Chapter 1
Some fundamentals of empirical research

> When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind.
> It may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.
> William Thomson, Lord Kelvin.
> (<http://hum.uchicago.edu/~jagoldsm/Webpage/index.html>)

1. Introduction

This book is an introduction to statistics. However, there are already very many introductions to statistics – why do we need another one? Well, this book is different from most other introductions to statistics in several ways:

- it has been written especially for linguists: there are many introductions to statistics for psychologists, economists, biologists etc., but only very few which, like this one, explain statistical concepts and methods on the basis of linguistic questions and for linguists;
- it explains how to do nearly all of the statistical methods both 'by hand' as well as with statistical software, but it requires neither mathematical expertise nor hours of trying to understand complex equations – many introductions devote much time to mathematical foundations (and, thus, make everything more difficult for the novice), others do not explain any foundations and immediately dive into some nicely designed software, which often hides the logic of statistical tests behind a nice GUI;
- it not only explains statistical concepts, tests, and graphs, but also the design of tables to store and analyze data, summarize previous literature, and some very basic aspects of experimental design;
- it only uses open source software (mainly R): many introductions use SAS or in particular SPSS, which come with many disadvantages such that (i) users must to buy expensive licenses that are restricted in how many functions they offer and how many data points they can handle)

and how long they can be used; (ii) students and professors may be able to use the software only on campus; (iii) they are at the mercy of the software company with regard to bugfixes and updates etc.;
- it does all this in an accessible and informal way: I try to avoid jargon wherever possible; the use of software will be illustrated in very much detail, and there are think breaks, warnings, exercises (with answer keys on the companion website), and recommendations for further reading etc. to make everything more accessible.

So, this book aims to help you do scientific quantitative research. It is structured as follows. Chapter 1 introduces the foundations of quantitative studies: what are variables and hypotheses, what is the structure of quantitative studies and what kind of reasoning underlies it, how do you obtain good experimental data, and in what kind of format should you store your data?

Chapter 2 provides an overview of the programming language and environment R, which will be used in all other chapters for statistical graphs and analyses: how do you create, load, and manipulate data to prepare for your analysis?

Chapter 3 explains fundamental methods of descriptive statistics: how do you describe your data, what patterns can be discerned in them, and how can you represent such findings graphically? Chapter 4 explains fundamental methods of analytical statistics: how do you test whether the obtained results actually mean something or have just arisen by chance? Chapter 5 introduces several multifactorial procedures, i.e. procedures, in which several potential cause-effect relations are investigated simultaneously. This chapter can only deal with a few selected methods and I will point you to additional references quite a few times.

Apart from the following chapters with their think breaks and exercises etc., the companion website for this book at <http://www.linguistics.ucsb.edu/faculty/stgries/research/sflwr/sflwr.html> (or its mirror at <http://groups.google.com/group/statforling-with-r/web/statistics-for-linguists-with-r>) is an important resource. You will have to go there anyway to download exercise files, data files, answer keys, errata etc., but at <http://groups.google.com/group/statforling-with-r> you will also find a newsgroup "StatForLing with R". I would like to encourage you to become a member of that newsgroup so that you can

- ask questions about statistics for linguists (and hopefully also get an answer from some kind soul);

- send suggestions for extensions and/or improvements or data for additional exercises;
- inform me and other readers of the book about bugs you find (and of course receive such information from other readers). This also means that if R commands, or *code*, provided in the book differs from that on the website, then the latter is most likely going to be correct.

Lastly, I have to mention one important truth right at the start: you cannot learn to do statistical analyses by reading a book about statistical analyses. You must *do* statistical analyses. There is no way that you read this book (or any other serious introduction to statistics) 15 minutes in bed before turning off the light and learn to do statistical analyses, and book covers or titles that tell you otherwise are, let's say, 'distorting' the truth for marketing reasons. I strongly recommend that, as of the beginning of Chapter 2, you work with this book directly at your computer so that you can immediately enter the R code that you read and try out all relevant functions from the code files from the companion website; often (esp. in Chapter 5), the code files for this chapter will provide you with important extra information, additional code snippets, further suggestions for explorations using graphs etc., and sometimes the exercise files will provide even more suggestions and graphs. Even if you do not understand every aspect of the code right away, this will still help you to learn all this book tries to offer.

2. On the relevance of quantitative methods in linguistics

Above I said this book introduces you to scientific quantitative research. But then, what are the goals of such research? Typically, one distinguishes three goals, which need to be described because (i) they are part of a body of knowledge that all researchers within an empirical discipline should be aware of and (ii) they are relevant for how this book is structured.

The first goal is the *description* of your data on some phenomenon and means that your data and results must be reported as accurately and revealingly as possible. All statistical methods described below will help you achieve this objective, but particularly those described in Chapter 3.

The second goal is the *explanation* of your data, usually on the basis of hypotheses about what kind(s) of relations you expected to find in the data. On many occasions, this will already be sufficient for your purposes. However, sometimes you may also be interested in a third goal, that of *prediction*: what is going to happen in the future or when you look at different data. Chapters 4 and 5 will introduce you to methods to pursue the goals of explanation and prediction.

When you look at these goals, it may appear surprising that statistical methods are not that widespread in linguistics. This is all the more surprising because such methods are very widespread in disciplines with similarly complex topics such as psychology, sociology, economics. To some degree, this situation is probably due to how linguistics has evolved over the past decades, but fortunately it is changing now. The number of studies utilizing quantitative methods has been increasing (in all linguistic subdisciplines); the field is experiencing a paradigm shift towards more empirical methods. Still, even though such methods are commonplace in other disciplines, they still often meet some resistance in linguistic circles: statements such as "we've never needed something like that before" or "the really interesting things are qualitative in nature anyway and are not in need of any quantitative evaluation" or "I am a field linguist and don't need any of this" are far from infrequent.

Let me say this quite bluntly: such statements are not particularly reasonable. As for the first statement, it is not obvious that such quantitative methods were not needed so far – to prove that point, one would have to show that quantitative methods could impossibly have contributed something useful to previous research, a rather ridiculous point of view – and even then it would not necessarily be clear that the field of linguistics is not now at a point where such methods are useful. As for the second statement, in practice quantitative and qualitative methods go hand in hand: qualitative considerations precede and follow the results of quantitative methods anyway. To work quantitatively does not mean to just do, and report on, some number-crunching – of course, there must be a qualitative discussion of the implications – but as we will see below often a quantitative study allows to identify what merits a qualitative discussion. As for the last statement: even a descriptive (field) linguist who is working to document a near-extinct language can benefit from quantitative methods. If the chapter on tense discusses whether the choice of a tense is correlated with indirect speech or not, then quantitative methods can show whether there is such a correlation. If a study on middle voice in the Athabaskan language Dena'ina tries to identify how syntax and semantics are related to middle voice marking, quantitative methods can reveal interesting things (cf. Berez and Gries 2010).

The last two points lead up to a more general argument already alluded to above: often only quantitative methods can separate the wheat from the chaff. Let's assume a linguist wanted to test the so-called aspect hypothesis according to which imperfective and perfective aspect are preferred in present and past tense respectively (cf. Shirai and Andersen 1995). Strictly speaking, the linguist would have to test all verbs in all languages, the so-called *population*. This is of course not possible so the linguist studies a *sample* of sentences to investigate their verbal morphology. Let's further assume the linguist took and investigated a small sample of 39 sentences in one language and got the results in Table 1.

Table 1. A fictitious distribution of tenses and aspects in a small corpus

	Imperfective	Perfective	Totals
Present tense	12	6	18
Past tense	7	13	20
Totals	19	19	38

These data look like a very obvious confirmation of the aspect hypothesis: there are more present tenses with imperfectives and more paste tenses with perfectives. However, the so-called chi-square test, which could be used for these data, shows that this tense-aspect distribution can arise by chance with a probability *p* that exceeds the usual threshold of 5% adopted in quantitative studies. Thus, the linguist would not be allowed to accept the aspect hypothesis for the population on the basis of this sample. The point is that an intuitive eye-balling of this table is insufficient – a statistical test is needed to protect the linguist against invalid generalizations.

An more eye-opening example is discussed by Crawley (2007: 314f.). Let's assume a study showed that two variables *x* and *y* are correlated such that the larger the value of *x*, the larger the value of *y*; cf. Figure 1.

Note, however, that the data actually also contain information about a third variable (with seven levels *a* to *g*) on which *x* and *y* depend. Interestingly, if you now inspect what the relation between *x* and *y* looks like for each of the seven levels of the third variable, you see that the relation suddenly becomes "the larger *x*, the smaller *y*"; cf. Figure 2, where the seven levels are indicated with letters. Such patterns in data are easy to overlook – they can only be identified through a careful quantitative study, which is why knowledge of statistical methods is indispensible.

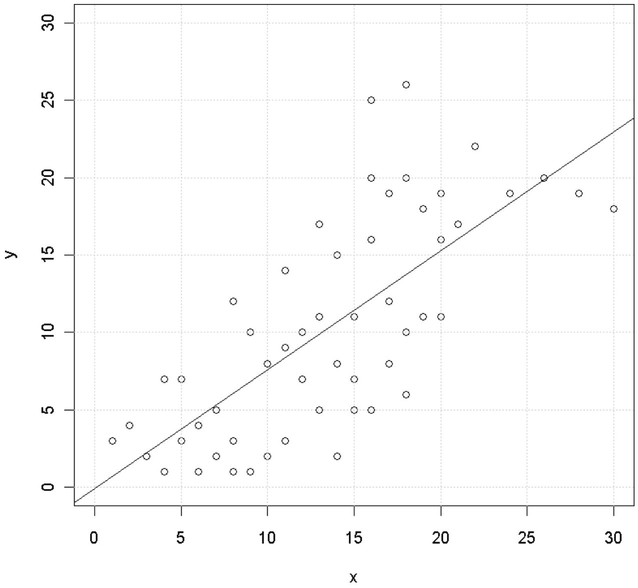


Figure 1. A fictitious correlation between two variables *x* and *y*

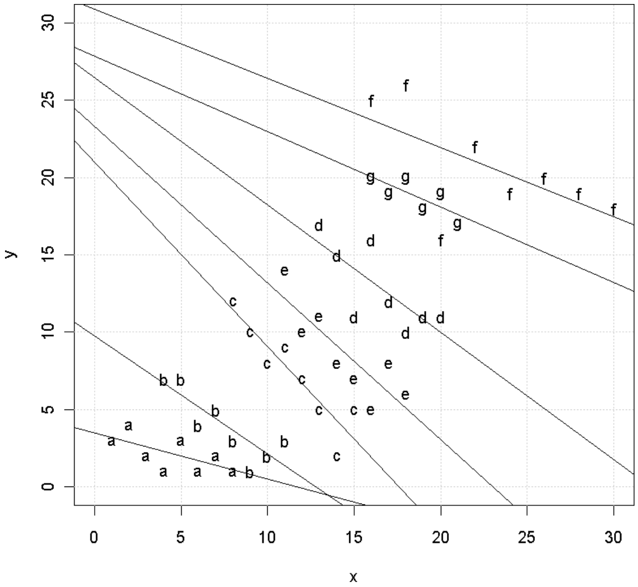


Figure 2. A fictitious correlation between two *variables x* and *y*, controlled for a third variable

For students of linguistics – as opposed to experienced practitioners – there is also a very practical issue to consider. Sometime soon you will want to write a thesis or dissertation. Quantitative methods can be extremely useful and powerful if only to help you avoid the pitfalls posed by the data in Table 1 and Figure 1. In the light of all this, it is hopefully obvious now that quantitative methods have a lot to offer, and I hope this book will provide you with some good and practical background knowledge.

This argument has an additional aspect to it. Contrary to, say, literary criticism, linguistics is an empirical science. Thus, it is necessary – in particular for students – to know about basic methods and assumptions of empirical research and statistics to be able to understand both scientific argumentation in general and linguistic argumentation in particular. This is especially relevant in the domains of, for example, contemporary quantitative corpus linguistics or psycholinguistics, where data are often evaluated with such a high degree of sophistication that a basic knowledge of the relevant terminology is required. Without training, what do you make of statements such as "The interaction between the size of the object and the size of the reference point does not reach standard levels of significance: $F_{1, 12} = 2.18$; $p = 0.166$; *partial eta*^{2} = 0.154."? Who knows off the top of their head whether the fact that the average sentence length of ten female second language learners in an experiment was about two words larger than the average sentence length of ten male second language learners is more likely to mean something or whether this is more likely a product of chance? Again, such data need serious statistical analysis.

3. The design and the logic of quantitative studies

In this section, we will have a very detailed look at the design of, and the logic underlying, quantitative studies. I will distinguish several phases of quantitative studies and consider their structure and discuss the reasoning employed in them. The piece of writing in which you then describe your quantitative research should usually have four parts: *introduction*, *methods*, *results*, and *discussion* – if you discuss more than one case study in your writing, then typically each case study gets its own methods, results, and discussion sections, followed by a general discussion.

With few exceptions, the discussion and exemplification in this section will be based on a linguistic example. The example is the phenomenon of particle placement in English, i.e. the constituent order alternation of transitive phrasal verbs exemplified in (1).

