That probability is shown in the last row: p=0.125, which makes the observed result not significantly different from chance.

Obviously, the strategy of writing up every possible result, and so on, is not feasible with continuous data, or even with the binary data from above if the sample size becomes large. However, consider Figure 15.1 to see what happens as the number of trials increases. The top left panel shows the probability distribution for the data in Table 15.1: p(0 times verb)=0.125, p(1 times verb)=0.375, p(2 times verb)=0.375, and p(3 times verb)=0.125. If you perform six or twelve trials, you obtain the other distributions in the upper panel, and if you perform twenty-five, fifty, or one hundred trials, you obtain the distributions in the lower panel: clearly, as the number of trials increases, the discrete probability distribution looks more and more like a bell-shaped curve, whose distribution can therefore be modeled on the basis of the equation underlying a Gaussian normal distribution, as shown in (1).

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2} \quad (1)$$

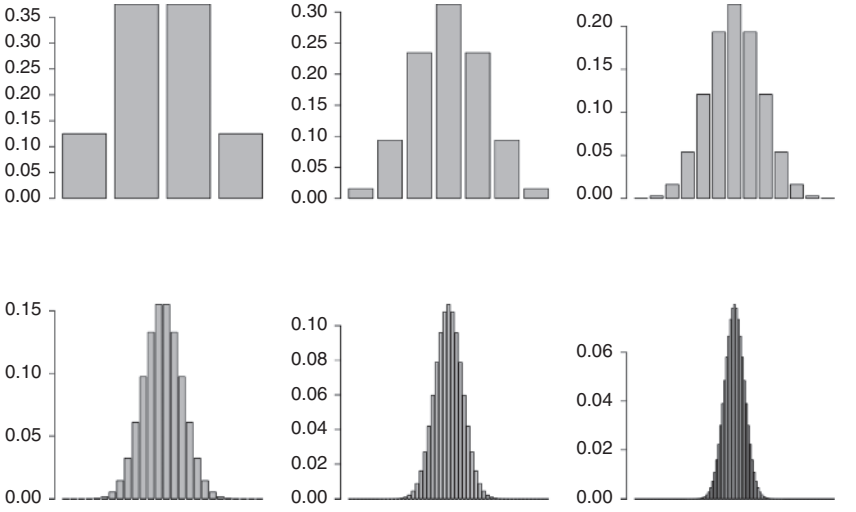

Figure 15.1. *Probability distributions for outcomes of equally likely binary trials (top row: 3, 6, and 12 trials; bottom row: 25, 50, and 100 trials)*

Thus, if the data under investigation are distributed in a way that is sufficiently similar to the normal distribution (or another one of several widely used probability density functions, such as the *F*-, *t*-, or χ^2-distribution), then one does not have to compute, and sum over, exact probabilities as in Table 15.1, but can approximate the *p*-value from parameters of equations underlying the above distributions (such as (1)); this is often called using *parametric tests*. Crucially, this approximation of a *p*-value on the basis of a function can be only as good as the data's distribution is similar to the corresponding function; the next section illustrates the relevance of this issue, as well as a few others, for selecting the right statistical test.

A second advantage of your data being distributed similarly to a known distribution is that this sometimes allows you to compute a so-called *confidence interval* on top of a descriptive statistic (such as a mean or a correlation). A 95 percent confidence interval helps you assess the precision of a statistic describing your sample; it

> identifies a range of values a researcher can be 95% confident contains the true value of a population parameter . . . Stated in probabilistic terms, the researcher can state that there is a probability/likelihood of 0.95 that the confidence interval contains the true value of the population parameter. (Sheskin 2007: 74)

Section 4 will provide two examples for confidence intervals.

3 Choosing significance tests

The decision for a particular statistical test is typically made on the basis of a set of questions that cover various aspects of the study you are

conducting, the number and types of variables that are involved, and the size and distribution of the dataset(s) involved. The remainder of this section discusses these questions in (1) to (6), and their possible answers and implications.

1. What kind of study is being conducted?

This question is usually easy to answer. At the risk of a slight simplification, studies are either *exploratory/hypothesis-generating* or *hypothesis-testing*. The former means that you are approaching a (typically large) dataset with the intentions of detecting structure(s) and developing hypotheses for future studies; your approach to the data is therefore data-driven, or bottom-up. The latter means you are approaching the dataset with a specific hypothesis in mind which you want to test. In this chapter, I will discuss only the latter type of study (see Chapter 14 for a discussion of the former type).

2. How many and what kinds of variables are involved?

There are essentially four different possible answers. First, you may only have one dependent variable. In that case, you normally want to compute a *goodness-of-fit test* to test whether the results from your data correspond to other results (from a previous study) or correspond to a known distribution (such as a normal distribution). Examples include the following:

- Does the ratio of *no*-negations (e.g., *He is no stranger*) and *not*-negations (e.g., *He is not a stranger*) in your data correspond to a uniform distribution?
- Does the average acceptability judgment you receive for a sentence correspond to that of a previous study?

