

Statistics

Simple methods for research in linguistics

Lukas Sönning

Lehrstuhl für englische Sprachwissenschaft einschließlich Sprachgeschichte
Universität Bamberg

lukas.soenning@uni-bamberg.de

Introductory remarks

This booklet was written for students doing quantitative research for a term paper or a final thesis in linguistics. It aims to explain the usage of a number of simple descriptive and inferential tools for data analysis. Four commonly occurring types of research design are illustrated and serve as examples throughout the text. In principle, the information given here should equip you with everything you need for the statistical analysis of your data (if your research design does not differ from the ones illustrated here). The appendix shows you how to carry out calculations by hand. To make things easier for students, an Excel module was created to accompany this text. It makes calculations and plotting easier (copy-and-paste) and is available on the server of the Chair of English Linguistics. Everything you see in here, you can do in Excel (including some Excel innovations, such as the dot plot and the boxplot proper).

Please consider this a working version of the booklet as well as the Excel module. Feedback on errors and lack of clarity is very much welcomed to improve future versions.

Lukas Sönning
November 2015

Contents

1. Basic concepts

- 1.1. Populations and samples
- 1.2. Variables

2. Descriptive statistics

- 2.1. Statistical measures
- 2.2. Plots

3. Inferential statistics: Confidence intervals

- 3.1. Basic concepts
- 3.2. Interpretation of confidence intervals
- 3.3. Reporting confidence intervals
- 3.4. Understanding confidence intervals: Some properties

4. The new statistics in practice: Effect sizes and confidence intervals

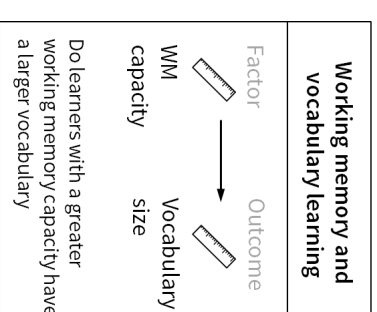
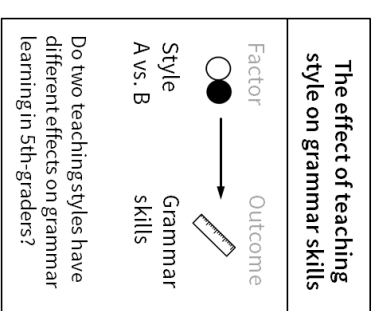
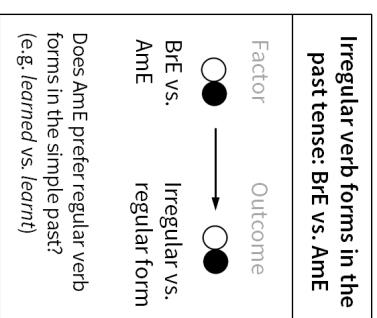
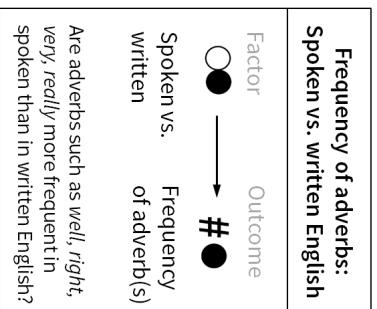
- 4.1. The new statistics vs. significance testing
- 4.2. Why the new statistics are better
- 4.3. Application to corpus data

5. Further reading

6. Constructing confidence intervals

- 6.1. Frequency data
- 6.2. Proportions (binary variables)
- 6.3. Quantitative data

References



Descriptive statistics

Single group/sample

Statistical measures
Relative frequency (per million words)

Plots
Dot plot or bar chart

Proportion, percentage

Dot plot or bar chart

Normal data: Mean, standard deviation
Non-normal: Median, trimmed mean (20%), interquartile range

Boxplot

Correlation coefficient
Normal data: Pearson
Non-normal: Spearman + Mean (median, trimmed mean) + Standard deviation (IQR)

Scatterplot

Comparing groups

Statistical measures
Ratio of frequencies

Plots
Dot plot with appended panel showing ratios

Difference of proportions

Dot plot with appended panel showing differences

Difference between means/medians/trimmed means
Parallel boxplots

Parallel boxplots

Difference between correlation coefficients
Groups in the same or in separate scatterplots

Groups in the same or in separate scatterplots

Inferential statistics

Single group/sample

Confidence interval
95% CI for a (relative) frequency

Plots
Dot plot with error bars

95% CI for a proportion

Dot plot with error bars

95% CI for a mean/median/trimmed mean
Text (or dot plot with error bars)

Text (or dot plot with error bars)

95% CI for a correlation coefficient
Text (or dot plot with error bars)

Text (or dot plot with error bars)

Confidence interval

95% CI for the ratio of frequencies

Plots
Dot plot with appended panel showing ratios + 95% CI

95% CI for a difference of proportions

Dot plot with appended panel showing differences + 95% CI

95% CI for a difference between means/trimmed means
Text (or dot plot with error bars)

Text (or dot plot with error bars)

95% CI for the difference between correlation coefficients
Text (or dot plot with error bars)

Text (or dot plot with error bars)

1. Basic concepts

1.1. Populations and samples

Researchers are interested in making statements that apply to **populations** – however, we only study **samples** and use this knowledge to learn about the population we are interested in.