(1) a. He picked up [$_{NP}$ the book].
CONSTRUCTION: *VPO* (verb - particle - object)
b. He picked [$_{NP}$ the book] up.
CONSTRUCTION: *VOP* (verb - object - particle)

An interesting aspect of this alternation is that, most of the time, both constructions appear to be quite synonymous and native speakers of English usually cannot explain why they produce (1a) on one occasion and (1b) on some other occasion. In the past few decades, linguists have tried to describe, explain, and predict the alternation (cf. Gries 2003a for a recent overview). In this section, we will use this alternation to illustrate the structure of a quantitative study.

3.1. Scouting

At the beginning of your study, you want to get an overview of previous work on the phenomenon you are interested in, which also gives you a sense of what still can or needs to be done. In this phase, you try to learn of existing theories that can be empirically tested or, much more infrequently, you enter uncharted territory in which you are the first to develop a new theory. This is a list of the activities that is typically performed in this phase:

- a first (maybe informal) characterization of the phenomenon;
- studying the relevant literature;
- observations of the phenomenon in natural settings to aid first inductive generalizations;
- collecting additional information (e.g., from colleagues, students, etc.);
- deductive reasoning on your part.

If you take just a cursory look at particle placement, you will quickly notice that there is a large number of variables that influence the constructional choice. A variable is defined as a symbol for a set of states, i.e., a characteristic that – contrary to a constant – can exhibit at least two different states or levels (cf. Bortz and Döring 1995: 6 or Bortz 2005: 6) or, more intuitively, as “descriptive properties” (Johnson 2008: 4) or as measurements of an item that can be either numeric or categorical (Evert, p.c.).

Variables that might influence particle placement include the following:[1]

- COMPLEXITY: is the direct object a *SIMPLE DIRECT OBJECT* (e.g., *the book*), a *PHRASALLY-MODIFIED DIRECT OBJECT* (e.g., *the brown book* or *the book on the table*) or a *CLAUSALLY-MODIFIED DIRECT OBJECT* (e.g., *the book I had bought in Europe*) (cf., e.g., Fraser 1966);
- LENGTH: the length of the direct object (cf., e.g., Chen 1986, Hawkins 1994);
- DIRECTIONAL OBJECT: the *PRESENCE* of a directional prepositional phrase (PP) after the transitive phrasal verb (e.g. in *He picked the book up from the table*) or its *ABSENCE* (cf. Chen 1986);
- ANIMACY: whether the referent of the direct object is *INANIMATE* as in *He picked up the book* or *ANIMATE* as in *He picked his dad up* (cf. Gries 2003a: Ch. 2);
- CONCRETENESS: whether the referent of the direct object is *ABSTRACT* as in *He brought back peace to the region* or *CONCRETE* as in *He picked his dad up at the station* (cf. Gries 2003a: Ch. 2);
- TYPE: is the part of speech of the head of the direct object a *PRONOUN* (e.g., *He picked him up this morning*), a *SEMIPRONOUN* (e.g., *He picked something up from the floor*), a *LEXICAL NOUN* (e.g., *He picked people up this morning*) or a *PROPER NAME* (e.g., *He picked Peter up this morning*) (cf. Van Dongen 1919).

During this early phase, it is often useful to summarize your findings in tabular format. One possible table summarizes which studies (in the columns) discussed which variable (in the rows). On the basis of the above list, this table could look like Table 2 and allows you to immediately recognize (i) which variables many studies have already looked at and (ii) the studies that looked at most variables. Another table summarizes the variable levels and their preferences for one of the two constructions. Again, on the basis of the above list, this table would look like Table 3, and you can immediately see that, for some variables, only one level has been associated with a particular constructional preference.

[1] I print variables in small caps; their levels in italicized small caps.

Table 2. Summary of the literature on particle placement I

	Fraser (1966)	Chen (1986)	Hawkins (1994)	Gries (2003a)	Van Dongen (1919)
Complexity	x				
Length		x	x		
DirectionalPP		x			
Animacy				x	
Concreteness				x	
Type					x

Table 3. Summary of the literature on particle placement II

	Variable level for Construction: *vpo*	Variable level for Construction: *vop*
Complexity	*phrasally-modified* *clausally modified*	
Length	*long*	
DirectionalPP	*absence*	*presence*
Animacy	*inanimate*	*animate*
Concreteness	*animate*	*concrete*
Type		*pronominal*

This table already suggests that Construction: *vpo* is used with cognitively more complex direct objects: long complex noun phrases with lexical nouns referring to abstract things. Construction: *vop* on the other hand is used with the opposite preferences. For an actual study, this first impression would of course have to be phrased more precisely. In addition, you should also compile a list of other factors that might either influence particle placement directly or that might influence your sampling of sentences or experimental subjects or … Much of this information would be explained and discussed in the first section of the empirical study, the introduction.

3.2. Hypotheses and operationalization

Once you have an overview of the phenomenon you are interested in and have decided to pursue an empirical study, you usually have to formulate hypotheses. What does that mean and how do you proceed? To approach this issue, let us see what hypotheses are and what kinds of hypotheses there are.

3.2.1. Scientific hypotheses in text form

Following Bortz and Döring (1995: 7), we will consider a hypothesis to be a statement that meets the following three criteria:

- it is a general statement that is concerned with more than just a singular event;
- it is a statement that at least implicitly has the structure of a conditional sentence (*if …, then …* or *the …, the …*) or can be paraphrased as one;
- it is potentially falsifiable, which means it must be possible to think of events or situations or states of affairs that contradict the statement. Most of the time, this implies that the scenario described in the conditional sentence must also be testable. However, these two characteristics are not identical. There are statements that are falsifiable but not testable such as "If children grow up without any linguistic input, then they will grow up to speak Latin." This statement is falsifiable, but for obvious ethical reasons not testable (cf. Steinberg 1993: Section 3.1).

The following statement is a scientific hypothesis according to the above criteria: "Reducing the minimum age to obtain a driver's license from 18 years to 17 years in European countries will double the number of traffic accidents in these countries." This statement is a general statement that is not restricted to just one point of time, just one country, etc. Also, this statement can be paraphrased as a conditional sentence: "If one reduces the minimum age …, then the number of traffic accidents will double." Lastly, this statement is falsifiable because it is conceivable – actually, very likely – that if one reduced the minimum age, that the number of traffic accidents would not double. Accordingly, the following statement is not a scientific hypothesis: "Reducing the minimum age to obtain a driver's license from 18 years to 17 years in European countries may double the number of traffic accidents in these countries." This statement is testable because the minimum age could be reduced, but it is not falsifiable since the word *may* basically means 'may or may not': the statement is true if the number of traffic accidents doubles, but also if it does not. Put differently, whatever one observed after the reduction of the minimum age, it would be compatible with the statement.

With regard to particle placement, the following statements are examples for scientific hypotheses:

- if the direct object of a transitive phrasal verb is syntactically complex, then native speakers will produce the constituent order *VPO* more often than when the direct object is syntactically simple;

- if the direct object of a transitive phrasal verb is long, then native speakers will produce the constituent order *VPO* more often than when the direct object is short;
- if a verb-particle construction is followed by a directional PP, then native speakers will produce the constituent order *VOP* more often than when no such directional PP follows (and analogously for all other variables mentioned in Table 3).

When you formulate a hypothesis, it is also important that the notions that you use in the hypothesis are formulated precisely. For example, if a linguistic theory uses notions such as *cognitive complexity* or *availability in discourse* or even something as seemingly straightforward as *constituent length*, then it will be necessary that the theory can define what exactly is meant by this; in Section 1.3.2.2 we will deal with this aspect in much more detail.

According to what we have said so far, a hypothesis consists of two parts, an *if* part (*I*) and a *then* part (*D*). The *I* stands for *independent variable*, the variable in the *if* part of the hypothesis that is often, but not necessarily, the cause of the changes/effects in the *then* part of the hypothesis. The *D* on the other hand stands for *dependent variable*, the variable in the *then* part of the hypothesis and whose values, variation, or distribution is to be explained.[2]

With this terminology, we can now paraphrase the above example hypotheses. In the first, *I* is the syntactic complexity of the direct object (COMPLEXITY with the three levels *SIMPLE*, *PHRASALLY-MODIFIED*, and *CLAUSALLY-MODIFIED*), and *D* is the choice of construction (CONSTRUCTION with the two levels *VPO* and *VOP*). In the second hypothesis, *I* is the length of the direct object (LENGTH with values from 1 to *x*), and *D* is again the choice of construction (CONSTRUCTION with the two levels *VPO* and *VOP*), etc.

A second kind of hypothesis only contains one (dependent) variable, but no independent variable with which the dependent variable's behavior is explained. In such cases, the hypothesis is 'only' a statement about what the values, variation, or distribution of the dependent variable look like. In

[2] Variables such as moderator or confounding variables will not be discussed here; cf. Matt and Cook (1994).

what follows, we will deal with both kinds of hypotheses (with a bias toward the former).

Thus, we can also define a scientific hypothesis as a statement about one variable, or a statement about the relation(s) between two or more variables in some context which is expected to also hold in similar contexts and/or for similar objects in the population. Once the potentially relevant variables to be investigated have been identified, you formulate a hypothesis by relating the relevant variables in the appropriate conditional sentence or some paraphrase thereof.

Once your hypothesis has been formulated in the above text form, you also have to define – before you collect data! – which situations or states of affairs would falsify your hypothesis. Thus, in addition to your own hypothesis – the so-called *alternative hypothesis* H_1 – you now also formulate another hypothesis – the so-called *null hypothesis* H_0 – which is the logical opposite to your alternative hypothesis. Often, that means that you get the null hypothesis by inserting the word *not* into the alternative hypothesis. For the first of the above three hypotheses, this is what both hypotheses would look like:

H_1: If the direct object of a transitive phrasal verb is syntactically complex, then native speakers will produce the constituent order *VPO* more often than when the direct object is syntactically simple.

H_0: If the direct object of a transitive phrasal verb is syntactically complex, then native speakers will *not* produce the constituent order *VPO* more often than when the direct object is syntactically simple.

In the vast majority of cases, the null hypothesis states that the dependent variable is distributed randomly (or in accordance with some well-known mathematically definable distribution such as the normal distribution), or it states that there is no difference between (two or more) groups or no relation between the independent variable(s) and the dependent variable(s) and that whatever difference or effect you get is only due to chance or random variation. However, you must distinguish two kinds of alternative hypotheses: *directional alternative hypotheses* not only predict that there is some kind of effect or difference or relation but also the direction of the effect – note the expression “more often” in the above alternative hypothesis. On the other hand, *non-directional alternative hypotheses* only predict that there is some kind of effect or difference or relation without specifying the direction of the effect. A non-directional alternative hypothesis for the above example would therefore be this:

$H_{1\ \text{non-directional}}$: If the direct object of a transitive phrasal verb is syntactically complex, then native speakers will produce the constituent order *VPO differently often* than when the direct object is syntactically simple.

Thus, H_0 states that there is no correlation between the syntactic complexity of a direct object and the constructional choice and that if you nevertheless find one in the sample, then this is only a chance effect. Both H_1's state that there is a correlation – thus, you should also find one in your sample. Both of these hypotheses must be formulated *before* the data collection so that one cannot present whatever result one gets as the 'predicted' one. Of course, all of this has to be discussed in the introduction of the written version of your paper.

3.2.2. Operationalizing your variables

Formulating your hypotheses in the above text form is not the last step in this part of the study, because it is as yet unclear how the variables invoked in your hypotheses will be investigated. For example and as mentioned above, a notion such as cognitive complexity can be defined in many different and differently useful ways, and even something as straightforward as constituent length is not always as obvious as it may seem: do we mean the length of, say, a direct object in letters, phonemes, syllables, morphemes, words, syntactic nodes, etc.? Therefore, you must find a way to *operationalize* the variables in your hypothesis. This means that you decide what will be observed, counted, measured etc. when you investigate your variables.

For example, if you wanted to operationalize a person's KNOWLEDGE OF A FOREIGN LANGUAGE, you could do this, among other things, as follows:

- COMPLEXITY OF THE SENTENCES that a person can form in the language in a test (only main clauses? also compound sentence, also complex sentences?);
- AMOUNT OF TIME in seconds between two errors in conversation;
- NUMBER OF ERRORS PER 100 WORDS in a text that the person writes in 90 minutes.

What is wrong with the following two proposals for operationalization?

- AMOUNT OF ACTIVE VOCABULARY;
- AMOUNT OF PASSIVE VOCABULARY.

THINK BREAK

These proposals are not particularly useful because, while knowing these amounts would certainly be very useful to assess somebody's knowledge of a foreign language, they are not directly observable: it is not clear what exactly you would count or measure. If you in turn operationalized the amount of passive vocabulary on the basis of the number of words a person knows in a vocabulary test (involving, say, words from different frequency bands) or synonym finding test, then you know what to count – but the above is too vague.