Second, you may have one dependent and one independent variable, in which case you want to compute a *monofactorial test for independence* to determine whether the values of the independent variable are correlated with those of the dependent variable. For example:

- Does the animacy of the referent of the direct object (a categorical independent variable) correlate with the choice of one of two post-verbal constituent orders (a categorical dependent variable)?
- Does the average acceptability judgment (a mean of a ratio/interval/ordinal dependent variable) vary as a function of whether the subjects providing the ratings are native or non-native speakers (a categorical independent variable)?

Third, you may have one dependent and two or more independent variables, in which case you want to compute a *multifactorial* analysis to determine whether the individual independent variables *and* their interactions correlate with the dependent variable. For example:

- Does the frequency of a negation type (a categorical dependent variable with the levels *no* vs *not*; see above) depend on the mode (a binary independent variable with the levels *spoken* vs *written*), the type of verb that is negated (a categorical independent variable with the levels *copula*, *have*, or *lexical*), and/or the interaction of these independent variables?
- Does the reaction time to a word w in a lexical decision task (a ratio/interval dependent variable) depend on the word class of w (a categorical independent variable), the frequency of w in a reference corpus (a ratio/interval independent variable), whether the subject has seen a word semantically related to w on the previous trial or not (a binary independent variable), whether the subject has seen a word phonologically similar to w on the previous trial or not (a binary independent variable), and/or the interactions of these independent variables?

Such multifactorial tests are discussed in Chapters 16 and 20.

Fourth, you have two or more dependent variables, in which case you want to perform a *multivariate* analysis, which can be exploratory (such as hierarchical cluster analysis, principal components analysis, factor analysis, multidimensional scaling) or hypothesis-testing in nature (MANOVA).

3. Are data points in your data related such that you can associate data points to each other meaningfully and in a principled way?

This question is concerned with whether you have what are called independent or dependent samples. For example, your two samples (e.g., the numbers of mistakes made by ten male and ten female non-native speakers in a grammar test) are *independent* of each other if you cannot connect each male subject's value to that of one female subject on a meaningful and principled basis. This would be the case if you randomly sampled ten men and ten women and let them take the same test.

There are two ways in which samples can be *dependent*. One is if you test subjects more than once (e.g., before and after a treatment). In that case, you could meaningfully connect each value in the before-treatment sample to a value in the after-treatment sample, namely connect each subject's two values. The samples are dependent because, for instance, if subject #1 is very intelligent and good at the language tested, then these characteristics will make his results better than average in both tests, especially compared to a subject who is less intelligent and proficient in the language and who will perform worse in both tests. Recognizing that the samples are dependent this way will make the test of before-vs-after treatments more precise.

The second way in which samples can be dependent can be explained using the above example of ten men and ten women. If the ten men were the husbands of the ten women, then one would want to consider the samples dependent. Why? Because spouses are on average more similar to each other than randomly chosen people: they often have similar IQs, similar professions, they spend more time with each

other than with randomly selected people, and so on. Thus, it would be useful to associate each husband with his wife, making this two dependent samples.

Independence of data points is often a very important criterion: many tests assume that data points are independent, and for many tests you must choose your test depending on what kind of samples you have. For instance, below I will discuss a *t*-test for independent samples and one for dependent samples.

4. What is the statistic of the dependent variable in the statistical hypotheses?

There are essentially five different answers to this question. Your dependent variable may involve *frequencies/counts* (e.g., when you study which level(s) of a categorical variable are attested more/less often than others), *central tendencies* (e.g., when you explore whether the mean or median of a ratio/interval or ordinal variable is as high as you expected), *dispersions* (e.g., when you investigate whether the variability of a ratio/interval or ordinal variable around its mean or median is higher in one group than another), *correlations* (e.g., when you ask whether changing the values of one variable bring about changes in another), and *distributions* (e.g., whether samples of two ratio/interval variables are both normally distributed or not). Obviously, the nature of your dependent variable has important consequences for your statistical analysis; below, we will discuss examples involving frequencies and central tendencies.

5. What does the distribution of the data look like? Normally or another way that can be described by a probability density function (or a way that can be transformed to correspond to a probability density function; see Section 5), or some other way?
6. How big are the samples to be collected? $n<30$ or $n\geq30$?

These final two questions are related to each other and to the above notion of parametric (vs non-parametric/distribution-free) tests. Parametric tests involve statistical approximations and rely on the sampled data being distributed in a particular way (for example, normally as represented in Figure 15.1 or the left panel of Figure 15.2). Sometimes, the data do not even have to be distributed normally as long as the sample size is large enough. However, the more the data violate distributional assumptions of the test you are considering (e.g., word lengths are often distributed as in the right panel of Figure 15.2), the safer it is to use a *non-parametric/distribution-free* alternative that does not rely on assumptions you know your data violate; see Section 4.2.2 for an example. See Chapter 3 for an example of a case where the distinction between parametric and non-parametric tests is important for analyzing grammaticality judgments.

Sometimes, tests have yet other requirements, such as particular minimal sample sizes or more complicated ones. In all cases, you must check your data for all of these to make an informed decision in favor of some test; ideally this involves a visual exploration of the data.