Population: Collection of entities we are interested in and want to make generalizations about
Usually far too large to be studied exhaustively (this is especially true for language)

Sample: A much smaller collection of units from the population which we use to make generalizations
→ A sample should be **representative** of the population
→ It should be selected **randomly** (i.e. each unit has an equal chance of being chosen)



Estimation → Based on a sample we can make **estimates** about the population
→ The accuracy of our estimates depends on the size and representativity of the sample

1.2. Variables

Definition in plain words: Something we can observe
 Something that takes on different values/levels

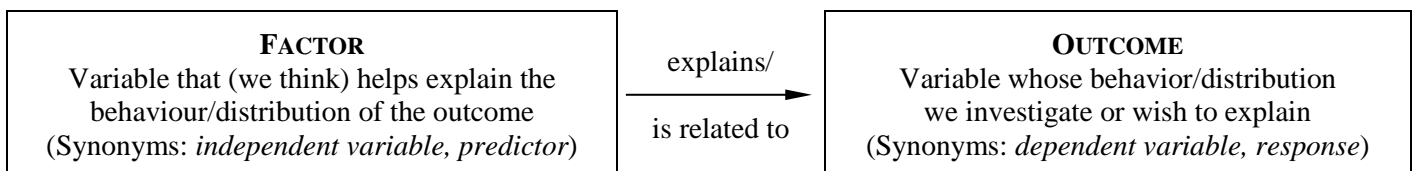
► Scales of measurement

Variables are classified in terms of scale of measurement, which basically expresses what type of information we get. We can distinguish 5 types for linguistic research.

Symbol	Scale of measurement	Description	Example
#●	Count/Frequency	Frequency of an event	Frequency of a word in a corpus
○●	Binary	Only 2 levels	Sex,
●●●	Nominal	3 or more levels	Native language, hair color
	Ordinal	Ordered levels	Questionnaire scales (agree...disagree)
	Quantitative	Measurements	Age, reaction time, test score

► Role in research design: Factor vs. outcome

We can also classify variables depending on the role they play in our research design. Unfortunately, several synonymous terms exist for each type.



► Operationalization

This fancy word refers to how we decide to measure a variable (e.g. specifying what to count, what to measure, how to measure it, etc.). Abstract (or latent) variables are more difficult to operationalize (e.g. PRONUNCIATION PROFICIENCY). Concrete variables are easier to operationalize (e.g. AGE, VERB TENSE). Note that **variables can be operationalized in different ways**. For example, LENGTH OF SUBJECT can be measured in terms of (i) number of words, (ii) number of morphemes, (iii) number of syllables, or (iv) number of characters. Look at how **other studies** have measured the variable you are interested in. This will make it easier for you to compare your results to those obtained in previous research.

2. Descriptive statistics

The tools offered by descriptive statistics are important in that they allow you to understand your data and to communicate your results. They thus serve as a medium of communication between data and researcher as well as between researcher and audience (readers/listeners). There are two types of tools for descriptive data analysis: (i) **statistical measures**, which summarize the information in your data using numbers and (ii) **plots** (=graphs, diagrams), which offer different ways of looking at and presenting data. What you need to bear in mind is that statistical measures **reduce your data** to a single number; plots can **retain much more information** if needed and therefore tell you much more about your data.

2.1. Statistical measures

Note 1. When reporting statistical measures you should **round (a lot)**. Do not report too many decimal places; maybe round them away completely (especially for percentages). They distract the eyes and make it much harder (for you and your audience) to compare the relevant digits. Decimal places also create the illusion of precise measurements; this is usually not the case (see section on inferential statistics).

Note 2. In the following we are going to distinguish between **two types of comparisons** which are often involved in (corpus) linguistic data analysis. Sometimes we investigate several **items**, e.g. different words or units such as constructions, collocations, lexical bundles, etc. For example, we may be interested in the frequency of the top 20 adverbs in spoken English. We are then comparing different items (adverbs) in the same group (spoken English). The term **group** is used in a *general* sense here to be applicable to different research designs. If we compare the frequency of adverbs in spoken vs. written English, we would say we are comparing two groups – the two “groups” being different language modes (in practical terms, two different (sub-)corpora).

▪ Frequency data

▷ Simple frequencies

The absolute number of occurrences observed (or counts) is called the **raw frequency**. Counts are usually expressed as a **relative frequency**, such as “per hundred” (%), “per thousand”, or “per million”. A relative frequency makes it possible to compare counts across different studies or samples. In corpus linguistics this is important, since the size of (sub-)corpora often differs. A typical measure in corpus linguistics is **per million words** (or per thousand words).

▷ Comparing two frequencies

Frequencies are usually compared using a **ratio** (e.g. 1.6 – word X occurred 1.6 times more often in corpus A than in corpus B). This is called a **frequency ratio**. It is better to always use relative frequencies for the calculation of such a ratio. Beware that the ratio changes depending on which frequency you divide by which. A ratio of 4.0 (4/1) corresponds to a ratio of 0.25 (1/4). Note that 4.0 may sound like a larger difference to some people than 0.25.

Technical note: The log ratio does not suffer from these drawbacks (see section 4.3 for an example and explanation).

▪ Binary and nominal data

▷ Simple proportions/percentages

Binary and nominal data are usually communicated using relative frequencies in the form of **percentages** (per 100) or **proportions** (per 1). For example, if we sample 200 instances of the verb learn in the simple past form, and 180 out of 200 of those occurrences were in the regular form (learned vs. learnt), we would summarize this finding saying that 180/200 occurrences were regular, which is a proportion of 0.9 or a percentage of 90 percent. Given this number we automatically know that 20/200 or 0.1 or 10 percent of the occurrences were irregular. You can always convert proportions to percentages (multiply by 100) if you prefer this way of reading the data.

▷ Comparing two proportions/percentages

We usually use the **difference of proportions** (or percentages) to compare two groups. For example if we observe that (in simple past contexts) learn occurs in the regular form 90% of the cases in American English, but only in 60% of the cases in British English, we can express this difference as 0.3 or 30 percent. It does not matter which category we focus on. For the irregular verb form, the difference between 10% (AmE) and 40% (BrE) is the same.

▪ Quantitative data

Note 1. While it is strictly speaking not appropriate to treat ordinal data in the same way as quantitative data, this is typically what researchers do. One argument against doing so is that the distance between the ordinal categories is not necessarily equal. A typical type of ordinal data in linguistics comes from **questionnaires**, where the response to items is usually indicated on a 3(+)-point scale (e.g. a 5-point scale: “I strongly agree”, “I agree”, “Undecided”, “I disagree”, “I strongly disagree”). When designing the questionnaire, try to create response categories that are as **equidistant** as possible – then you can justify the use of the following methods.