From the above it follows that operationalizing involves using numbers to represent states of variables. Such a number may be a measurement (402 ms reaction time, 12 words in a synonym finding test, the direct object is four syllables long), but discrete non-numerical states can also be coded using numbers. Thus, variables are not only distinguished according to their role in the hypothesis – independent vs. dependent – but also according to their level of measurement:

- nominal or categorical variables are variables with the lowest information value. Different values of these variables only reveal that the objects with these different values exhibit different characteristics. Such variables are called *nominal variables* (or *binary variables*) when they can take on only two different levels; such variables are called *categorical variables* when they can take on three or more different levels. In our example of particle placement, the variable DIRECTIONALPP could be coded with 1 for the *ABSENCE* and 2 for *PRESENCE*, but note that the fact that the value for *PRESENCE* is twice as large as that for *ABSENCE* does not mean anything (other than that the values are different) – theoretically, you could code absence with 34.2 and presence with 7.[3] Other

[3] Usually, nominal variables are coded using 0 and 1. There are two reasons for that: (i) a conceptual reason: often, such nominal variables can be understood as the presence of something (= 1) or the absence of something (= 0) or even as a ratio variable (cf. below); i.e., in the example of particle placement, the nominal variable CONCRETENESS could be understood as a ratio variable NUMBER OF CONCRETE REFERENTS; (ii) for rea-

typical examples for nominal or categorical variables are ANIMACY (*ANIMATE* vs. *INANIMATE*), CONCRETENESS (*CONCRETE* vs. *ABSTRACT*), STRESS (*STRESSED* vs. *UNSTRESSED*), AKTIONSART (*ACTIVITY* vs. *ACCOMPLISHMENT* vs. *ACHIEVEMENT* vs. *STATE*) etc.

– *ordinal variables* not only distinguish objects as members of different categories the way that nominal and categorical variables do – they also allow to rank-order the objects in a meaningful way. However, differences between ranks cannot be meaningfully compared. Grades are a typical example: a student with an A (4 grade points) scored a better result than a student with a C (2 grade points), but just because 4 is two times 2, that does not necessarily mean that the A-student did exactly twice as well as the C-student – depending on the grading system, the A-student may have given three times as many correct answers as the C-student. In the particle placement example, the variable COMPLEXITY is an ordinal variable if you operationalize it as above: *SIMPLE NP* (1) vs. *PHRASALLY-MODIFIED* (2) vs. *CLAUSALLY-MODIFIED* (3). It is useful to make the ranks compatible with the variable: if the variable is called SYNTACTIC COMPLEXITY, then large rank numbers should represent large degrees of complexity, i.e., complex direct objects. If, on the other hand, the variable is called SYNTACTIC SIMPLICITY, then large rank numbers should represent large degrees of simplicity, simple direct objects. Other typical examples are SOCIO-ECONOMIC STATUS or DEGREE OF IDIOMATICITY or PERCEIVED VOCABULARY DIFFICULTY (e.g., *LOW*/1 vs. *INTERMEDIATE*/2 vs. *HIGH*/3).
– *ratio variables* not only distinguish objects as members of different categories and with regard to some rank ordering – they also allow to meaningfully compare the differences and ratios between values. For example, LENGTH IN SYLLABLES is such a ratio variable: when one object is six syllables long and another is three syllables long, then the first is of a different length than the second (the nominal information), the first is longer than the second (the ordinal information), and it is exactly twice as long as the second. Other typical examples are annual salaries, reaction times in milliseconds.[4]

These differences can be clearly illustrated in a table. Table 4 is a part of a fictitious data set on lengths and degrees of complexity of subjects and

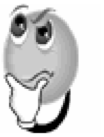
sons I will not discuss here, it is computationally useful to use 0 and 1 and, somewhat counterintuitively, some statistical software even requires that kind of coding.

[4] Strictly speaking, there is also a class of so-called *interval variables*, which we are not going to discuss here separately from ratio variables.

objects – which column contains which kind of variable?

Table 4. A fictitious data set of subjects and objects

DATA POINT	COMPLEXITY	DATA SOURCE	SYLLLENGTH	GRMRELATION
1	*HIGH*	*D8Y*	6	*OBJECT*
2	*HIGH*	*HHV*	8	*SUBJECT*
3	*LOW*	*KB0*	3	*SUBJECT*
4	*INTERMEDIATE*	*KB2*	4	*OBJECT*



THINK BREAK

DATA POINT is a categorical variable: every data point gets its own number so that you can uniquely identify it, but the number as such may represent little more than the order in which the data points were entered. COMPLEXITY is an ordinal variable with three levels. DATA SOURCE is another categorical variable: the levels of this variable are file names from the British National Corpus. SYLLLENGTH is a ratio variable since the third object can correctly be described as half as long as the first. GRMRELATION is a nominal/categorical variable. These distinctions are very important since these levels of measurement determine which statistical tests can and cannot be applied to a particular question and data set.

The issue of operationalization is one of the most important of all. If you do not operationalize your variables properly, then the whole study might be useless since you may actually end up not measuring what you want to measure. Without an appropriate operationalization, the *validity* of your study is at risk. Let us briefly return to an example from above. If we investigated the question of whether subjects in English are longer than direct objects and looked through sentences in a corpus, we might come across the following sentence:

(2) The younger bachelors ate the nice little parrot.

The result for this sentence depends on how LENGTH is operationalized. If LENGTH is operationalized as number of morphemes, then the subject is longer than the direct object: the subject gets the value 5 (*The*, *young*, comparative *-er*, *bachelor*, plural *s*) and the direct object gets the value 4 (*the*, *nice*, *little*, *parrot*). However, if LENGTH is operationalized as number of words, the subject (3 words) is shorter than the direct object (4 words). And, if LENGTH is operationalized as number of characters without spaces, the subject and the direct object are equally long (19 characters). In this case, thus, the operationalization alone determines the result.

3.2.3. Scientific hypotheses in statistical/mathematical form

Once you have formulated both your own (alternative) hypothesis and the logically complementary (null) hypothesis and have defined how the variables will be operationalized, you also formulate two statistical versions of these hypotheses. That is, you first formulate the two text hypotheses, and in the statistical hypotheses you then express the numerical results you expect on the basis of the text hypotheses.

Statistical hypotheses usually involve one of three different mathematical forms: there are hypotheses about *frequencies or frequency differences*, hypotheses about *means or differences of means*, and hypotheses about *correlations*. (Rather infrequently, we will also encounter hypotheses about dispersions and distributions.) We begin by looking at a simple example regarding particle placement: if a verb-particle construction is followed by a directional PP, then native speakers will produce the constituent order *VOP* more often than when no such directional PP follows. To formulate the statistical hypothesis counterpart to this text form, you have to answer the question, if I investigated, say, 200 sentences with verb-particle constructions in them, how would I know whether this hypothesis is correct or not? (As a matter of fact, you actually have to proceed a little differently, but we will get to that later.) One possibility of course is to count how often CONSTRUCTION: *VPO* and CONSTRUCTION: *VOP* are followed by a directional PP, and if there are more directional PPs after CONSTRUCTION: *VOP* than after CONSTRUCTION: *VPO*, then this provides support to the alternative hypothesis. Thus, this possibility involves frequencies and the statistical hypotheses are:

$H_{1\ \text{directional}}$: $n_{\text{dir. PPs after CONSTRUCTION: }\textit{VPO}} < n_{\text{dir. PPs after CONSTRUCTION: }\textit{VOP}}$

$H_{1\ \text{non-directional}}$: $n_{\text{dir. PPs after CONSTRUCTION: }\textit{VPO}} \neq n_{\text{dir. PPs after CONSTRUCTION: }\textit{VOP}}$

H_0: $n_{\text{dir. PPs after CONSTRUCTION: }\textit{VPO}} = n_{\text{dir. PPs after CONSTRUCTION: }\textit{VOP}}$[5]

[5] Note: I said above that you often obtain the null hypothesis by inserting *not* into the alternative hypothesis. Thus, when the statistical version of the alternative hypothesis involves a "<", then you might expect the statistical version of the null hypothesis to contain a "≥". However, we will follow the usual convention also mentioned above that

Just in passing: what do these statistical hypotheses presuppose?

THINK BREAK

They presuppose that you investigate equally many instances of both constructions because otherwise a small observed frequency of directional PPs after CONSTRUCTION: *VOP* – the frequency we expect to be large – could simply be due to a small overall frequency of CONSTRUCTION: *VOP*.

For the variable COMPLEXITY, you could formulate similar hypotheses based on frequencies, if COMPLEXITY is operationalized on the basis of, for example, the three levels mentioned above.

Let us now turn to an example involving statistical hypotheses based on means: if the direct object of a transitive phrasal verb is long, then native speakers will produce the constituent order *VPO* more often than when the direct object is short. One probably obvious way to proceed is to measure the average lengths of direct objects in CONSTRUCTION: *VPO* and CONSTRUCTION: *VOP* and then compare these average lengths to each other. You could therefore write:

$H_{1\ \text{directional}}$: $mean_{\text{Length of the direct object in CONSTRUCTION: } vpo} >$
$mean_{\text{Length of the direct object in CONSTRUCTION: } vop}$

$H_{1\ \text{non- directional}}$: $mean_{\text{Length of the direct object in CONSTRUCTION: } vpo} \neq$
$mean_{\text{Length of the direct object in CONSTRUCTION: } vop}$

H_0: $mean_{\text{Length of the direct object in CONSTRUCTION: } vpo} =$
$mean_{\text{Length of the direct object in CONSTRUCTION: } vop}$

With similarly obvious operationalizations, the other text hypotheses from above can be transformed into analogous statistical hypotheses. Now, and only now, we finally know what needs to be observed in order for us to reject the null hypothesis. (We will look at hypotheses involving correlations, dispersion, and distributions later.)

All hypotheses discussed so far were concerned with the simple case where a sample of verb-particle constructions was investigated regarding whether the two constructions differ with regard to one independent varia-

a null hypothesis states the absence of a difference/effect/correlation etc., which is why we write “=“. You will see below that the cases covered by “≥” will still be invoked in the computations that are based on these statistical hypotheses.

ble (e.g., DIRECTIONALPP: *PRESENT* vs. *ABSENT*). The statistical methods to handle such cases are the subject of Chapter 4. However, things are often not that simple: most phenomena are multifactorial in nature, which means dependent variables are usually influenced by, or at least related to, more than one independent variable. There are basically two different ways in which several independent and dependent variables may be related, which we will explore on the basis of the example involving constituent lengths.

Let us assume you wished to study whether the lengths of constituents – captured in the dependent variable LENGTH – are dependent on two independent variables, the variable GRMRELATION (with the two levels *SUBJECT* and *OBJECT*) and the variable CLAUSE TYPE (with the two levels *MAIN* and *SUBORDINATE*). Let us further assume you did a small a pilot study in which you investigated 120 constituents that are distributed as shown in Table 5. Let us finally assume you determined the syllabic lengths of all 120 constituents to compute the means for the four variable level combinations – subjects in main clauses, subjects in subordinate clauses, objects in main clauses, and objects in subordinate clauses – and obtained the following results:

- on average, subjects are shorter than direct objects;
- on average, the subjects and objects in main clauses are shorter than the subjects and objects in subordinate clauses.

Table 5. A fictitious data set of subjects and objects

	GRMRELATION: *SUBJECT*	GRMRELATION: *OBJECT*	Totals
CLAUSETYPE: *MAIN*	30	30	60
CLAUSETYPE: *SUBORD*	30	30	60
Totals	60	60	120

The interesting thing is that these results can come in different forms. On the one hand, the effects of the two independent variables can be *additive*. That means the combination of the two variables has the effect that you would expect on the basis of each variable's individual effect. Since subjects are short, as are constituents in main clauses, according to an additive effect subjects in main clauses should be the shortest constituents, and objects in subordinate clauses should be longest. This result, which is the one that a null hypothesis would predict, is represented in Figure 3.

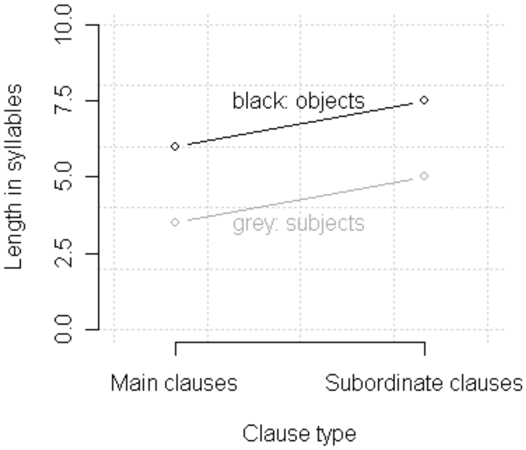


Figure 3. Interaction plot for GRMRELATION × CLAUSE TYPE 1

However, it is also possible that the two independent variables *interact*. Two or more variables interact if their joint effect on the dependent variable is not predictable from their individual effects on the same dependent variable. In our case: when objects are longer than subjects and constituents in subordinate clauses longer than constituents in main clauses, but at the same time we find that objects in subordinate clauses are fairly short, then we have an interaction. This scenario is represented in Figure 4.

As before, subjects are on average shorter than objects. As before, main clause constituents are shorter on average than subordinate clause constituents. But the combination of these variable levels does not have the expected additive effect: objects in general are longer than subjects – but objects in subordinate clauses are not. This kind of interaction can often be recognized easily in a plot because the lines connecting means intersect or at least have slopes with different signs.

Yet another kind of interaction is shown in Figure 5. At first sight, this does not look like an interaction: the means of all individual variable levels are as before and even the null hypothesis expectation that objects in subordinate clauses are the longest constituents is borne out. Why is this still an interaction?