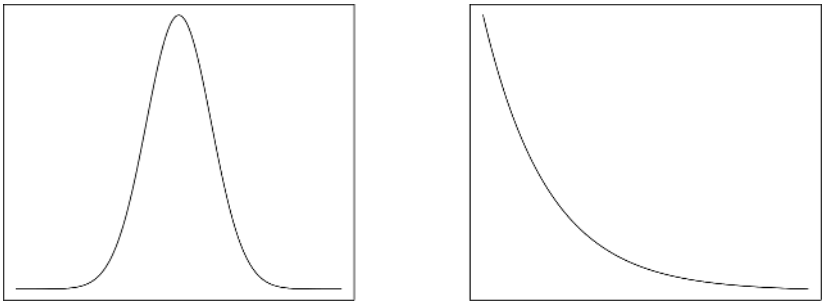

Figure 15.2. *A normal distribution (left panel); an exponential distribution (right panel)*

Once all the above questions have been answered and all other requirements have been checked, they usually point to one or two tests that address your question exactly. The following sections exemplify the choice of statistical tests and how they are then performed using some small examples. I am using the open-source language and programming environment R (www.r-project.org). Just as in many other scripting languages or spreadsheet applications, you perform (statistical) operations with *functions* (which tell R what to do, such as compute a log, a sum, or a mean), which take arguments (which tell R what to apply a function to and how). For example, `sum(c(1, 2, 3, 4, 5))` applies the function sum to one argument (a vector containing the numbers from 1 to 5), `mean(c(1, 2, 3, 4, 5))` computes the mean of the numbers from 1 to 5, and so on. The sections below will clarify this.

4 Performing significance tests and computing confidence intervals

This section exemplifies a small selection of frequently used tests; Section 4.1 exemplifies tests where the dependent variable is categorical; Section 4.2 exemplifies cases where central tendencies of ratio/interval and ordinal variables are tested.

4.1 Frequencies

This section introduces a goodness-of-fit test (Section 4.1.1), a test for independence (Section 4.1.2), and confidence intervals for percentages of categorical variables (Section 4.1.3).

4.1.1 The chi-square test for goodness of fit

This section discusses the test to use if you have answered the above questions as follows: you are conducting a study of one dependent categorical

variable and you want to test whether the observed frequencies of the variable's levels – which are independent of each other – are distributed as expected from a particular probability distribution (e.g., the uniform distribution) or previous results. For example, you asked fifty subjects to indicate whether they think that *walk* is a noun or a verb (of course it can be both – you are interested in the subjects' first responses), and you obtained responses such that thirty subjects said *verb* and twenty said *noun*. If you want to test whether these two observed frequencies, thirty and twenty, differ significantly from the chance expectation that subjects would have responded *verb* and *noun* equally often, then you compute a chi-square test for goodness of fit. In addition to the above criteria, this test also requires that 80 percent of expected frequencies are greater than or equal to five.

First, you enter the frequencies into R in the form of a so-called *vector* (a sequence of elements such as numbers or strings) and give names to the frequencies, using <- as an arrow-like assignment operator and the function c (for "concatenate"); anything in a line after a pound sign is ignored and merely serves to provide commentary.

```
walk <- c(30, 20) # create a vector with the observed frequencies
names(walk) <- c("verb", "noun") # name the observed frequencies
```

Then you compute the test using the function chisq.test with two arguments: the vector walk you just created, and a vector p of the expected probabilities, and since your H_0 expects the two parts of speech to be equally frequent, this is two times 0.5. The result of this test you assign to a data structure you can call, say, walk.test:

```
walk.test <- chisq.test(walk, p=c(0.5, 0.5)) # compute the chi-square test
```

Nothing is returned, but walk.test now contains all relevant results:

```
walk.test # show the result
      Chi-squared test for given probabilities
data: walk
X-squared = 2, df = 1, p-value = 0.1573
```

The results shows that the distribution of verbs and nouns does not differ significantly from chance: $p>0.05$. However, you should also make sure that the assumptions of the test were met so you compute the expected frequencies (which in this case you do not really need R for). Obviously, both expected frequencies exceed five so the use of the chi-square test was legitimate.

```
walk.test$expected # show the expected frequencies
verb noun
  25 25
```

If the assumption regarding the expected frequencies is not met, exact alternatives for dependent variables with 2 or 3+ levels are the binomial test and the multinomial test respectively; the former is already implemented in the R function binom.test.

Table 15.2 *Fictitious data from a forced-choice part-of-speech selection task*

	College education = yes	College education = no	Totals
walk = noun	16	4	20
walk = verb	9	21	30
Totals	25	25	50

4.1.2 The chi-square test for independence

The following scenario arises more frequently: you are conducting a study involving independent observations of two categorical variables, one dependent and one independent, and you want to test whether the observed frequencies of the levels of the dependent variable vary across the levels of the independent variable. For example, if in the above example you not only registered how often subjects considered *walk* a verb or noun, but also whether each subject had a college education or not, then you may have obtained the following result:

To determine whether the frequencies with which *walk* was classified as a noun or a verb are correlated with the subjects' level of education, you compute a chi-square test for independence, which has the same assumption regarding the expected frequencies as the chi-square test for goodness of fit.