Note 2. The classic statistical measures for quantitative variables (mean and standard deviation) have one important weakness: They are very easily influenced by **outliers**. Outliers are unusual measurements in that they have a very high or very low value compared to the other measurements in the sample. Outliers commonly occur with variables that are bounded by zero (i.e., they only have positive values). Examples are duration and reaction time. You can detect outliers with the help of a **boxplot** (see below). If there is an unusually high number of outliers (more than 5-10 percent of the data) it is better to use robust statistical measures such as the median and the trimmed mean or the interquartile range. Another simple option is to compare the mean and the median (described below) – if they differ considerably, outliers are present.

▷ A single sample

Statistical measures for quantitative data are divided into **measures of location** and **measures of spread**. Measures of location tell us where the “typical” (i.e. the most representative) score of the sample lies. Measures of spread are just as important. They tell us how much the scores vary in the sample. Variation is frequently ignored when presenting and discussing results. However, it deserves the same amount of attention as the commonly discussed measures of location. These two types of measures always go **hand in hand** – both should be reported and interpreted.

Measures of location

Mean	<i>M</i>	The average of the scores
Median	<i>Mdn</i>	The score in the middle: half of the scores are higher, and half are lower
Trimmed mean	<i>M_{tr}</i>	The highest and lowest 20% of the scores are removed. Then you calculate the mean of the remaining 60% of the scores (20% trimming is common).

Measures of spread

Standard deviation	<i>SD</i>	If you add and subtract 1 standard deviation to the mean, you get an interval. Roughly 2/3 of the scores in the sample are in this interval. In other words, around 2/3 of the scores are less than 1 SD away from the mean.
Interquartile range	<i>IQR</i>	The interquartile range contains the middle 50 % of the scores. See below (boxplots) for an explanation.

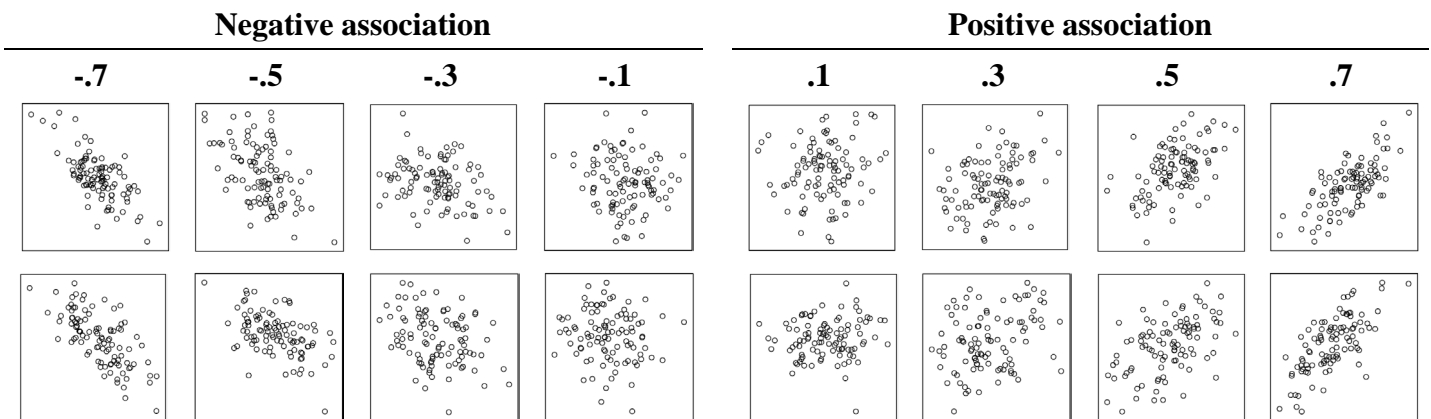
▷ Comparing two groups

An easy-to-interpret measure for comparing two groups is the **difference between their means/medians/trimmed means**.

Technical note: Another option are standardized measures (*r*-family or *d*-family measures), which make it easier to compare results across different variables and/or studies, especially if measurements are made on different scales.

▷ Association between two quantitative variables

The association between two quantitative variables is measured with a **correlation coefficient** r . Correlation coefficients may be unfamiliar to many people. They can range between -1 and +1, which reflects the strength (0 vs. 1) and the direction (+ vs. -) of an association. A **positive association** means: the higher the X, the higher the Y. This might be the case when comparing motivation with level of pronunciation proficiency – subjects with a higher motivation will on average also have higher proficiency. A **negative association**, on the other hand, means: the higher the X, the lower the Y (or vice versa). This might be observed if we measure exam anxiety and exam score. Other things being equal, students that feel more nervous and worried about an exam might perform worse than the more relaxed and confident ones. To get a feel for the correlation coefficient, it is best to look at pictures of data. The following plots show a few examples for a number of values for the correlation coefficient.



Sometimes, **benchmark values** are used to interpret a correlation coefficient as indicating a small (.1), medium (.3) or large (.5) effect (Cohen 1988). It is generally better to interpret the size of an effect by comparing it with the (i) the effect of other variables and/or (ii) the results obtained in other studies dealing with the same or a similar phenomenon. If no such information is available, these benchmarks may be used.

▷ Association between two quantitative variables: Comparing two groups

It is also possible to compare the correlation coefficients of two groups (or two studies). This comparison is usually expressed as the **difference of the correlation coefficients**.

▷ Effect of outliers

Like the mean and the standard deviation, the correlation coefficient is very **sensitive to outliers**. It is thus essential to check whether there are problems with outliers. Before calculating a correlation coefficient, inspect the two variables (i) separately using **boxplots** and (ii) together in a **scatterplot** (see below). If there are no unusual patterns in the data, you can go ahead and calculate the (Pearson) correlation coefficient (Excel function KORR). If the data look problematic, you should use Spearman's correlation coefficient, which is robust against outliers. The appendix explains how to calculate it.

2.2. Plots

Note 1. Data visualization is a broad and fascinating subject. In the following, I will try to put major recommendations into a nutshell. One thing you should keep in mind is that there is often **more than one way** of looking at/showing your data. It makes sense to try out different arrangements of the variables in the graph and alternate between combining groups in a single display and showing them separately. Different arrangements can show different aspects of the data. Avoid information overload.