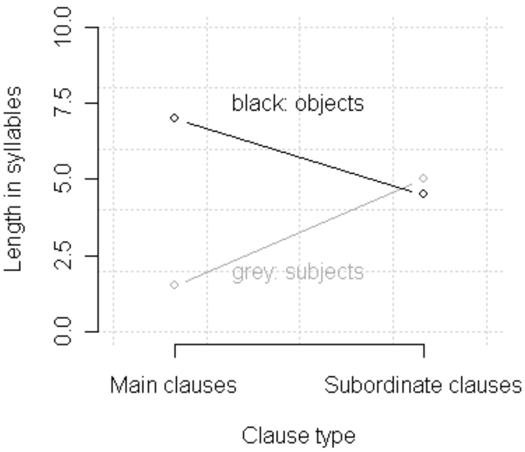


Figure 4. Interaction plot for GRMRELATION × CLAUSE TYPE 2

THINK BREAK

This is an interaction because even though the lines do not intersect and both have a positive slope, the slope of the line for objects is still much higher than that for subjects. Put differently, while the difference between main clause subjects and main clause objects is only about one syllable, that between subordinate clause subjects and subordinate clause objects is approximately four syllables. This unexpected difference is the reason why this scenario is also considered an interaction.

Thus, if you have more than two independent variables, you often need to consider interactions for both the formulation of hypotheses and the subsequent evaluation. Such issues are often the most interesting but also the most complex. We will look at some such methods in Chapter 5.

One important recommendation following from this is that, when you read published results, you should always consider whether other independent variables may have contributed to the results. To appreciate how important this kind of thinking can be, let us look at a non-linguistic example.

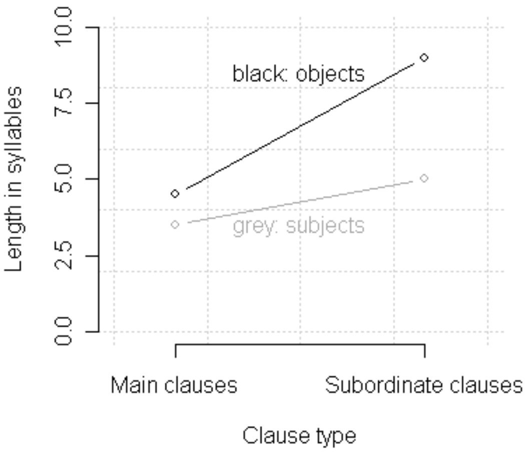


Figure 5. Interaction plot for GRMRELATION × CLAUSE TYPE 3

The 11/29/2007 issue of the Santa Barbara Independent discussed a study according to which illegal Mexican immigrants use Californian emergency rooms only half as much per capita as legal Mexican immigrants. This finding was used to support the position that illegal Mexican immigrants do not put a strain on the US American health system. In response to this article, a letter to the editor of 12/6/2007, however, points out that said illegal Mexican immigrants are on average much younger than legal Mexican immigrants: of the illegal Mexican immigrants, only 5% are older than 50, but of the legal Mexican immigrants 37% are. Since younger people are in general less likely to need medical assistance, the letter writer argues that the fact that illegal Mexican immigrants require less medical assistance does not prove that they do not overuse the medical system – they are just too young to require as much medical help. You should recognize what happened here. Technically speaking, the letter writer introduced an additional independent variable, AGE, to the analysis of the dependent variable NUMBER OF ER PATIENTS in terms of the independent variable IMMIGRATION STATUS: *LEGAL* vs. *ILLEGAL*. In this particular case, it is obvious that younger people need less medical care than older people – so AGE appears to play some role. Note, however, that contrary to what the letter to the editor appears to assume, it is still far from obvious that that is also relevant for a discussion of ER patients since one might just as well argue that the larger amount of medical assistance required by older people is unlikely to consist mainly of ER care, but that is a different story.

While we are not going to solve the issue here, it should have become obvious that considering more than one variable or more variables than are mentioned in some context can be interesting and revealing. However, this does not mean that you should always try to add as many additional variables as possible. An important principle that limits the number of additional variables to be included is called *Occam's razor*. This rule ("entia non sunt multiplicanda praeter necessitatem") states that additional variables should only be included when it's worth it, i.e., when they increase the explanatory and predictive power of the analysis substantially. How exactly that decision is made will be explained especially in Chapter 5.

Recommendation(s) for further study

- Good and Hardin (2006: 154f.) for another example to 'think multifactorially'
- <http://en.wikipedia.org/wiki/Occam's_Razor> and Crawley (2007: 325) on *Occam's razor*

3.3. Data collection and storage

Only after all variables have been operationalized and all hypotheses have been formulated do you actually collect your data. For example, you run an experiment or do a corpus study or … However, you will hardly ever study the whole population of events but a sample so it is important that you choose your sample such that it is representative and balanced with respect to the population to which you wish to generalize. Here, I call a sample *representative* when the different parts of the population are reflected in the sample, and I call a sample *balanced* when the sizes of the parts in the population are reflected in the sample. Imagine, for example, you want to study the frequencies and the uses of the discourse marker *like* in the speech of Californian adolescents. To that end, you want to compile a corpus of Californian adolescents' speech by asking some Californian adolescents to record their conversations. In order to obtain a sample that is representative and balanced for the population of all the conversations of Californian adolescents, the proportions of the different kinds of conversations in which the subjects engage would ideally be approximately reflected in the sample. For example, a good sample would not just include the conversations of the subjects with members of their peer group(s), but also conversations with their parents, teachers, etc., and if possible, the proportions that all these different kinds of conversations make up in the sample would correspond to their proportions in real life, i.e. the population.

While it is important you try to stick to these rules as much as possible, why are they often more of a theoretical ideal?

THINK BREAK

This is often just a theoretical ideal because we don’t know all parts and their proportions in the population. Who would dare say how much of an average Californian adolescent’s discourse – and what is an average Californian adolescent? – takes place within his peer group, with his parents, with his teachers etc.? And how would we measure the proportion – in words? sentences? minutes? Still, even though these considerations will often only result in estimates, you must think about the composition of your sample(s) just as much as you think about the exact operationalization of your variables. If you do not do that, then the whole study may well fail because you may be unable to generalize from whatever you find in your sample to the population. One important rule in this connection is to choose the elements that enter into your sample randomly, to randomize. For example, if the adolescents who participate in your study receive a small recording device with a lamp and are instructed to always record their conversations when the lamp lights up, then you could perhaps send a signal to the device at random time intervals (as determined by a computer). This would make it more likely that you get a less biased sample of many different kinds of conversational interaction, which would then reflect the population better.

Let us briefly look at a similar example from the domain of first language acquisition. It was found that the number of questions in recordings of caretaker-child interactions was surprisingly high. Some researchers suspected that the reason for that was parents’ (conscious or unconscious) desire to present their child as very intelligent so that they asked the child “And what is that?” questions all the time so that the child could show how many different words he knows. Some researchers then changed their sampling method such that the recording device was always in the room, but the parents did not know exactly when it would record caretaker-child interaction. The results showed that the proportion of questions decreased considerably …

In corpus-based studies you will often find a different kind of randomization. For example, you will find that a researcher first retrieved all instances of the word he is interested in and then sorted all instances according to random numbers. When the researcher then investigates the first 20% of the list, he has a random sample. However you do it, randomization is one of the most important principles of data collection.

Once you have collected your data, you have to store them in a format that makes them easy to annotate, manipulate, and evaluate. I often see people – students as well as seasoned researchers – print out long lists of data points, which are then annotated by hand, or people annotate concordance lines from a corpus in a text processing software. This may seem reasonable for small data sets, but it doesn't work or is extremely inconvenient for larger ones, and the generally better way of handling the data is in a spreadsheet software (e.g., OpenOffice.org Calc) or a database, or in R. However, there is a set of ground rules that needs to be borne in mind.

First, the first row contains the names of all variables. Second, each of the other rows represents one and only one data point. Third, the first column just numbers all *n* cases from 1 to *n* so that every row can be uniquely identified and so that you always restore one particular ordering (e.g., the original one). Fourth, each of the remaining columns represents one and only one variable with respect to which every data point gets annotated. In a spreadsheet for a corpus study, for example, one additional column may contain the name of the corpus file in which the word in question is found; another column may provide the line of the file in which the word was found. In a spreadsheet for an experimental study, one column should contain the name of the subject or at least a unique number identifying the subject; other columns may contain the age of the subject, the sex of the subject, the exact stimulus or some index representing the stimulus the subject was presented with, the order index of a stimulus presented to a subject (so that you can test whether a subject's performance changes systematically in the course of the experiment), …

To make sure these points are perfectly clear, let us look at two examples. Let's assume for your study of particle placement you had looked at a few sentences and counted the number of syllables of the direct objects. First, a question: in this design, what is the dependent variable and what is the independent variable?

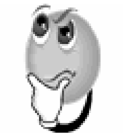
THINK BREAK

The independent variable is the ratio variable Length (in syllables), which can take on all sorts of positive integer values. The dependent variable is the nominal variable Construction, which can be either *vpo* or *vop*. When all hypotheses were formulated and, subsequently, data were collected and coded, then I sometimes see a format such as the one represented in Table 6.

Table 6. A not-so-good table 1

	Length: *2*	Length: *3*	Length: *5*	Length: *6*
Construction: *vpo*	||	||	|||	||
Construction: *vop*	||||	|||	||	|

As a second example, let's look at the hypothesis that subjects and direct objects are differently long (in words). Again the question: what is the dependent variable and what is the independent variable?



THINK BREAK

The independent variable is the nominal variable Relation, which can be either *subject* or *object*. The dependent variable is Length, which can take on positive integer values. If you formulated all four hypotheses (alternative hypothesis: text form and statistical form; null hypothesis: text form and statistical form) and then looked at the small corpus in (3),

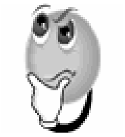
(3) a. The younger bachelors ate the nice little cat.
b. He was locking the door.
c. The quick brown fox hit the lazy dog.

then your spreadsheet should *not* look like Table 7.

Table 7. A not-so-good table 2

Sentence	Subj	Obj
The younger bachelors ate the nice little cat.	3	4
He was locking the door.	1	2
The quick brown fox hit the lazy dog.	4	3

Both Table 6 and Table 7 violate all of the above rules. In Table 7, for example, every row represents two data points, not just one, namely one data point representing some subject’s length and one representing the length of the object from the same sentence. Also, not every variable is represented by one and only column – rather, Table 7 has two columns with data points, each of which represents one level of an independent variable, not one variable. Before you read on, how would you have to reorganize Table 7 to make it compatible with the above rules?



THINK BREAK

This is a much better way to store the data.

Table 8. A much better coding of the data in Table 7

Case	Sent#	Sentence	Relation	Length
1	1	The younger bachelors ate the nice little cat.	subj	3
2	1	The younger bachelors ate the nice little cat.	obj	4
3	2	He was locking the door.	subj	1
4	2	He was locking the door.	obj	2
5	3	The quick brown fox hit the lazy dog.	subj	4
6	3	The quick brown fox hit the lazy dog.	obj	3

In this version, every data point has its own row and is characterized according to the two variables in their respective columns. An even more comprehensive version may now even include one column containing just the subjects and objects so that particular cases can be found more easily. In the first row of such a column, you would find *The younger bachelor*, in the second row of the same column, you would find *the nice little cat* etc. The same logic applies to the improved version of Table 6, which should look like Table 9.

With very few exceptions, this is the format in which you should always save your data.[6] Ideally, you enter the data in this format into a spreadsheet

[6] There are some more complex statistical techniques which can require different formats, but in the vast majority of cases, the standard format discussed above is the one that you

software and save the data (i) in the native file format of that application (to preserve colors and other formattings you may have added) and (ii) into a tab-delimited text file, which is often smaller and easier to import into other applications (such as R).

One important aspect to note is that data sets are often not complete. Sometimes, you can't annotate a particular corpus line, or a subject does not provide a response to a stimulus. Such 'data points' are not simply omitted, but are entered into the spreadsheet as so-called missing data with the code "NA" in order to (i) preserve the formal integrity of the data set (i.e., have all rows and columns contain the same number of elements) and (ii) be able to do follow-up studies on the missing data to see whether there is a pattern in the data points which needs to be accounted for.

Table 9. A much better coding of the data in Table 6

CASE	CONSTRUCTION	LENGTH
1	*VPO*	2
2	*VPO*	2
3	*VOP*	2
4	*VOP*	2
5	*VOP*	2
6	*VOP*	2
7	*VPO*	3
8	*VPO*	3
9	*VOP*	3
10	*VOP*	3
11	*VOP*	3
…	…	…

All these steps having to do with the data collection must be described in the methods part of your written version: what is the population to which you wanted to generalize, how did you draw your (ideally) representative and balanced sample, which variables did you collect data for, etc.

3.4. The decision

When the data have been stored in a format that corresponds to that of Table 8/Table 9, you can finally do what you wanted to do all along: evaluate

will need and that will allow you to easily switch to another format. Also, for reasons that will only become obvious much later in Chapter 5, I myself always use capital letters for variables and small letters for their levels.

the data. As a result of that evaluation you will obtain frequencies, means, or correlation coefficients. However, one central aspect of this evaluation is that you actually do not simply try to show that your alternative hypothesis is correct – contrary to what you might expect you try to show that the statistical version of the null hypothesis is wrong, and since the null hypothesis is the logical counterpart to the alternative hypothesis, this supports your own alternative hypothesis. The obvious question now is, why this 'detour'? The answer to this question can be approached by considering the following two questions:

- how many subjects and objects do you maximally have to study to show that the alternative hypothesis "subjects are shorter than direct objects" is correct?
- how many subjects and objects do you minimally have to study to show that the null hypothesis "subjects are as long as direct objects" is incorrect?