Again, you begin by entering the data. This time, because the data are tabular, you use the function matrix with two arguments: a vector of observed frequencies by column and the table's number of columns (ncol). It is again also useful to provide names to the data by adding row and column names in the form of vectors using the function list:

```
walk <- matrix(c(16, 9, 4, 21), ncol=2) # create a matrix with the observed
      frequencies
dimnames(walk) <- list(Walk=c("noun", "verb"), Education=c(">= college",
       "< college")) # name the dimensions of the matrix
walk # look at the matrix
      Education
Walk   >= college < college
  noun          16         4
  verb           9        21
```

You then use the function chisq.test with the matrix walk as its only argument and assign the results to walk.test again (overwriting the earlier results):

```
walk.test <- chisq.test(walk) # compute the chi-square test
walk.test # show the result
       Pearson's Chi-squared test with Yates' continuity correction
data:  walk
X-squared = 10.0833, df = 1, p-value = 0.001496
```

For 2×2 tables R automatically applies a continuity correction to the data (see Sheskin 2007: 628f.); if that is not desired, use correct=FALSE as another

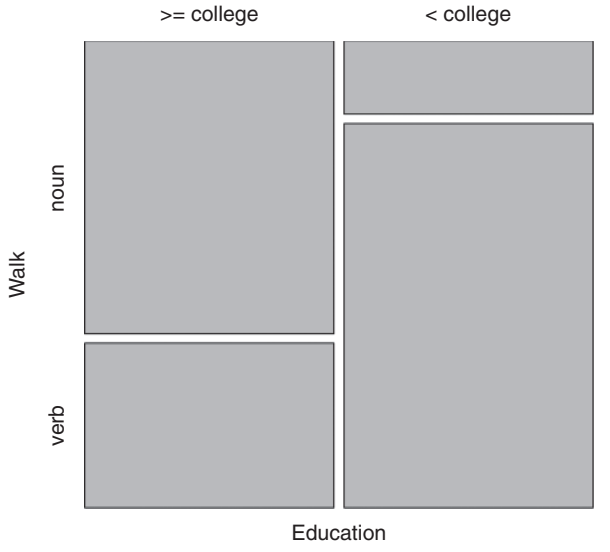

Figure 15.3. *Mosaic plot for the data in* walk

argument to chisq.test. The result shows there is a clear correlation between the part of speech assigned to *walk* and the education level of the subjects: $p<0.05$.

Before you explore what the correlation looks like, you should again test whether the expected-frequencies assumption is met, and it turns out it is:

```
walk.test$expected # show the expected frequencies
      Education
Walk   >= college < college
  noun         10        10
  verb         15        15
```

Finally, what kind of correlation do the data support? The quickest way to find out involves the so-called Pearson residuals, which correspond to the difference between each cell's observed minus its expected frequency, divided by the square root of the expected frequency. If a Pearson residual is positive/negative, then the corresponding observed frequency is greater/less than its expected frequency. Second, the more the Pearson residual deviates from 0, the stronger that effect. In R, this is easy to compute:

```
walk.test$residuals # show the Pearson residuals
      Education
Walk   >= college < college
  noun    1.897367 -1.897367
  verb   -1.549193  1.549193
```

The effect is that subjects with college education assigned the part of speech *noun* more often than expected, whereas subjects without a college degree assigned the part of speech *verb* more often than expected. This effect is also obvious from a graphical representation of the data (e.g., a so-called mosaic plot),

in which the large areas for the ">=college:noun" and "<college:verb" combinations represent the effect you already inferred from the residuals:

```
mosaicplot(t(walk)) # create a mosaic plot
```

(The t() just transposes the table so its row–column organization corresponds to the above matrix.) If the assumption regarding the expected frequencies is not met, an exact alternative for 2×2 tables is the Fisher-Yates exact test; this test, as well as extensions to variables with more than two levels, is implemented in the R function fisher.test.

4.1.3 Confidence intervals for percentages

This section is concerned with how to compute a confidence interval for an observed percentage. For example, in a corpus sample of 815 instances of the verb *to run*, you may have found that 203 of these (24.91 percent) involve the prototypical sense "fast pedestrian motion." To better evaluate that percentage in the population, you now want to determine its 95 percent confidence interval. The required R function is called prop.test, and it needs three arguments: the number of relevant instances in the sample that make up the percentage (a.k.a. successes), the overall sample size, and the argument correct=FALSE, which means that you do not apply a continuity correction (for the sake of comparison with other software):

```
run.ci <- prop.test(203, 815, correct=FALSE) # compute the confidence
      interval
run.ci$conf.int # show the confidence interval
[1] 0.2206115 0.2799023
attr(,"conf.level")
[1] 0.95
```

That is, following Sheskin's logic from above, you can be 95 percent confident that the true percentage of this sense out of all instances of *to run* is between 22.06 and 27.99 percent. If you apply this approach to the *walk* data from Section 4.1.1, you obtain the result shown below. Importantly, the non-significant result from above is suggested by the fact that the confidence intervals overlap.

```
walk.verb <- prop.test(30, 50, correct=FALSE) # compute the confidence
      interval
walk.verb$conf.int # show the confidence interval
[1] 0.4618144 0.7239161
attr(,"conf. level")
[1] 0.95

walk.noun <- prop.test(20, 50, correct=FALSE) # compute the confidence
      interval
walk.noun$conf.int # show the confidence interval
[1] 0.2760839 0.5381856
attr(,"conf. level")
[1] 0.95
```

4.2 Central tendencies

This section introduces a test of means from independent samples (Section 4.2.1), the corresponding test for medians (Section 4.2.2), a test of means from dependent samples (Section 4.2.3), and the computation of confidence intervals for means (Section 4.2.4).