Note 2. You may be unfamiliar with some of the chart types (especially dot plots and boxplots). The Chair of English Linguistics provides a free Excel module (accessible on the server), which makes it easy to create such charts by copy-and-pasting your data into a prefabricated spreadsheet.

▪ **General guidelines**

No pie charts. Pie charts make it difficult for the human eye to judge and compare percentages. The bar chart and the dot plot always outperform the pie.

No 3-D effects. 3-D charts are a waste of ink and more difficult to read than ordinary bar charts.

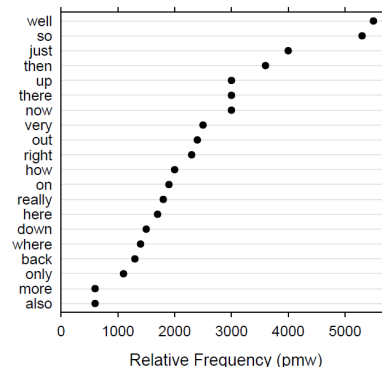
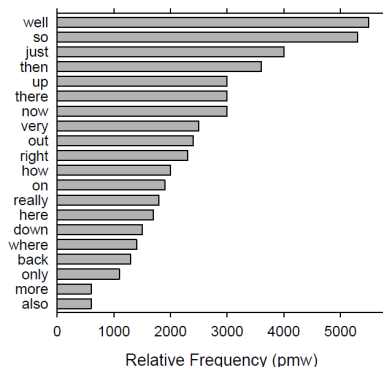
Be minimalistic. It is generally recommended to try and reduce the amount of ink so the relevant data-based information is foregrounded. Reduce the number of tick marks, do not use grid lines (or very lightly colored ones). Also reduce the size of the chart to a sensible level.

Avoid color. Color should only be used if it serves a purpose. It usually doesn't – grey scales can do the job just as well. Your plot should be pretty because of the interesting data it shows. Avoiding color will: (i) save printing costs, (ii) make sure black-and-white reproductions (print, copy) are unambiguous (converted to greyscales some colors may not be distinguishable), (iii) allow you to use color in a presentation to highlight things

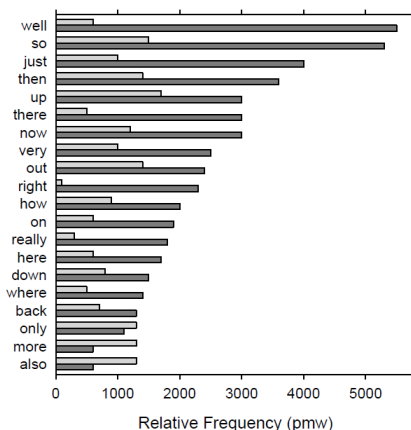
Order. This is in fact an underestimated and underutilized recommendation. If the groups/categories/ items you plot do not have a natural/logical order, then you should order them based on some aspect of the data, for example value (see examples below).

▪ **Frequency data: Bar charts and dot plots**

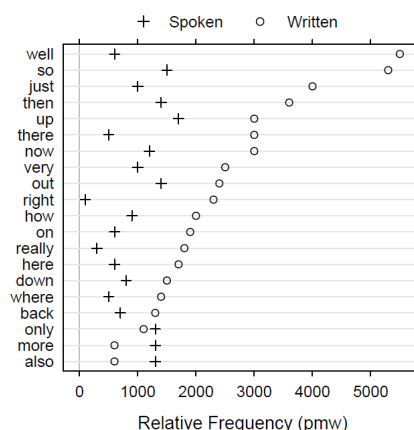
Frequency data can be shown using **bar charts** and **dot plots**. While bar charts are encountered very often, dot plots have a few important advantages. When comparing frequencies of several items (here: frequency of top 20 adverbs in spoken English) bar charts yield acceptable results. Note the **horizontal** format, which ensures readability of the labels (adverbs). The items have been **ordered**.



If we wish to compare two groups (corpora), bar charts are arguably less suitable. The minimalistic design of dot plots manages to keep clear vision and makes it easier to see patterns and focus on just one group while mentally filtering out the other. The plots below compare the frequency of adverbs in spoken vs. written English. Again, a horizontal format is chosen and the words are ordered. There are **different options for ordering** here. We could order them based on the frequency of spoken and written combined, the frequency of written or the frequency of spoken. Here the same ordering as above was chosen.

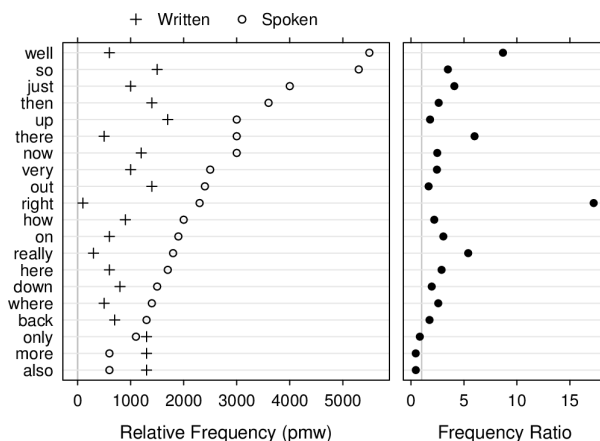


spoken
written



When **comparing frequencies** we usually use a **ratio** to express the comparison in numbers. We can add this information to the plot easily by **adding a second panel** on the right. It shows the ratio for spoken vs. written English for each of the 20 adverbs. We can immediately see that *right* sticks out; *well*, *there* and *really* also appear to be very typical for spoken English.