THINK BREAK

You probably figured out quickly that the answer to the first question is "infinitely many." Strictly speaking, you can only be sure that the alternative hypothesis is correct if you have studied all subjects and direct objects and found not a single counterexample. The answer to the second question is "one each" because if the first subject is longer or shorter than the first object, we know that, strictly speaking, the null hypothesis is not correct. However, especially in the humanities and in the social sciences you do not usually reject a hypothesis on the basis of just one counterexample. Rather, you evaluate the data in your sample and then determine whether your null hypothesis H_0 and the empirical result are sufficiently incompatible to reject H_0 and thus accept H_1. More specifically, you assume the null hypothesis is true and compute the probability *p* that you would get your observed result or all other results that deviate from the null hypothesis even more strongly. When that probability *p* is equal to or larger than 5%, then you stick to the null hypothesis because, so to speak, the result is still too compatible with the null hypothesis to reject it and accept the alternative hypothesis. If, however, that probability *p* is smaller than 5%, then you can reject the null hypothesis and adopt the alternative hypothesis.

For example, if in your sample subjects and direct objects are on average 4.2 and 5.6 syllables long, then you compute the probability *p* to find this difference of 1.4 syllables or an even larger difference when you in fact don't expect any such difference (because that is what the null hypothesis predicts). Then, there are two possibilities:

- if this probability *p* is smaller than 5% (or 0.05, as this is usually written), then you can reject the null hypothesis that there is no difference between subjects and direct objects in the population. In the results section of your paper, you can then write that you found a significant difference between the means in your sample, and in the discussion section of your paper you would discuss what kinds of implications this has, etc.
- if this probability *p* is equal to or larger than 5%, then you cannot reject the null hypothesis that there is no difference between subjects and direct objects in the population. In the results section of your paper, you must then admit that you have not found a significant difference between the means in your sample. In the discussion part of your paper, you should then discuss the implications of this finding as well as speculate or reason about why there was no significant difference – there may have been outliers in the corpus data or in the experiment (because subjects reacted strangely to particular stimuli, coding errors, etc. (*Outliers* are values in the sample that are rather untypical given the rest of the sample.)

Two aspects of this logic are very important: First, the fact that an effect is significant does not necessarily mean that it is an important effect despite what the everyday meaning of *significant* might suggest. The word *significant* is used in a technical sense here, meaning the effect is large enough for us to assume that, given the size of the sample(s), it is probably not random. Second, just because you accept an alternative hypothesis given a significant result, that does not mean that you have *proven* the alternative hypothesis. This is because there is still the probability *p* that the observed result has come about although the null hypothesis is correct – *p* is small enough to warrant accepting the alternative hypothesis, but not to prove it.

This line of reasoning may appear a bit confusing at first especially since we suddenly talk about two different probabilities. One is the probability of 5% (to which the other probability is compared), that other probability is the probability to obtain the observed result when the null hypothesis is correct.

Warning/advice

You must never change your hypotheses *after* you have obtained your results and then sell your study as successful support of the 'new' alternative hypothesis. Also, you must never explore a data set – the nicer way to say 'fish for something useable' – and, when you then find something significant, sell this result as a successful test of a previously formulated alternative hypothesis. You may of course explore a data set in search of patterns and hypotheses, but if a data set generates a hypothesis, you must test that hypothesis on the basis of different data.

The former is the so-called *level of significance*, which defines a threshold level. If that threshold is exceeded, you must stick to the null hypothesis. This value is *defined before data are obtained* (and usually set to 5%). The latter probability is the so-called *p*-value or probability of error and is *computed on the basis of the data*. Why is this probability called probability of error? It is because it is the probability to err when you accept the alternative hypothesis given the observed data. Sometimes, you will find that people use different wordings for different *p*-values:

- $p < 0.001$ is sometimes referred to as *highly significant* and indicated with ***;
- $0.001 \leq p < 0.01$ is sometimes referred to as *very significant* and indicated with **;
- $0.01 \leq p < 0.05$ is sometimes referred to as *significant* and indicated with *;
- $0.05 \leq p < 0.1$ is sometimes referred to as *marginally significant* and indicated with *ms* or a period but since such *p*-values are larger than the usual standard of 5%, calling such results marginally significant amounts, polemically speaking at least, to saying "Look, I didn't really get the significant results I was hoping for, but they are still pretty nice, don't you agree?", which is why I typically discourage the use of this expression.

But while we have seen above how this comparison of the two probabilities contributes to the decision in favor of or against the alternative hypothesis, it is still unclear how this *p*-value is computed.

3.4.1. Overview: discrete probability distributions

Let’s assume you and I decided to toss a coin 100 times. If we get heads, I get one dollar from you – if we get tails, you get one dollar from me. Before this game, you formulate the following hypotheses:

Text H_0: Stefan does not cheat: the probability for heads and tails is 50% vs. 50%.
Text H_1: Stefan cheats: the probability for heads is larger than 50%.

This scenario can be easily operationalized using frequencies:

Statistical H_0: Stefan will win just as often as I will, namely 50 times.
Statistical H_1: Stefan will win more often than I, namely more than 50 times.

Now my question: when we play the game and toss the coin 100 times, after which result will you suspect that I cheated?

- when you lost 51 times (probably not …)?
- when you lost 55 times ? when you lost 60 times? (maybe …)?
- when you lost 80 times or even more often? (most likely …)?

THINK BREAK

Maybe without realizing it, you are currently doing some kind of significance test. Let’s assume you lost 60 times. Since the expectation from the null hypothesis was that you would lose only 50 times, you lost more often than you thought you would. Let’s finally assume that, given the above explanation, you decide to only accuse me of cheating when the probability *p* to lose 60 or even more times in 100 tosses is smaller than 5%. Why “or even more times”? Well, above we said

> you expect the null hypothesis to be true and compute the probability *p* to get the observed result or all other results that deviate from the null hypothesis even more strongly.

Thus, you must ask yourself how and how much does the observed result deviate from the result expected from the null hypothesis. Obviously, the number of losses is larger: 60 > 50. Thus, the results that deviate from the null hypothesis that much or even more in the predicted direction are those where you lose 60 times or more often: 60 times, 61 times, 62, times, …, 99 times, and 100 times. In a more technical parlance, you set the significance level to 5% (0.05) and ask yourself "how likely is it that Stefan did not cheat but still won 60 times although he should only have won 50 times?" This is exactly the logic of significance testing.

It is possible to show that the probability *p* to lose 60 times or more just by chance – i.e., without me cheating – is 0.02844397, i.e., 2.8%. Since this *p*-value is smaller than 0.05 (or 5%), you can now accuse me of cheating. If we had been good friends, however, so that you would not have wanted to risk our friendship by accusing me of cheating prematurely and had set the significance level to 1%, then you would *not* be able to accuse me of cheating, since 0.02844397 > 0.01.

This example has hopefully clarified the overall logic even further, but what is probably still unclear is how this *p*-value is computed. To illustrate that, let us reduce the example from 100 coin tosses to the more manageable amount of three coin tosses. In Table 10, you find all possible results of three coin tosses and their probabilities provided that the null hypothesis is correct and the chance for heads/tails on every toss is 50%.

Table 10. All possible results of three coin tosses and their probabilities (when H_0 is correct)

Toss 1	Toss 2	Toss 3	# heads	# tails	p_{result}
heads	heads	heads	3	0	0.125
heads	heads	tails	2	1	0.125
heads	tails	heads	2	1	0.125
heads	tails	tails	1	2	0.125
tails	heads	heads	2	1	0.125
tails	heads	tails	1	2	0.125
tails	tails	heads	1	2	0.125
tails	tails	tails	0	3	0.125

More specifically, the three left columns represent the possible results, column 4 and column 5 show how many heads and tails are obtained in each of the eight possible results, and the rightmost column lists the probability of each possible result. As you can see, these are all the same, 0.125. Why is that so?

Two easy ways to explain this are conceivable, and both of them require you to understand the crucial concept of *independence*. The first one involves understanding that the probability of heads and tails is the same on every trial and that all trials are independent of each other. This notion of independence is a very important one: trials are independent of each other when the outcome of one trial (here, one toss) does not influence the outcome of any other trial (i.e., any other toss). Similarly, samples are independent of each other when there is no meaningful way in which you can match values from one sample onto values from another sample. For example, if you randomly sample 100 transitive clauses out of a corpus and count their subjects' lengths in syllables, and then you randomly sample 100 *different* transitive clauses from the same corpus and count their direct objects' lengths in syllables, then the two samples – the 100 subject lengths and the 100 object lengths – are independent. If, on the other hand, you randomly sample 100 transitive clauses out of a corpus and count the lengths of the subjects and the objects in syllables, then the two samples – the 100 subject lengths and the 100 object lengths – are dependent because you can match up the 100 subject lengths onto the 100 object lengths perfectly by aligning each subject with the object from the very same clause. Similarly, if you perform an experiment twice with the same subjects, then the two samples made up by the first and the second experimental results are dependent, because you match up each subject's data point in the first experiment with the same subject's data point in the second experiment. This distinction will become very important later on.

Returning to the three coin tosses: since there are eight different outcomes of three tosses that are all independent of each other and, thus, equally probable, the probability of each of the eight outcomes is $^1/_8$ = 0.125.

The second way to understand Table 10 involves computing the probability of each of the eight events separately. For the first row that means the following: the probability to get head in the first toss, in the second, in the third toss is always 0.5. Since the tosses are independent of each other, you obtain the probability to get heads three times in a row by multiplying the individual events' probabilities: $0.5 \cdot 0.5 \cdot 0.5 = 0.125$ (the multiplication rule in probability theory). Analogous computations for every row show that the probability of each result is 0.125. According to the same logic, we can show the null hypothesis predicts that each of us should win 1.5 times, which is a number that has only academic value since you cannot win half a time.

Now imagine you lost two out of three times. If you had again set the level of significance to 5%, could you accuse me of cheating?

THINK BREAK

No way. Let me first ask again which events need to be considered. You need to consider the observed result – that you lost two times – and you need to consider the result(s) that deviate(s) even more from the null hypothesis and the observed result in the predicted direction. This is easy here: the only such result is that you lose all three times. Let us compute the sum of the probabilities of these events.

As you can see in column 4, there are three results in which you lose two times in three tosses: H H T (row 2), H T H (row 3), and T H H (row 5). Thus, the probability to lose exactly two times is 0.125+0.125+0.125 = 0.375, and that is already much much more than your level of significance 0.05 allows. However, to that you still have to add the probability of the event that deviates even more from the null hypothesis, which is another 0.125. If you add this all up, the probability *p* to lose two or more times in three tosses when the null hypothesis is true is 0.5. This is ten times as much as the level of significance so there is no way that you can accuse me of cheating.

This logic can also be represented graphically very well. In Figure 6, the summed probabilities for all possible numbers of heads are represented as bars. The bars for two heads – the observed result – and for three heads – the even more extreme deviation from the null hypothesis in this direction – are shown in black, and their lengths indicate the probabilities of these outcomes. Visually speaking, you move from the expectation of the null hypothesis away to the observed result (at $x = 2$) and add the length of that bar to the lengths of all bars you encounter if you move on from there in the same direction, which here is only one bar at $x = 3$.

This actually also shows that you, given your level of significance, cannot even accuse me of cheating when you lose all three times because this result already comes about with a probability of $p = 0.125$, as you can see from the length of the rightmost bar.

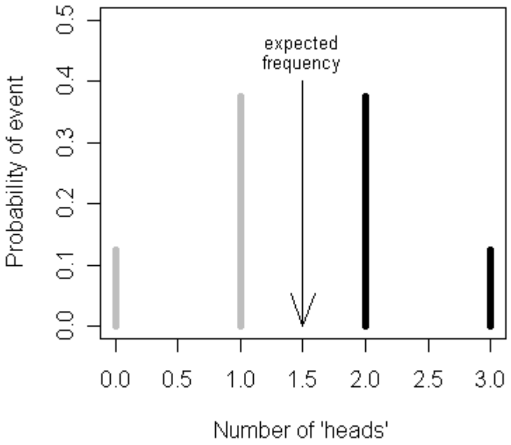


Figure 6. All possible results of three coin tosses and their probabilities (when H_0 is correct, one-tailed)

The final important aspect that needs to mentioned now involves the kind of alternative hypothesis. So far, we have always been concerned with *directional alternative hypotheses*: the alternative hypothesis was “Stefan cheats: the probability for heads is larger than 50% [and not just different from 50%].” The kind of significance test we discussed are correspondingly called *one-tailed tests* because we are only interested in one direction in which the observed result deviates from the expected result. Again visually speaking, when you summed up the bar lengths in Figure 6 you only moved from the null hypothesis expectation in one direction. This is important because the decision for or against the alternative hypothesis is based on the cumulative lengths of the bars of the observed result and the more extreme ones in that direction.