4.2.1 The *t*-test for independent samples

This section introduces one of the best-known tests for central tendencies, which you apply if you are studying data involving a normally distributed ratio/interval-scaled dependent variable and a binary independent variable (with independent data points), and you want to test whether the averages of the dependent variable in the two groups (i.e., the two means) defined by the independent variables differ significantly from each other. For example, you may be interested in two different subtractive word-formation processes, blending and complex clipping. The former typically involves the creation of a new word by joining the beginning of a source word with the end of another (*brunch*, *foolosopher*, and *motel* are cases in point), whereas the latter involves fusing the beginnings of two source words (*scifi* and *sysadmin* are examples). You are now comparing the two processes in terms of how similar the source words are to each other, where said similarity is operationalized on the basis of the Dice coefficient, essentially the percentage of shared letter bigrams out of all bigrams.

As usual, the first step is to get the data into R, but in cases like these, you usually load them from a tab-separated file that was created with a spreadsheet software and has the so-called case-by-variable format: the first row contains the column names, the first column contains the case numbers, and each row describes a single observation in terms of the variables defined by the columns. Table 15.3 exemplifies this format on the basis of an excerpt of data from Gries (2013: Section 4.3.2.1).

This is how you read such a .txt file like the above into a data frame word.form in R:

```
word.form <- read.delim(file.choose()) # load the data from a text file
```

Table 15.3 *Dice coefficients of source words for complex clippings and blends*

CASE	PROCESS	DICE
1	ComplClip	0.0678
2	ComplClip	0.0704
3	ComplClip	0.0483
...	...	...
79	Blend	0.1523
80	Blend	0.1507

The data from each column are now available by combining the name of the data frame (word.form) with a dollar sign ($) and the column name (e.g., PROCESS). It is usually advisable to briefly check the structure of the data frame to make sure that importing the data has been successful; the function str displays the structure of a data structure:

```
str(word.form) # inspect the structure of the data
'data.frame':    80 obs. of  3 variables:
 $ CASE   : int 41  42 43 44  45 46 47  48  49 50 ...
 $ PROCESS: Factor w/ 2  levels "Blend","ComplClip": 2  2  2  2 2  2  2 ...
 $ DICE   : num  0.0678 0.0704 0.0483  0.0871 0.0813  0.0532 0.0675 ...
```

Apart from the above criteria for the *t*-test, especially the assumption of normality, the *t*-test also requires that the variances of the data points in the two groups are homogeneous (i.e., not significantly different). Since the default test for variance homogeneity also requires the data points to be normally distributed, this criterion should be tested first. One R function that can be used is shapiro.test, which takes a vector of data points and tests whether these data points differ significantly from normality. However, since we have two groups of data points – one for blends, one for complex clippings – there is a better way, using the function tapply:

```
tapply(word.form$DICE, word.form$PROCESS, shapiro.test) # test for normality
```

This means: take the values of tapply's first argument (word.form$DICE), split them up into groups by tapply's second argument (word.form$PROCESS), and apply tapply's third argument (shapiro.test) to each group:

```
$Blend

      Shapiro-Wilk normality test

data:  X[[1L]]
W = 0.9727, p-value = 0.4363


$ComplClip

      Shapiro-Wilk normality test

data:  X[[2L]]
W = 0.9753,  p-value = 0.5186
```

Both *p*-values are not significant, indicating that the Dice coefficients in each group do not differ from normality. You can therefore proceed to test whether the variance of one group of Dice coefficients differs significantly from the other. The function var.test can take a formula as input, which consists of a dependent variable, a tilde, and (an) independent variable(s).[1] In this case, word.form $DICE is the dependent variable, and word.form$PROCESS is the independent variable:

[1] If you cannot test for homogeneity of variances with var.test because your data violate the normality assumption, you can use the function fligner.test, which requires the same kind of formula as var.test.

```
var.test(word.form$DICE ~ word.form$PROCESS) # test for variance homogeneity
       F test to compare two variances
data:  word.form$DICE by word.form$PROCESS
F = 0.6632, num df = 39, denom df = 39, p-value = 0.2042
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3507687 1.2539344
sample estimates:
ratio of variances
         0.663205
```

Again, the *p*-value indicates a non-significant result: the variances do not differ from each other significantly and you can finally use the *t*-test for independent samples. The function is called t.test and it is usually used just like var.test (i.e., with a formula):

```
t.test(word.form$DICE ~ word.form$PROCESS) # compute a t-test
       Welch Two Sample t-test
data:  word.form$DICE by word.form$PROCESS
t = 16.4104, df = 74.928, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.05991431 0.07647069
sample estimates:
    mean in group Blend mean in group ComplClip
              0.1381300               0.0699375
```