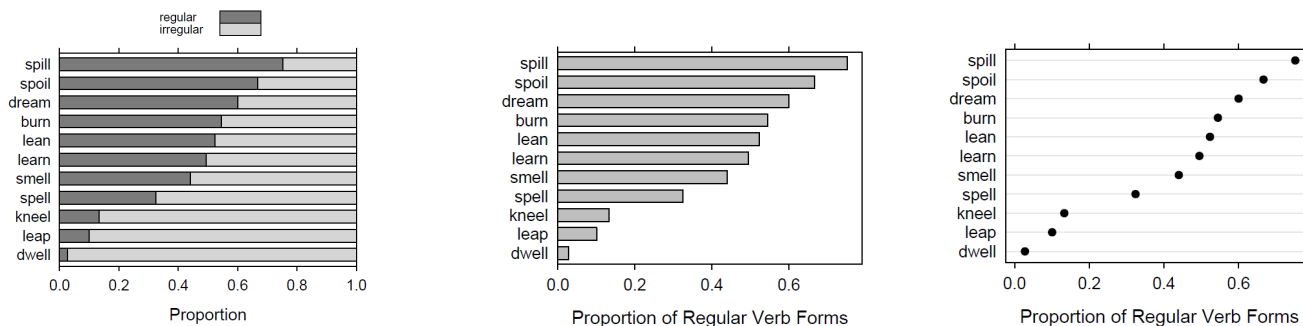
Adding a second panel that expresses comparisons is in fact a **powerful plotting strategy** and very much recommended. It offers a further **ordering option**: we could order the adverbs according to frequency ratio.



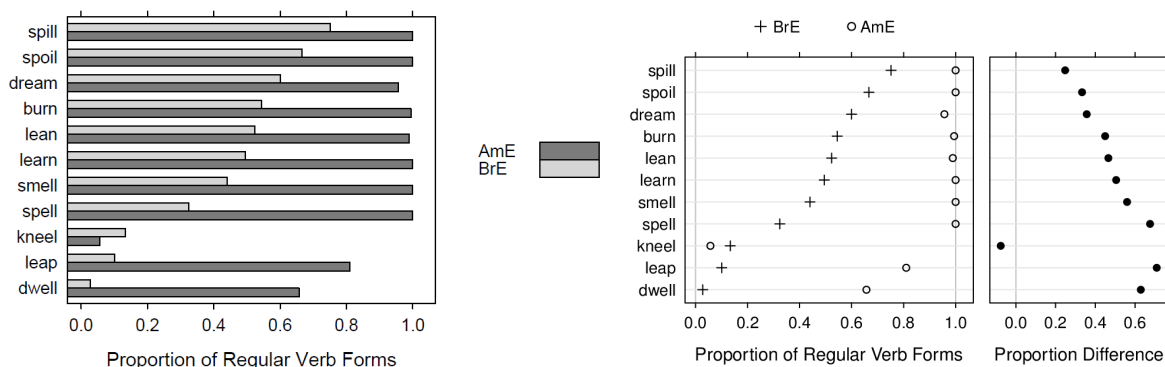
There are obviously different choices for **plotting symbols in dot plots**. What is most important is that the groups (here: spoken vs. written) remain visually distinguishable. Two sets of symbols can be recommended: if there is no overplotting (i.e. frequencies in the same line are never exactly the same), you can use (● and ○). The symbols chosen here (+ and ○) have the advantage that both are visible if overplotting occurs. The use of the more salient filled circle (●) is reserved for showing the magnitude of interest. Here this is the frequency ratio of spoken vs. written.

▪ **Binary and nominal data:** Bar charts and dot plots

Percentages are usually shown using **bar charts** (no pie charts!). Bar charts are a good choice when there are only a few groups/items to compare. For binary data (which we focus on here), the use of stacked bar charts makes sense if we only look at one group. The following plots show the occurrence of regular and irregular verb forms in the simple past in BrE newspaper texts (data from Levin 2009). If we **focus on one of the categories** (e.g. regular verb forms) we can also use a **simple bar chart** or a **dot plot**. The axis should unambiguously indicate which category is plotted. Note that the verbs are **ordered by “regularity”**.



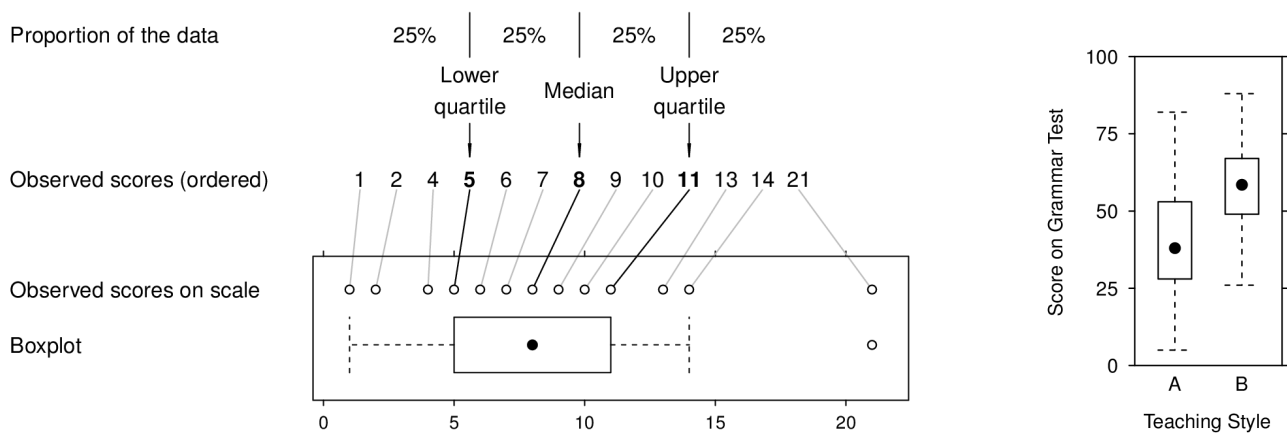
As soon as we **compare two groups**, for example BrE vs. AmE, stacked bar charts are a poor choice. We should instead focus on one category (here: regular verb forms) and use a grouped bar chart or a dot plot. As in the example above, the grouped bar chart is very busy. This happens when comparing a larger number of items, which is typically the case in corpus linguistics. By append a second panel to the dot plot we can directly show the comparison between the varieties, which is simply the difference between the proportions (of regular verb forms). An interesting pattern emerges.