However, often you only have a *non-directional alternative hypothesis*. In such cases, you have to look at both ways in which results may deviate from the expected result. Let us return to the scenario where you and I toss a coin three times, but this time we also have an impartial observer who has no reason to suspect cheating on either part. He therefore formulates the following hypotheses (with a significance level of 0.05):

Statistical H_0: Stefan will win just as often as the other player, namely 50 times (or “Both players will win equally often”).

Statistical H_1: Stefan will win more or less often than the other player (or “The players will not win equally often”).

Imagine now you lost three times. The observer now asks himself whether one of us should be accused of cheating. As before, he needs to determine which events to consider. First, he has to consider the observed result that *you* lost three times, which arises with a probability of 0.125. But then he also has to consider the probabilities of other events that deviate from the null hypothesis just as much or even more. With a directional alternative hypothesis, you moved from the null hypothesis only in one direction – but this time there is no directional hypothesis so the observer must also look for deviations just as large or even larger in the other direction of the null hypothesis expectation. For that reason – both tails of the distribution in Figure 6 must be observed – such tests are considered *two-tailed tests*. As you can see in Table 10 or Figure 6, there is another deviation from the null hypothesis that is just as extreme, namely that *I* lose three times. Since the observer only has a non-directional hypothesis, he has to include the probability of that event, too, arriving at a cumulative probability of 0.125+0.125 = 0.25. This logic is graphically represented in Figure 7.

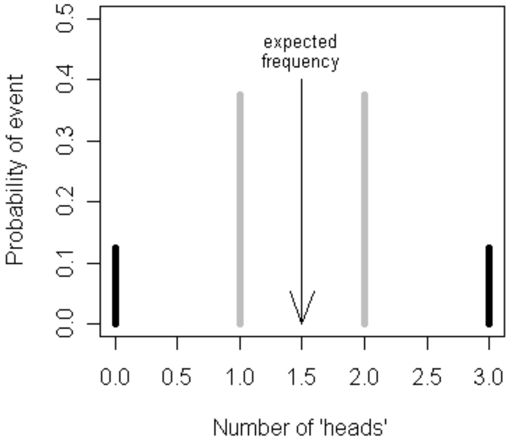


Figure 7. All possible results of three coin tosses and their probabilities (when H_0 is correct, two-tailed)

Note that when you tested your directional alternative hypothesis, you looked at the result 'you lost three times', but when the impartial observer tested his non-directional alternative hypothesis, he looked at the result 'somebody lost three times.' This has one very important consequence: when you have prior knowledge about a phenomenon that allows you to formulate a directional, and not just a non-directional, alternative hypothesis, then the result you need for a significant finding can be less extreme than if you only have a non-directional alternative hypothesis. In most cases, it will be like here: the *p*-value you get for a result with a directional alternative hypothesis is half of the *p*-value you get for a result with a non-directional alternative hypothesis. Prior knowledge is rewarded, which will be illustrated once more now.

With statistical software such as R these kinds of computations can easily be done for more than three tosses, which is why we can now return to the example game involving 100 tosses. Again, we look at the situation through your eyes (directional alternative hypothesis) and through those of an impartial observer (non-directional alternative hypothesis), but this time you and the observer try to determine *before the game* which results are so extreme that one will be allowed to adopt the alternative hypothesis. We begin with your perspective: In Figure 8, you find the by now familiar graph for 100 tosses with the expected frequency for heads of 50. (The meaning of the black lines will be explained presently.)

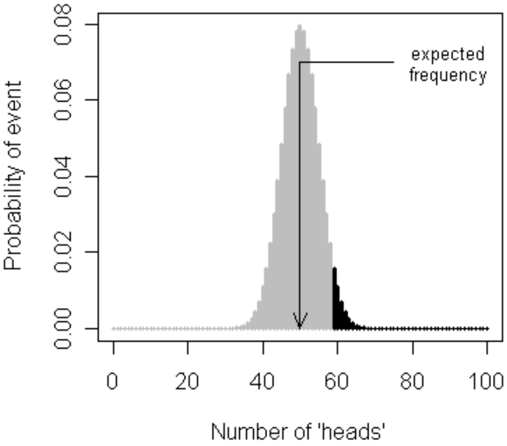


Figure 8. All possible results of 100 coin tosses and their probabilities (when H_0 is correct, one-tailed H_1)

Above, we had an empirical result whose *p*-value we were interested in, and in order to get that *p*-value, we moved from the expected null hypothesis results to the extreme values. Now we want to determine, but not exceed, a *p*-value before we have results and have to proceed the other way round: from an extreme point to the null hypothesis expectation. For example, to determine how many times you can lose without getting a cumulative probability exceeding 0.05, you begin at the most extreme result on the right – that you lose 100 times – and begin to add the lengths of the bars.

(Of course, you would compute that and not literally measure lengths.) The probability that you lose all 100 tosses is $7.8886 \cdot 10^{-31}$. To that you add the probability that you lose 99 out of 100 times, the probability that you lose 98 out of 100 times, etc. When you have added all probabilities until 59 times heads, then the sum of all these probabilities reaches 0.0443; all these are represented in black in Figure 8. Since the probability to get 58 heads out of 100 tosses amounts to 0.0223, you cannot add this event's probability to the others anymore without exceeding the level of significance value of 0.05. Put differently, if you don't want to cut off more than 5% of the summed bar lengths, then you must stop adding probabilities at $x = 59$. You conclude: if Stefan wins 59 times or more often, then I will accuse him of cheating, because the probability of that happening is the largest one that is still smaller than 0.05.

Now consider the perspective of the observer in Figure 9, which is very similar, but not completely identical to Figure 8. The observer also begins with the most extreme result, that I get heads every time: $p_{100 \text{ heads}} \approx 7{,}8886 \cdot 10^{-31}$. But since the observer only has a non-directional alternative hypothesis, he must also include the probability of the opposite, equally extreme result that you get tails all the time. For each additional number of heads – 99, 98, etc. – the observer must now also add the corresponding opposite results – 1, 2, etc. Once the observer has added the probabilities 61 times heads / 39 times tails and 39 times heads / 61 times tails, then the cumulative sum of the probabilities reaches 0.0352 (cf. the black bars in Figure 9). Since the joint probability for the next two events – 60 heads / 40 tails and 40 heads / 60 tails – is 0.0217, the observer cannot add any further results without exceeding the level of significance of 0.05. Put differently, if the observer doesn't want to cut off more than 5% of the summed bar lengths on both sides, then he must stop adding probabilities by going from right to the left at $x = 61$ and stop going from the left to right at $x = 39$. He concludes: if Stefan or his opponent wins 61 times or more often, then someone is cheating (most likely the person who wins more often).

Again, observe that in the same situation the person with the directional alternative hypothesis needs a less extreme result to be able to accept it than the person with a non-directional alternative hypothesis: with the same level of significance, you can already accuse me of cheating when you lose 59 times (only 9 times more often than the expected result) – the impartial observer needs to see someone lose 61 times (11 times more often than the expected result) before he can start accusing someone. Put differently, if you lose 60 times, you can accuse me of cheating, but the observer cannot. This difference is very important and we will use it often.

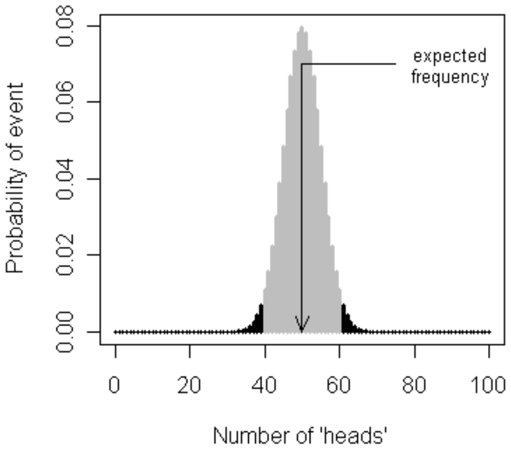


Figure 9. All possible results of 100 coin tosses and their probabilities (when H_0 is correct, two-tailed H_1)

While reading the last few pages, you probably sometimes wondered where the probabilities of events come from: How do we know that the probability to get heads 100 times in 100 tosses is $7.8886 \cdot 10^{-31}$? These values were computed with R on the basis of the so called binomial distribution. You can easily compute he probability that one out of two possible events occurs *x* out of *s* times when the event's probability is *p* in R with the function `dbinom`. The arguments of this function we deal with here are:

- *x*: the frequency of the event (e.g., three times heads);
- *s*: the number of trials in which the event could occur (e.g., three tosses);
- *p*: the probability of the event in each trial (e.g., 50%).

You know that the probability to get three heads in three tosses when the probability of head is 50% is 12.5%. In R:[7]

```
> dbinom(3,·3,·0.5)¶
[1]·0.125
```

As a matter of fact, you can compute the probabilities of all four possi-

[7] I will explain how to install R etc. in the next chapter. It doesn't really matter if you haven't installed R and/or can't enter or understand the above input yet. We'll come back to this …

ble numbers of heads – 0, 1, 2, and 3 – in one line (because, as we will see below, sequences of integers can be defined with a colon):

```
> dbinom(0:3,·3,·0.5)¶
[1]·0.125·0.375·0.375·0.125
```

In a similar fashion, you can also compute the probability that heads will occur two or three times by summing up the relevant probabilities:

```
> sum(dbinom(2:3,·3,·0.5))¶
[1]·0.5
```

Now you do the same for the probability to get 100 heads in 100 tosses,

```
> dbinom(100,·100,·0.5)¶
[1]·7.888609e-31
```

the probability to get heads 58 or more times in 100 tosses (which is larger than 5% and does not allow you to accept a one-tailed directional alternative hypothesis),

```
> sum(dbinom(58:100,·100,·0.5))¶
[1]·0.06660531
```

the probability to get heads 59 or more times in 100 tosses (which is smaller than 5% and does allow you to accept a one-tailed directional alternative hypothesis):

```
> sum(dbinom(59:100,·100,·0.5))¶
[1]·0.04431304
```

For two-tailed tests, you can do the same, e.g., compute the probability to get heads 39 times or less often, or 61 times and more often (which is smaller than 5% and allows you to accept a two-tailed non-directional alternative hypothesis):

```
> sum(dbinom(c(0:39,·61:100),·100,·0.5))¶
[1]·0.0352002
```

If we want to proceed the other way round in, say, the one-tailed case, then we can use the function `qbinom` to determine, for a given probability *q*, the number of occurrences of an event (or successes) in *s* trials, when the probability of the event in each trial is *p*. The arguments of this function are:

- `p`: the probability for which we want the frequency of the event (e.g., 12.51%);
- `size`: the number of trials in which the event could occur (e.g., three tosses);
- `prob`: the probability of the event in each trial (e.g., 50%);
- `lower.tail=TRUE`, if we consider the probabilities from 0 to `p` (i.e., probabilities from the lower/left tail of the distribution), or `lower.tail=FALSE` if we consider the probabilities from `p` to 1 (i.e., probabilities from the upper/right tail of the distribution); note, you can abbreviate `TRUE` and `FALSE` as `T` and `F` respectively (but cf. below).

Compare Figure 10 and the following code to the results discussed in the text above. On the *x*-axis, you find the numbers of heads (for three and 100 tosses in the left and the right panel respectively) and on the *y*-axis the cumulative probability of all possible numbers of heads included on the *x*-axis (i.e., 0.125, 0.5, 0.875, and 1 in the left panel).

```
> qbinom(0.1251,·3,·0.5,·lower.tail=FALSE)¶
[1]·2
> qbinom(0.1249,·3,·0.5,·lower.tail=FALSE)¶
[1]·3
> qbinom(0.05,·100,·0.5,·lower.tail=FALSE)¶
[1]·58
```

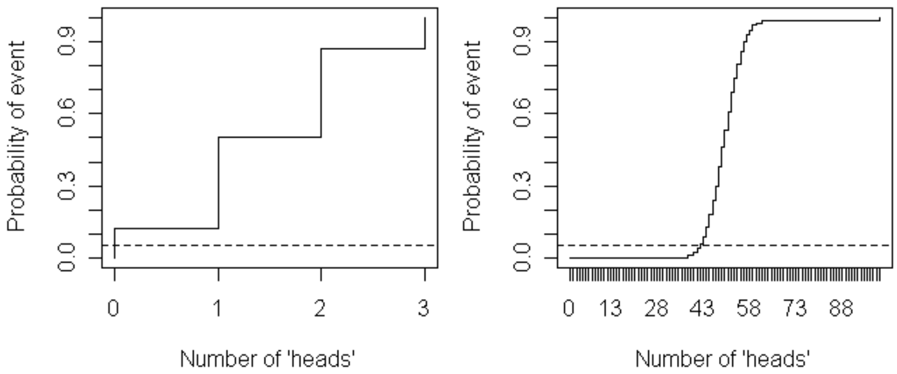


Figure 10. The probabilities for all possible results of three tosses (left panel) or 100 tosses (right panel); the dotted line is at 0.05

3.4.2. Extension: continuous probability distributions

In the above examples, we always had only one variable with two levels. Unfortunately, life is usually not that easy. On the one hand, we have seen above that our categorical variables will often involve more than two levels. On the other hand, if the variable in question is ratio-scaled, then the computation of the probabilities of all possible states or levels is not possible. For example, you cannot compute the probabilities of all possible reaction times to a stimulus. For this reason, many statistical techniques do not compute an exact *p*-value as we did, but are based on the fact that, as the sample size increases, the probability distributions of events begin to approximate those of mathematical distributions whose functions/equations and properties are very well known. For example, the curve in Figure 9 for the 100 coin tosses is very close to that of a bell-shaped normal distribution. In other words, in such cases the *p*-values are estimated on the basis of these equations, and such tests are called parametric tests. Four such distributions will be important for the discussion of tests in Chapters 4 and 5:

- the standard normal distribution with *z*-scores (`norm`);
- the *t*-distribution (`t`);
- the *F*-distribution (`f`);
- the chi-square- / χ^2-distribution (`chisq`).