Not only do you get the group means at the bottom, which show that the mean for blends is about twice as high as that for complex clippings, you also see that that result is highly significant: $p<<<0.05$. A graphical representation that summarizes such data in a very clear and comprehensive way is the so-called box plot, which requires the function box plot, a formula, and usually the argument notch=TRUE:

```
box plot(word.form$DICE ~ word.form$PROCESS, notch=TRUE) # create a box plot
```

This plot provides a lot of information and should be used much more often than it is:

- the thick horizontal lines correspond to the medians;
- the upper and lower horizontal lines indicate the central 50 percent of the data around the median (approximately the first and third quartiles);
- the upper and lower end of the whiskers extend to the most extreme data point which is no more than 1.5 times the length of the box away from the box;
- values outside of the range of the whiskers are marked individually as small circles;
- the notches of the boxes provide an approximate 95 percent confidence interval for the difference of the medians: if they do not overlap, then the medians are probably significantly different (see Sheskin 2007: 40–4 for very comprehensive discussion).

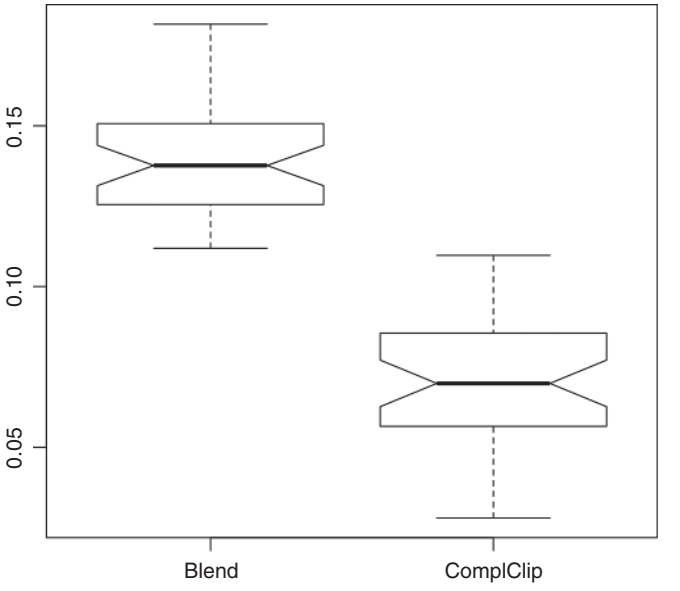

Figure 15.4. *Box plot of the Dice coefficients for the two subtractive word-formation processes*

If you cannot use the *t*-test because the data are not normally distributed, you can use the *U*-test instead, which is discussed in the following section. If you cannot use the *t*-test because the variances are not homogeneous, you can either use a version of the *t*-test which was designed to be less affected by unequal variances (the *t*-test by Welch, which is in fact R's default) or again the *U*-test.

4.2.2 The *U*-test

There are two main reasons to use a *U*-test. One is that you are studying data involving an ordinal-scaled dependent variable and a binary independent variable (with independent data points), and you want to test whether the averages of the dependent variable in the two groups (i.e., the two medians) defined by the independent variables differ significantly from each other. The other was mentioned at the end of the previous section: you have data that would usually be analyzed with a *t*-test for independent samples, but assumptions of the *t*-test are not met. The *U*-test also assumes that the data in the two groups are from populations that are distributed identically, but violations to this requirement affect the test results much less than those of the *t*-test (which is probably why this criterion is often not even mentioned in textbooks.)

Given the overall similarity of the two tests and in the interest of brevity, I will exemplify the *U*-test only on the basis of the same data as the *t*-test for independent

samples. The name of the required function is wilcox.test and it requires the by now already familiar formula as input:

```
wilcox.test(word.form$DICE ~ word.form$PROCESS) # compute a U-test
      Wilcoxon rank sum test with continuity correction
data: word.form$DICE by word.form$PROCESS
W = 1600, p-value = 1.434e-14
alternative hypothesis: true location shift is not equal to 0
```

Given the large difference between the medians (recall Figure 15.4) and the highly significant result of the *t*-test, it is not surprising that the *U*-test also returns a highly significant result. (R also returns a warning not shown above because of the fact that there are three ties – i.e., three Dice values that are attested more than once. However, this is no need for concern since R automatically adjusts the way the *p*-value is computed accordingly.)