▪ **Quantitative data:** Boxplots and scatterplots

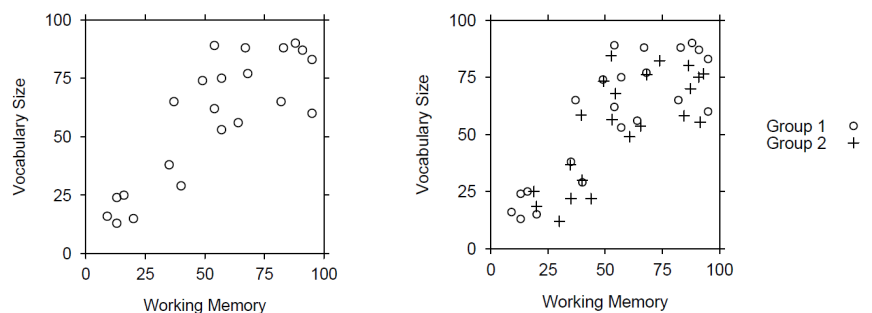
Single-number summaries (such as the mean and the standard deviation) discard a lot of information in quantitative data. A very useful chart type for showing the distribution of scores in a sample is the **boxplot**. It gives a wealth of information: central tendency is indicated by the **median** (in our version of the boxplot as a black dot). The box extends from the lower to the upper quartile. These scores cut off the bottom 25% and the top 25% of the scores, respectively. The box shows the **interquartile range**. It contains the middle 50% of the data. Unusually large or small scores, so-called **outliers**, are identified and plotted separately as individual open circles. Boxplots are a very useful tool to check whether there is a problem with outliers and whether a robust measure of location and spread should be chosen to summarize the data.

The following figure illustrates the boxplot. There are 13 scores in the sample, which range from 1 to 21. The ordered scores are shown in the upper part of the display. The median divides the ordered scores into two equal halves. The median, the upper quartile and the lower quartile divide the scores into four parts with a roughly equal number of scores. Between the upper and lower quartile we thus (roughly) find the central 50% of the scores.

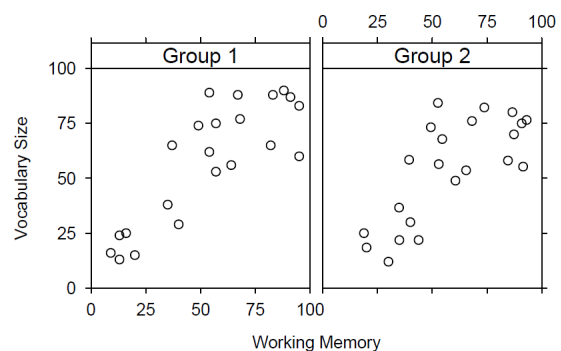


A further advantage of boxplots is that they make it very easy to **compare two or more groups** in the same plot. They enable us to contrast groups in terms of central tendency (comparison of the medians) and variability (comparison of the boxes). You can create such boxplots via copy-and-paste in the Excel module.

Scatterplots can be used to inspect the association between two quantitative variables. You should use **open circles** as plotting symbols, so overlapping data points are still distinguishable. A grid is usually not necessary and may distract from the pattern in the point cloud.

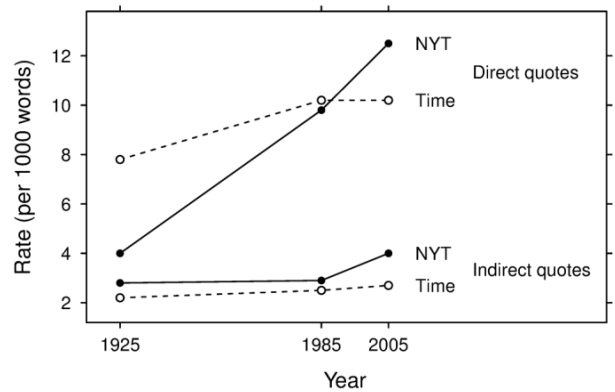


We can **compare two groups** by plotting them into the same panel using different plotting symbols or colors. Here the **use of color** often makes a plot more readable. Another strategy is to plot groups into different panels. Make sure the axes have the same range when showing the plots on the same page/slide.



▪ Time as a variable: Line plots

Line plots are the best choice when showing trends over time or quasi-time differences such as age groups or developmental stages. Line plots usually do not need color: vary the plotting symbols (\circ \bullet $+$) and line types (---- ____). What makes line plots much easier to read is when the lines are **labeled directly** in the plot. You should **avoid a legend** (if possible). It also makes sense to use **hierarchical labeling** (as in the example on the right). The plot on the right shows the diachronic change in frequency of direct vs. indirect quotes in two newspapers (data from Biber and Gray 2013).



3. Inferential statistics: Confidence intervals

As was mentioned above, researchers are usually interested in arriving at more general statements that do not only apply to the specific sample(s) they have collected but more generally, to the underlying population. The statistical measures, differences and ratios we calculate based on our sample(s) are only **estimates** of the true state of affairs in the population. Every estimate is connected with a certain degree of **uncertainty**. Confidence intervals express this uncertainty and help us judge the **precision** of our estimate. You have to understand 4 key concepts to know what inferential statistics is all about and how confidence intervals are used and interpreted.

3.1. Basic concepts

▪ Sample vs. population

Researchers study samples to make statements about populations. For example, a study investigating the development of pronunciation skills in German learners of English may analyze the speech of 60 learners. However, the goal would be to generalize to all German learners. If we want to go from statements about our sample to statements about a population we need the tools provided by inferential statistics.

▪ Sample statistic vs. population parameter

The statistical measures, differences and ratios we calculate based on our sample(s) are called **sample statistics**. The true value in the population we are trying to **estimate** is called a **population parameter**. We thus use sample statistics to advance our knowledge of population parameters, which we can never know for certain. A sample statistic is our **best guess** at the population parameter. Sample statistics are typically indicated with Roman letters, population parameters are denoted by Greek letters.