For each of these distributions, there is a function whose name begins with *q* and ends with the above function name (i.e. `qnorm`, `qt`, `qf`, `qchisq`), and a function whose name begins with *p* and ends with the above function name (i.e. `pnorm`, `pt`, `pf`, `pchisq`). By analogy to our adding up the lengths of the bars that make up the curve of the binomial distribution, the functions beginning with *q* compute how much in percent of the area under the curve (the whole of which is defined as 1) a particular *z*-/*t*-/*F*-/χ^2-value defines. For example, in the very similar-looking binomial distribution in Figure 8 above we added probabilities from the right hand side moving to the left in order to test how often you can lose without being able to adopt the alternative hypothesis, using the function `qbinom`. The continuous probability functions can be used rather similarly. If we want to perform a one-tailed test on the basis of the standard normal distribution – the bell-shaped normal distribution with a mean of 0 and a standard deviation of 1, I will explain these concepts in detail below – then we can determine which value on the *x*-axis cuts off 5% of the left or the right part of the curve:

```
> qnorm(0.05,·lower.tail=TRUE)·#·one-tailed·test,·left·panel¶
[1]·-1.644854
> qnorm(0.95,·lower.tail=TRUE)·#·one-tailed·test,·
      right·panel¶
[1]·1.644854
```

That means, the grey area under the curve in the left panel of Figure 11 in the range $-\infty \leq x \leq -1.644854$ corresponds to 5% of the total area under the curve. Since the standard normal distribution is symmetric, the same is true of the grey area under the curve in the right panel in the range $1.644854 \leq x \leq \infty$.

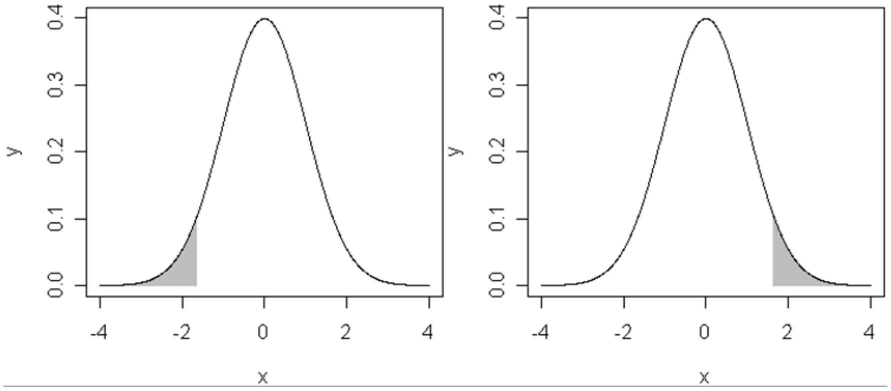


Figure 11. Density function of the standard normal distribution for $p_{\text{one-tailed}} = 0.05$

This corresponds to a one-tailed test since you only look at one side of the curve, and if you were to get a value of -1.7 for such a one-tailed test, then that would be a significant result. For a corresponding two-tailed test at the same significance level, you would have to consider both areas under the curve (as in Figure 9) and consider 2.5% on each edge to arrive at 5% altogether. To get the *x*-axis values that *jointly* cut off 5% under the curve, this is what you could enter into R; code lines two and three are different ways to compute the same thing (cf. Figure 12):

```
> qnorm(0.025,·lower.tail=TRUE)·#·two-tailed·test,·
      left·shaded·area:·∞≤x≤-1.96¶
[1]·-1.959964
> qnorm(0.975,·lower.tail=TRUE)·#·two-tailed·test,·
      right·shaded·area:·1.96≤x≤∞¶
[1]·1.959964
> qnorm(0.025,·lower.tail=FALSE)·#·two-tailed·test,·
      right·shaded·area:·1.96≤x≤∞¶
[1]·1.959964
```

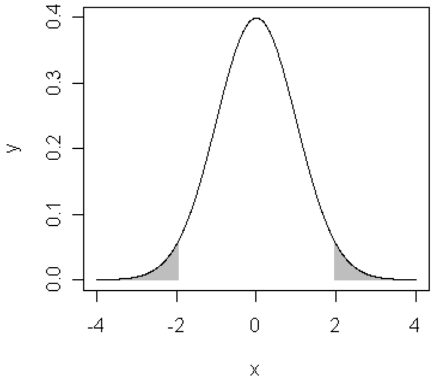


Figure 12. Density function of the standard normal distribution for $p_{\text{two-tailed}} = 0.05$

Again, you see that with non-directional two-tailed tests you need a more extreme result for a significant outcome: -1.7 would not be enough. In sum, with the q-functions we determine the minimum one- or two-tailed statistic that we must get to obtain a particular *p*-value. For one-tailed tests, you typically use $p = 0.05$; for two-tailed tests you typically use $p = {}^{0.05}/_{2} = 0.025$ on each side.

The functions whose names start with *p* do the opposite of those beginning with *q*: with them, you can determine which *p*-value our statistic corresponds to. The following two rows show how you get *p*-values for one-tailed tests (cf. Figure 11):

```
> pnorm(-1.644854,·lower.tail=TRUE)·#·one-
      tailed·test,·left·panel¶
[1]·0.04999996
> pnorm(1.644854,·lower.tail=TRUE)·#·one-
      tailed·test,·left·panel¶
[1]·0.95
```

For the two-tailed test, you of course must multiply the probability by two because whatever area under the curve you get, you must consider it on both sides of the curve. For example (cf. again Figure 12):

```
> 2*pnorm(-1.959964,·lower.tail=TRUE)·#·two-tailed·test¶
[1]·0.05
```

The following confirms what we said above about the value of -1.7: that value is significant in a one-tailed test, but not in a two-tailed test:

```
> pnorm(-1.7,·lower.tail=TRUE)·#·significant,·since·<0.05¶
[1]·0.04456546
> 2*pnorm(-
     1.7,·lower.tail=TRUE)·#·not·significant,·since·>0.05¶
[1]·0.08913093
```

The other p/q-functions work in the same way, but will require some additional information, namely so-called degrees of freedom. I will not explain this notion here in any detail but instead cite Crawley's (2002: 94) rule of thumb: "[d]egrees of freedom [*df*] is the sample size, *n*, minus the number of parameters, *p*, estimated from the data." For example, if you compute the mean of four values, then *df* = 3 because when you want to make sure you get a particular mean out of four values, then you can choose three values freely, but the fourth one is then set. If you want to get a mean of 8, then the first three values can vary freely and be 1, 1, and 1, but then the last one must be 29. Degrees of freedom are the way in which sample sizes and the amount of information you squeeze out of a sample are integrated into the significance test.

The parametric tests that are based on the above distributions are usually a little easier to compute (although this is usually not an important point anymore, given the computing power of current desktop computers) and more powerful, but they have one potential problem. Since they are only estimates of the real *p*-value based on the equations defining z-/t-/F-/χ^2-values, their accuracy is dependent on how well these equations reflect the distribution of the data. In the above example, the binomial distribution in Figure 9 and the normal distribution in Figure 12 are extremely similar, but this may be very different on other occasions. Thus, parametric tests make distributional assumptions – the most common one is in fact that of a normal distribution – and you can use such tests only if the data you have meet these assumptions. If they don't, then you must use a so-called *non-parametric test* or use a permutation test or other resampling methods. For nearly all tests introduced in Chapters 4 and 5 below, I will list the assumptions which you have to test before you can apply the test, explain the test itself with the computation of a *p*-value, and illustrate how you would summarize the result in the third (results) part of the written version of your study. I can already tell you that you should always provide the sample sizes, the obtained effect (such as the mean, the percentage, the difference between means, etc.), the name of the test you used, its statistical parameters, the *p*-value, and your decision (in favor of or against the alternative hypothesis). The interpretation of these findings will then be discussed in the fourth and final section of your study.

Recommendation(s) for further study

Good and Hardin (2006: Ch. 1 and 2) for many interesting and practically relevant tips as well as Good and Hardin (2006: Ch. 8) on information you should provide in your methods and results sections

Warning/advice

Do not give in to the temptation to use a parametric test when its assumptions are not met. What have you gained when you do the wrong test and maybe publish wrong results and get cited because of the methodological problems of your study?

4. The design of an experiment: introduction

In this section, we will deal with a few fundamental rules for the design of experiments.[8] The probably most central notion in this section is the token set (cf. Cowart 1997). I will distinguish two kinds of token sets, *schematic token sets* and *concrete token sets*. A schematic token set is typically a tabular representation of all experimental conditions. To explain this more clearly, let us return to the above example of particle placement.

Let us assume you do want to investigate particle placement not only on the basis of corpus data, but also on the basis of experimental data. For instance, you might want to determine how native speakers of English rate the acceptability of sentences (the dependent variable ACCEPTABILITY) that differ with regard to the constructional choice (the first independent variable CONSTRUCTION: *VPO* vs. *VOP*) and the part of speech of the head of the direct object (the second independent variable OBJPOS: *PRONOMINAL* vs. *LEXICAL*).[9] Since there are two independent variables for each of the two levels, there are 2·2 = 4 experimental conditions. This set of experimental conditions is the schematic token set, which is represented in two different forms in Table 11 and Table 12. The participants/subjects of course never get to see the schematic token set. For the actual experiment, you must develop concrete stimuli – a concrete token set that realizes the variable level combinations of the schematic token set.

[8] I will only consider the simplest and most conservative kind of experimental design, *factorial designs*, where every variable level is combined with every other variable level.

[9] For expository reasons, I only assume two levels of OBJPOS.

Table 11. Schematic *token set* for CONSTRUCTION × OBJPOS 1

	OBJPOS: *PRONOMINAL*	OBJPOS: *LEXICAL*
CONSTRUCTION: *VPO*	V Part pron. $NP_{dir. obj.}$	V Part lexical $NP_{dir. obj.}$
CONSTRUCTION: *VOP*	V pron. $NP_{dir. obj.}$ Part	V lexical $NP_{dir. obj.}$ Part

Table 12. Schematic *token set* for CONSTRUCTION × OBJPOS 2

Experimental condition	CONSTRUCTION	OBJPOS
1	*VPO*	*PRONOMINAL*
2	*VPO*	*LEXICAL*
3	*VOP*	*PRONOMINAL*
4	*VOP*	*LEXICAL*

However, both the construction of such concrete token sets and the actual presentations of the concrete stimuli are governed by a variety of rules that aim at minimizing undesired sources of noise in the data. Three such sources are particularly important:

- *knowledge of what the experiment is about*: you must make sure that the participants in the experiment do not know what is being investigated before or while they participate (after the experiment you should of course tell them). This is important because otherwise the participants might make their responses socially more desirable or change the responses to 'help' the experimenter.
- *undesirable experimental effects*: you must make sure that the responses of the subjects are not influenced by, say, habituation to particular variable level combinations. This is important because in the domain of, say, acceptability judgments, Nagata (1987, 1989) showed that such judgments can change because of repeated exposure to stimuli and this may not be what you're interested in.
- *evaluation of the results*: you must make sure that the responses of the subjects can be interpreted unambiguously. Even a large number of willing and competent subjects is useless if your design does not allow for an appropriate evaluation of the data.

In order to address all these issues, you have to take the rules in (4) to (12) under consideration. Here's the first one in (4):

(4) The stimuli of each individual concrete token set differ with regard

to the variable level combinations under investigation (and ideally only with regard to these and nothing else).

Consider Table 13 for an example. In Table 13, the stimuli differ only with respect to the two independent variables. If this was not the case (for example, because the left column contained the stimuli *John picked up it* and *John brought it back*) and you found a difference of acceptability between them, then you would not know what to attribute this difference to – the different construction (which would be what this experiment is all about), the different phrasal verb (that might be interesting, but is not what is studied here), to an interaction of the two … The rule in (4) is therefore concerned with the factor 'evaluation of the results'.

Table 13. A concrete token set for CONSTRUCTION × OBJPOS 1

	OBJPOS: *PRONOMINAL*	OBJPOS: *LEXICAL*
CONSTRUCTION: *VPO*	John picked up it.	John picked up the keys.
CONSTRUCTION: *VOP*	John picked it up.	John picked the keys up.

When creating the concrete token sets, it is also important to control for variables which you are not interested in and which make it difficult to interpret the results with regard to the variables that you are interested in. In the present case, for example, the choice of the verbs and the direct objects may be important. For instance, it is well known that particle placement is also correlated with the concreteness of the referent of the direct object. There are different ways to take such variables, or sources of variation, into account. One is to make sure that 50% of the objects are abstract and 50% are concrete for each experimental condition in the schematic token set (as if you introduced an additional independent variable). Another one is to use only abstract or only concrete objects, which would of course entail that whatever you find in your experiment, you could strictly speaking only generalize to that class of objects.