4.2.3 The *t*-test for dependent samples

The *t*-test discussed in Section 4.2.1 above involved a test of means from independent samples – in this section, I will discuss its counterpart for dependent samples. More specifically, you use the *t*-test for dependent samples if your data involve two groups of pairwise-associated data points on a ratio/interval scale and you want to test whether the means of the two groups are significantly different. The *t*-test for dependent samples also comes with the additional requirement that the pairwise differences between the samples' data points are normally distributed.[2]

As an example, consider a case where ten students take a grammar test and score a particular number of points. Then, they participate in an exercise session on the tested grammar topic and take a second grammar test; the question is whether their scores have changed. First, you enter the data of the ten subjects into two vectors before and after; crucially, the data points have to be in the same order for both before and after. That is, if the first data point of before belongs to subject 1, then so must the first data point of after, and so on.

```
before <- c(4 ,17, 8, 7, 13, 13, 3, 6, 12, 13) # create the 1st vector
after <- c(16, 16, 16, 17, 23, 22, 8, 20, 23, 11) # create the 2nd vector
```

To compute the pairwise differences between the two tests, you just subtract one vector from the other; R will perform a pairwise computation for you:

```
differences <- before-after # compute pairwise differences
differences # show the pairwise differences
[1] -12   1  -8 -10 -10  -9  -5 -14 -11   2
```

[2] Reference works differ with regard to this criterion. Some cite the criterion mentioned above (that the pairwise differences must be normally distributed); others state that the data points in the two populations represented by the two samples must be normally distributed.

To test whether these differences are normally distributed, you can proceed as before:

```
shapiro.test(differences) # test for normality
      Shapiro-Wilk normality test
data:  differences
W = 0.874, p-value = 0.1112
```

Obviously, they are: $p>0.05$, which means you can perform a *t*-test for dependent samples. The function for this test is again t.test, but there are two small changes. First, to indicate that this time you need a *t*-test for dependent samples, you add the argument paired=TRUE. Second, when you did the *t*-test for independent samples and the *U*-test, you had one vector/factor per variable: the vector DICE for the dependent variable and the factor PROCESS for the independent variable, and then you used a formula. This time, you have one vector per level of the independent variable: one for the level "test before the treatment" (before) and one for the level "test after the treatment" (after). That means you cannot use the formula notation, but you just separate the two vectors with a comma:

```
t.test(before, after, paired=TRUE) # compute t-test for dependent samples
      Paired t-test
data: before and after
t = -4.4853, df = 9, p-value = 0.001521
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-11.433079 -3.766921
sample estimates:
mean of the differences
                  -7.6
```

The result of the second test is on average 7.6 points better than the first, and that difference is very significant; $p<0.01$: it seems as if the treatment led to a substantial increase – in fact, to an increase of nearly 80 percent (since the means of before and after are 9.6 and 17.2 respectively). How can this result be represented graphically? One way would be to plot the vector differences in the form of a histogram. You use the function plot with the argument type="h" to plot the histogram, and the argument sort(differences) sorts the differences to be plotted in ascending order; the remaining arguments define the x- and y-axis labels to yield the left panel of Figure 15.5. It is plain to see that most differences are highly negative, which shows that the after values are larger.

```
plot(sort(differences), type="h", xlab="Subject", ylab="Difference: before -
      after") # plot a histogram of the differences
```

A second graphical representation is shown in the right panel: Each subject is represented by an arrow from that subject's score in the before-treatment test to the subject's score in the after-treatment test. The improvement is reflected in the fact that most arrows go upward, and the two numbers on the left indicate the speakers whose results did not improve.

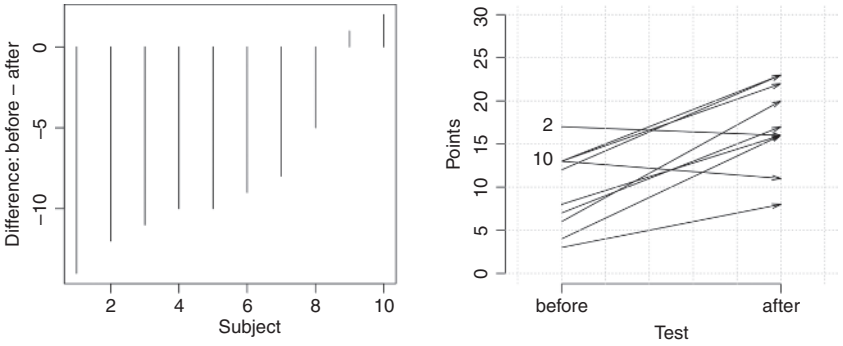

Figure 15.5. *Graphical representation of the differences between* before *and* after

If you cannot use a *t*-test for dependent samples, a non-parametric alternative is the Wilcoxon test. The function for this test is again wilcox.test, just add the argument paired=TRUE.

4.2.4 Confidence intervals for means

This section is concerned with how to compute a confidence interval for an observed mean. For example, you may have conducted a second experiment of the type in the previous section, with more subjects participating in a before- and an after-treatment test. You now want to know the mean of this second after-treatment test, as well as its 95 percent confidence interval. First, you enter the data:

```
after.2 <- c(10, 21, 8, 15, 23, 11, 12, 11, 13, 15, 21, 10, 9, 14, 9, 14, 12,
      4, 16, 13, 19, 19, 22, 18, 19) # enter the data
```

The function to compute confidence intervals for means is again t.test, and it requires the vector with the data points and conf.level with the desired confidence level. However, the computation of such a confidence interval requires that the data are distributed normally, which is why you need to run shapiro.test again.

```
shapiro.test(after.2) # test for normality
data: after.2
W = 0.9698, p-value = 0.6406
```

The data are distributed normally so you can proceed:

```
t.test(after.2, conf.level=0.95) # compute the confidence interval
data:  after.2
t = 14.5218, df = 24, p-value = 2.194e-13
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 12.28478 16.35522
sample estimates:
mean of x
    14.32
```

The mean number of points of this second group's after-treatment results after.2 is 14.32, with a 95 percent confidence interval of 12.28 (lower bound) and 16.36 (upper bound). If you compute the same kind of confidence interval for after, you will see that the confidence intervals of after and after.2 overlap, which suggests – but not demonstrates – that the means are not significantly difference (which is confirmed by a *t*-test).