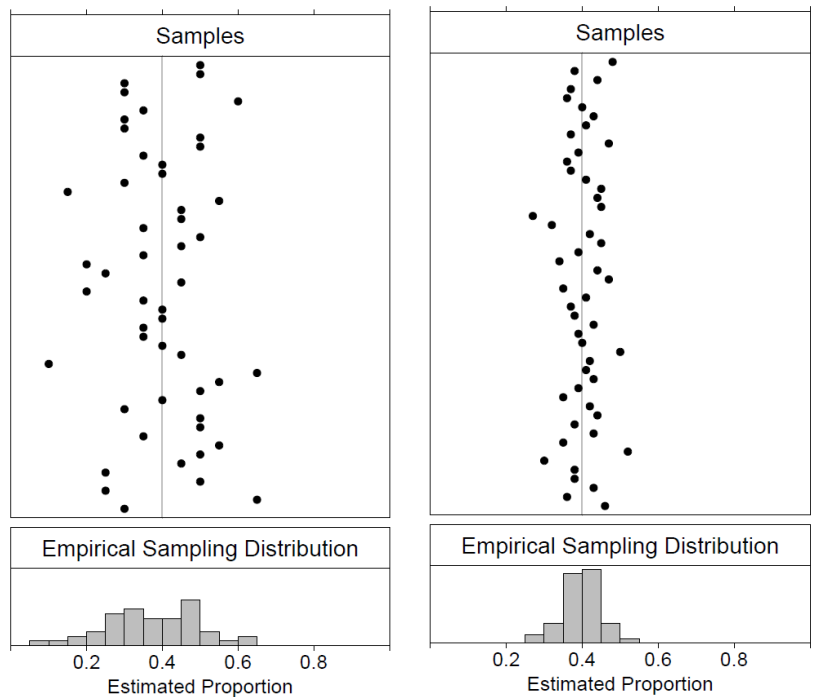
▪ Sampling variation

Sampling variation is the researcher's enemy. If we keep drawing samples from a population and calculate a sample statistic, this **sample statistic will vary from sample to sample**, and thus from study to study. Sampling variation is always involved when samples are used – we just don't see it because we have only drawn one of the many possible samples from this population. Sampling variation can be illustrated via simulation. Such simulations are a useful didactic tool. However, for illustrative purposes they assume that we know the true population parameter. This is not the case in reality. Think of the simulations as something that some higher power might witness when watching researchers down on earth doing their job. The diagrams below each show sample statistics from 50 different random samples drawn from the same population and estimating the same population parameter, in this case a proportion/percentage.

The true proportion is .4 (40%). There is variation between the 50 samples in both plots. This is called **sampling variation**. The histograms on the bottom show the distribution of the sample statistics.

The crucial difference between the two plots is that in the left plot the size of all samples was 20 and the right plot it was 100. The sampling variation in the right plot is much smaller. This is because **larger samples produce sample statistics that are on average closer to the true population parameter**.

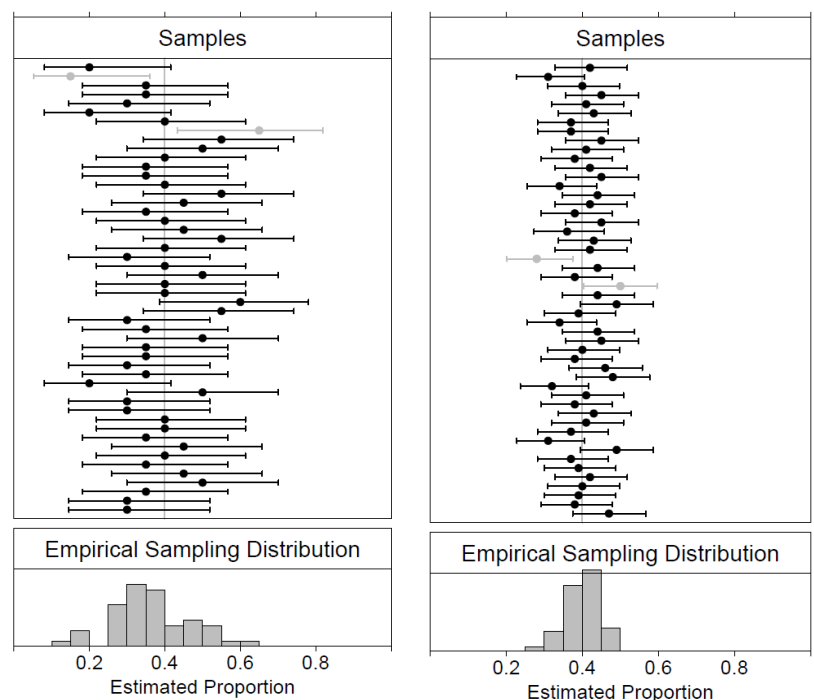
Remember that your study produces one of those dots, and it is very likely that you over- or underestimate the true population parameter.



▪ Confidence intervals

Confidence intervals (CIs) take into account sampling variation and tell us how much we can rely on our estimate, i.e. our sample statistic. It is again easiest to explain confidence intervals via simulation. We will take the two examples from above and show new estimates (sample statistics) with 95% CIs.

The confidence intervals in the left plot are much wider, which indicates the lack of precision of these estimates. Sample statistics based on small samples are less precise than those based on bigger samples. The idea behind a 95% confidence interval is the following: if we keep on sampling (or simulating), then in the long run 95% of the confidence intervals will capture the true population parameter. You can see that in each plot 2 out of 50 intervals miss the true value of 0.4. These are shown in grey. Of course, it is possible to use other levels of confidence. The interpretation changes accordingly. Due to convention, the 95% level is used most frequently.



3.2. Interpretation of confidence intervals

In practice, you would calculate and report a confidence interval for the statistical measures that give a quantitative answer to your research question(s). Depending on your research design this could be a relative frequency/proportion/mean/correlation coefficient/frequency ratio/proportion difference /mean difference/difference between correlation coefficients. The appendix explains how confidence intervals for these measures are calculated. The Excel module, however, makes such calculations much easier.