Recommendation(s) for further study

Good and Hardin (2006: 38f.) on handling extraneous variables

(5) You must use more than one concrete token set, ideally as many concrete token sets as there are variable level combinations (or a multiple thereof).

One reason for the rule in (5) is that, if you only used the concrete token set in Table 13, then a conservative point of view would be that you could only generalize to other sentences with the transitive phrasal verb *pick up* and the objects *it* and *the book*, which would probably not be the most interesting study ever. Thus, the first reason for (5) is again concerned with the factor 'evaluation of results', and the remedy is to create different concrete token sets with different verbs and different objects such as those shown in Table 14 and Table 15, which also must conform to the rule in (4). For your experiment, you would now just need one more.

A second reason for the rule in (5) is that if you only used the concrete token set in Table 13, then subjects would probably able to guess the purpose of the experiment right away: since our token set had to conform to the rule in (4), the subject can identify the relevant variable level combinations quickly because those are the only things according to which the sentences differ. This immediately brings us to the next rule:

(6) Every subject receives maximally one item out of a concrete token set.

Table 14. A concrete token set for CONSTRUCTION × OBJPOS 2

	OBJPOS: *PRONOMINAL*	OBJPOS: *LEXICAL*
CONSTRUCTION: *VPO*	Mary brought back him.	Mary brought back his dad.
CONSTRUCTION: *VOP*	Mary brought him back.	Mary brought his dad back.

Table 15. A concrete token set for CONSTRUCTION × OBJPOS 3

	OBJPOS: *PRONOMINAL*	OBJPOS: *LEXICAL*
CONSTRUCTION: *VPO*	I eked out it.	I eked out my living.
CONSTRUCTION: *VOP*	I eked it out.	I eked my living out.

As I just mentioned, if you do not follow the rule in (6), the subjects might guess from the minimal variations within one concrete token set what the whole experiment is about: the only difference between *John picked up it* and *John picked it up* is the choice of construction. Thus, when subject X gets to see the variable level combination (CONSTRUCTION: *VPO* × OBJPOS: *PRONOMINAL*) in the form of *John picked up it*, then the other experimental items of Table 13 must be given to other subjects. In that regard, the rules in both (5) and (6) are (also) concerned with the factor 'knowledge of what the experiment is about'.

(7) Every subject is presented every variable level combination.

The motivation for the rule in (7) are the factors 'undesirable experimental effects' and 'evaluation of the results'. First, if several experimental items you present to a subject only instantiate one variable level combination, then habituation effects may distort the results. Second, if you present one variable level combination to a subject very frequently and another one only rarely, then whatever difference you find between these variable level combinations may theoretically just be due to the different frequencies of exposure and not due to the effects of the variable level combinations under investigation.

(8) Every subject gets to see every variable level combination more than once and equally frequently.

(9) Every experimental item is presented to more than one subject and to equally many subjects.

These rules are motivated by the factor 'evaluation of the results'. You can see what their purpose is if you think about what happens when you try to interpret a very unusual reaction by a subject to a stimulus. On the one hand, that reaction could mean that the item itself is unusual in some respect in the sense that every subject would react unusually – but you can't test that if that item is not also given to other subjects, and this is the reason for the rule in (9). On the other hand, the unusual reaction could mean that only this particular subject reacts unusually to that variable level combination in the sense that the same subject would react more 'normally' to other items instantiating the same variable level combination – but you can't test that if that subject does not see other items with the same variable level combination, and this is the reason for the rule in (8).

(10) The experimental items are interspersed with distractors / filler items; there are minimally as many filler items as real experimental items per subject, but ideally two or three times as many filler items as real experimental items per subject.

The reason for this rule is obviously 'knowledge of what the experiment is about': you do not want the subjects to be able to guess the purpose of the experiment (or have them think they know the purpose of the experiment) so that they cannot distort the results.[10]

An additional well-known factor that can distort results is the order in which items and distractors are presented. To minimize such effects, you must take into consideration the final two rules:

(11) The order of experimental and filler items is pseudorandomized.
(12) The order of experimental and filler items is pseudorandomized differently for every subject.

The rule in (11) requires that the order of experimental items and filler items is randomized using a random number generator, but it is not completely random – hence *pseudo*randomized – because the ordering resulting from the randomization must usually be 'corrected' such that

- the first stimulus (e.g., the first question on a questionnaire) is not an experimental item but a distractor;
- experimental items do not follow each other directly;
- experimental items exhibiting the same variable level combinations do not follow each other, which means that, after *John picked it up*, the next experimental item must not be *Mary brought him back* even if the two are interrupted by distractors.

The rule in (12) means that the order of stimuli must vary pseudorandomly across subjects so that whatever you find cannot be attributed to systematic order effects: every subject is exposed to a different order of experimental items and distractors. Hence, both (11) and (12) are concerned with 'undesirable experimental effects ' and 'evaluation of the results'.

Only after all these steps have been completed properly can you begin to print out the questionnaires and have subjects participate in an experiment. It probably goes without saying that you must carefully describe how you set up your experimental design in the methods section of your study. Since this is a rather complex procedure, we will go over it again in the following section.

[10] In many psychological studies, not even the person actually conducting the experiment (in the sense of administering the treatment, handing out the questionnaires, …) knows the purpose of the experiment. This is to make sure that the experimenter cannot provide unconscious clues to desired or undesired responses. An alternative way to conduct such so-called double-blind experiments is to use standardized instructions in the forms of videotapes or have a computer program provide the instructions.

Warning/advice

You must be prepared for the fact that usually not all subjects answer all questions, give all the acceptability judgments you ask for, show up for both the first and the second test, etc. Thus, you should plan conservatively and try to get more subjects than you thought you would need in the first place. As mentioned above, you should still include these data in your table and mark them with “NA”. Also, it is often very useful to carefully examine the missing data for whether their patterning reveals something of interest (it would be very important if, say, 90% of the missing data exhibited only one variable level combination or if 90% of the missing data were contributed by only two out of, say, 60 subjects).

One final remark about this before we look at another example. I know from experience that the previous section can have a somewhat discouraging effect. Especially beginners read this and think “how am I ever going to be able to set up an experiment for my project if I have to do all this? (I don’t even know my spreadsheet well enough yet …)” And it is true: I myself still need a long time before a spreadsheet for an experiment of mine looks the way it is supposed to. But if you do not go through what at first sight looks like a terrible ordeal, your results might well be, well, let’s face it, crap! Ask yourself what is more discouraging: spending maybe several days on getting the spreadsheet right, or spending maybe several weeks for doing a simpler experiment and then having unusable results …

5. The design of an experiment: another example

Let us assume you want to investigate which variables determine how many elements a quantifier such as *some* refers to; consider (13):

(13) a. [$_{NP}$ some balls [$_{PP}$ in front of [$_{NP}$ the cat]]
b. [$_{NP}$ some balls [$_{PP}$ in front of [$_{NP}$ the table]]
c. [$_{NP}$ some cars [$_{PP}$ in front of [$_{NP}$ the building]]

Thus, the question is: are *some balls in front of the cat* as many balls as *some balls in* front *of the table*? Or: does *some balls in front of the table* mean as many balls as *some cars in front of the building* means cars? What – or more precisely, how many – does *some* mean? Your study of the literature may have shown that at least the following two variables influence the quantities that *some* denotes:

- OBJECT: the size of the object referred to by the first noun: *SMALL* (e.g. *cat*) vs. *LARGE* (e.g. *car*);
- REFPOINT: the size of the object introduced as a reference in the PP: *SMALL* (e.g. *cat*) vs. *LARGE* (e.g. *building*).[11]

Obviously, a study of *some* with these two variables results in a schematic token set with four variable level combinations, as represented in Table 16.

The (non-directional) hypotheses for this study are:

H_0: The average estimate of how many *some* denotes is independent of the sizes of the objects (OBJECT: *SMALL* vs. *LARGE*) and the sizes of the reference points (REFPOINT: *SMALL* vs. *LARGE*) in the utterances for which subjects provide estimates: $mean_{\text{SMALL+SMALL}} = mean_{\text{SMALL+LARGE}} = mean_{\text{LARGE+SMALL}} = mean_{\text{LARGE+LARGE}}$.

H_1: The average estimate of how many *some* denotes is dependent on the sizes of the objects (OBJECT: *SMALL* vs. *LARGE*) and/or the sizes of the reference points (REFPOINT: *SMALL* vs. *LARGE*): there is at least one ≠ in the above equation ("and/or" because of the possibility of an interaction; cf. above Section 1.3.2.3).

Table 16. Token sets (schematic + concrete) for OBJECT × REFPOINT

	REFPOINT: *SMALL*	REFPOINT: *LARGE*
OBJECT: *SMALL*	*SMALL + SMALL*: *some dogs next to a cat*	*SMALL + LARGE*: *some dogs next to a car*
OBJECT: *LARGE*	*LARGE + SMALL*: *some cars next to a cat*	*LARGE + LARGE*: *some cars next to a fence*

Let us now also assume you want to test these hypotheses with a questionnaire: subjects will be shown phrases such as those in Table 16 and then asked to provide estimates of how many elements a speaker of such a phrase would probably intend to convey – how many dogs were next to a cat etc. Since you have four variable level combinations, you need at least four concrete token sets (the rule in (5)), which are created according to the rule in (4). According to the rules in (6) and (7) this also means you need at least four subjects: you cannot have fewer because then some subject

[11] I will not discuss here how to decide what is 'small' and what is 'large'. In the study from which this example is taken, the sizes of the objects were determined on the basis of a pilot study prior to the real experiment.

would see more than one stimulus from one concrete token set. You can then assign experimental stimuli to the subjects in a rotating fashion. The result of this is shown in the sheet <Phase 1> of the file <C:/_sflwr/_inputfiles/01-5_ExperimentalDesign.ods> (just like all files, this one too can be found on the companion website at <http://groups.google.com/group/statforling-with-r/web/statistics-for-linguists-with-r> or its mirror). The actual experimental stimuli are represented only schematically as a unique identifying combination of the number of the token set and the variable levels of the two independent variables (in column F).

As you can easily see in the table on the right, the rotation ensures that every subject sees each variable level combination just once and each of these from a different concrete token set. However, we know you have to do more than that because in <Phase 1> every subject sees every variable level combination just once (which violates the rule in (8)) and every experimental item is seen by only one subject (which violates the rule in (9)). Therefore, you first re-use the experimental items in <Phase 1>, but put them in a different order so that the experimental items do not occur together with the very same experimental items (you can do that by rotating the subjects differently). One possible result of this is shown in the sheet <Phase 2>.

The setup in <Phase 2> does not yet conform to the rule in (8), though. For that, you have to do a little more. You must present more experimental items to, say, subject 1, but you cannot use the existing experimental items anymore without violating the rule in (6). Thus, you need four more concrete token sets, which are created and distributed across subjects as before. The result is shown in <Phase 3>. As you can see in the table on the right, every experimental item is now seen by two subjects (cf. the row totals), and in the columns you can see that each subjects sees each variable level combination in two different stimuli.

Now every subjects receives eight experimental items, you must now create enough distractors. In this example, let's use a ratio of experimental items to distractors of 1:2. Of course, 16 distractors are enough, which are presented to all subjects – there is no reason to create 8·16 = 128 distractors. Consider <Phase 4>, where the filler items have been added to the bottom of the table.

Now you must order the all stimuli – experimental items *and* distractors – for every subject. To that end, you can add a column called "RND", which contains random numbers ranging between 0 and 1 (you can get those from R or by writing "=RAND()" (without double quotes, of course) into a cell in OpenOffice.org Calc. If you now sort the whole spreadsheet

(i) according to the column “SUBJ” and then (ii) according to the column “RAND”, then all items of one subject are grouped together, and within each subject the order of items is random. This is required by the rule in (12) and represented in <Phase 5>.

When you look at <Phase 5>, you also see that the order of some elements must still be changed: red arrows in column H indicate problematic sequences of experimental items. To take care of these cases, you can arbitrarily pick one distractor out of a series of distractors and exchange their positions. The result is shown in <Phase 6>, where the green arrows point to corrections. If we had used actual stimuli, you could now create a cover sheet with instructions for the subjects and a few examples (which in the case of, say, judgments would ideally cover the extremes of the possible judgments!), paste the experimental stimuli onto the following page(s), and hand out the questionnaires. To evaluate this experiment, you would then have to compute a variety of means:

- the means for the two levels of OBJECT (i.e., $mean_{\text{OBJECT: }\textit{SMALL}}$ and $mean_{\text{OBJECT: }\textit{LARGE}}$);
- the means for the two levels of REFPOINT (i.e., $mean_{\text{REFPOINT: }\textit{SMALL}}$ and $mean_{\text{REFPOINT: }\textit{LARGE}}$);
- the four means for the interaction of OBJECT and REFPOINT.

We will discuss the method that is used to test these means for significant differences – analysis of variance or ANOVA – in Section 5.3.

Recommendation(s) for further study

Good and Hardin (2006: Ch. 13) and, when you have more experience with R, the website <http://cran.r-project.org/src/contrib/Views/ExperimentalDesign.html>

Now you should do the exercises for Chapter 1 (which you can find on the website) …