5 Some final remarks

The above discussion could only discuss a small selection of tests and their assumptions and application. In this final section, I will briefly discuss four notions that are worth exploring: directional hypotheses, transformations, missing data, and multiple/post hoc tests.

First, given space constraints, all of the above discussed so-called *non-directional* alternative hypotheses and *two-tailed tests* – that is, tests of hypotheses that postulate a difference/an effect, but not the direction of said difference (e.g., *a* is not equal to *b*). However, if you not only expect some difference, but also the direction of that difference (*a* is larger than *b*), you can formulate a *directional hypothesis* and compute a *one-tailed test*. This is advantageous because, if you have a directional hypothesis, the effect you need to find in order to get a significant result is smaller; in other words, your prior knowledge will be rewarded. Thus, this should be among the first topics for further study.

Second, we have seen that parametric tests of ratio/interval data rely on distributional assumptions that need to be tested before, say, a *t*-test for independent samples can be computed. If those assumptions are not met, then one way to proceed is to use a test for ordinal data, as was discussed above at the end of Section 4.2.1. However, not only are tests that only utilize the ordinal information of data less powerful than their parametric counterparts, but for many more complex tests, non-parametric or exact alternatives are also not readily available. Therefore, an alternative to non-parametric tests is to apply a *transformation* to the original data, which, if the right transformation is applied correctly, can reduce the impact of outliers, normalize distributions, and homogenize variances. The most frequently used transformations of a vector x are the square-root transformation (sqrt(x)), the logarithmic transformation (log(x)), the reciprocal transformation (1/x), the arcsine transformation (2*asin(sqrt(0.25)) or asin(sqrt(0.25))), and the square transformation (x^2); if your data violate distributional assumptions, such transformation may be quite useful.

Third, observational and experimental data are often incomplete: particular types of corpus examples are not attested or cannot be annotated unambiguously; subjects do not respond to particular stimuli or do not show up for the after-treatment test. While a detailed treatment of the analysis of *missing data* is beyond the scope of this chapter, it is important to point out that missing data must not be ignored: they should be carefully recorded and investigated for patterns to

determine whether they are in fact already a noteworthy and interpretable finding in and of themselves. For example, if a particular experimental stimulus exhibits a large number of non-responses, this may reveal something interesting about that stimulus or the studied hypothesis, or it may lead to you discarding the data from that stimulus from the statistical analysis. Thus, an analysis of missing data should be an indispensable analytical step.

Finally, a word on multiple/post hoc tests. Multiple testing arises when you perform several significance tests on the same dataset, and they are post hoc if you (decide to) perform these multiple tests only after you have performed a first test. An example of the first situation would be if you collected reaction times to words as well as, say, six predictors describing the words, and then ran all possible (six) pairwise correlations between the predictors and the reaction times as opposed to one multifactorial study. An example of the second situation would be if you tested the effect of one categorical independent variable with four levels *a*, *b*, *c*, and *d* on a ratio/interval dependent variable, obtained a significant result, and then ran all six pairwise comparisons of means: *a* vs *b*, *a* vs *c*, *a* vs *d*, *b* vs *c*, *b* vs *d*, and *c* vs *d*. The first situation is problematic because you might be accused of "fishing for results," but also for an additional statistical reason which also applies to the second situation: If you perform one significance test with a significance level of 95 percent, there is a probability of 0.05 that the decision to reject the null hypothesis is wrong. However, if you perform n independent significance tests each with a significance level of 95 percent, there is a probability of $1–0.95^n$ that at least one rejection of a null hypothesis is wrong; for n=6, this probability is already $1–0.95^6 \approx 0.265$. Thus, when you perform multiple tests, it is common practice to adjust your significance level from 95 percent for each test. For six post hoc tests, it would be necessary to adjust the significance level to 99.14876 percent, because $0.9914876^6 = 0.95$. However, even with such a so-called *correction for multiple testing*, testing all possible null hypotheses is to be discouraged. More discussion of these topics and all previous ones can be found in Crawley (2007); Baayen (2008); Johnson (2008); and Gries (2013).

References

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Cohen, J. 1994. The earth is round (p<.05). *American Psychologist* 49.12: 997–1003.

Crawley, M. J. 2007. *The R Book*. Chichester: John Wiley and Sons.

Gries, S. Th. 2013. *Statistics for Linguistics Using R: A Practical Introduction*, 2nd edn. Berlin and New York: Mouton de Gruyter.

Johnson, K. 2008. *Quantitative Methods in Linguistics*. Oxford and Malden, MA: Blackwell.

Sheskin, D. J. 2007. *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th edn. Boca Raton, FL: Chapman & Hall/CRC.