But let us now focus on a very important question: How do we interpret a 95% confidence interval? There are several ways of interpreting it, which foreground different aspects. Note that it is strictly speaking not correct to say that the 95% CI contains the population parameter with a probability of 0.95 (95%). A compromise is to say that we are **95% confident** that the parameter captures the true parameter. Here are three interpretations:

- Interpretation 1: Plausible values for the population parameter

The 95% CI covers a range of values. We can say that these values are plausible values for the population parameter. In other words, based on our sample and a confidence level of 95%, we cannot reject these values as candidates for the population parameter.

- Interpretation 2: Precision of our estimate

We can interpret the 95% CI as indicating the precision of our estimate. If it is wide, our estimate is not very precise and thus not very useful. If the 95% CI is narrow and covers a small range of values, we know that our sample statistic is precise and thus a reliable estimate of the population parameter.

- Interpretation 3: Relationship to statistical significance (p-values)

If the 95% CI does not include a certain value X, then – assuming we would have carried out a null hypothesis significance test – the difference between the sample statistic and the value X would have been statistically significant. For the comparison of groups (for example, with a frequency ratio, a difference of proportions, or a mean difference), this means that if the 95% CI does not include 0 as a plausible value, the difference between the groups is also statistically significant. CIs thus imply the result of a significance test.

3.3. Reporting confidence intervals

Confidence intervals are reported in the text using **square brackets**, with the confidence limits separated by a semicolon. If you are consistent in the level of confidence reported (e.g. 95%) it is enough to mention the level of confidence when reporting the first confidence interval.

Example: “The adverb *well* occurred 8.7 times more often in spoken than in written English, **95% CI [6.0; 12.5]**. The frequency ratio for *very* was **2.4 [1.6; 3.8]**.”

In plots, confidence intervals are shown using **error bars**. **Dot plots** are arguably the best choice for showing sample statistics and their confidence intervals. What you need to pay attention to is the fact that **error bars in diagrams may show different things**. Sometimes they are used to show the standard deviation or the standard error (which is in fact a 68% confidence interval). Therefore you **always** need to **explain** what the error bars in your charts show.

3.4. Understanding confidence intervals: Some properties

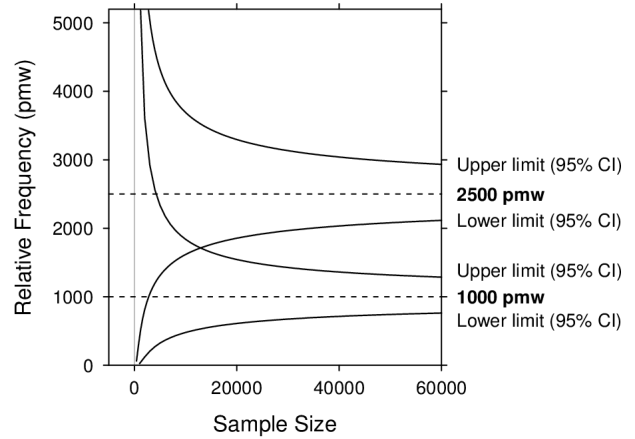
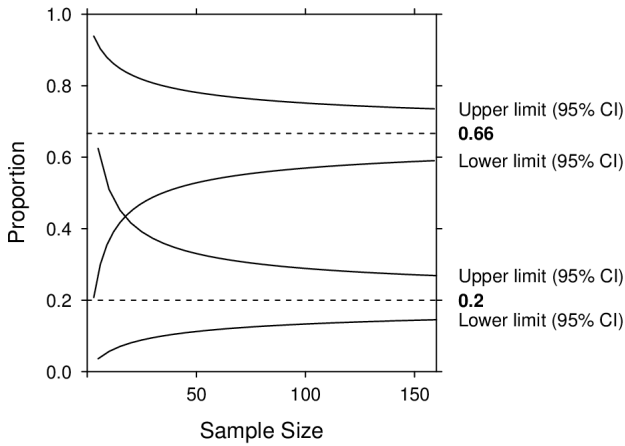
- Symmetry

Don't be surprised if a confidence interval is not symmetric around the point estimate.

Sample statistic	Symmetric?	Notes (technical)
Single sample		
Relative frequency	No	Approaches symmetry when $n \rightarrow \infty$
Proportion/percentage	No	Only when $p = .5$ (50%)
Mean	Yes	
Correlation coefficient	No	Only when $r = 0$
Comparing groups		
Frequency ratio	No	Symmetric on the log scale
Proportion difference	No	
Mean difference	Yes	
Difference between correlation coefficients	No	

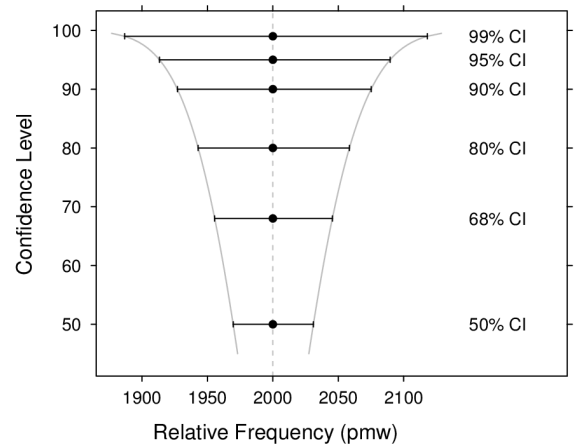
▪ **Width of confidence intervals: Effect of sample size**

Above we saw that larger samples yield more precise estimates. The following plots show the relationship between sample size and the width of a 95% CI for two proportions (on the left: 0.2 = 20% and 0.66 = 66%) and two relative frequencies (on the right: 1000 pmw and 2500 pmw). As you can see, the gain in precision is especially notable in the range of small sample sizes. For proportions, gains are large up to a sample size of 40 to 50, then the increase in precision levels off. The same is true for relative frequencies. Precision increases substantially for a sample size up to 15,000 to 20,000.



▪ **Width of confidence intervals: Effect of confidence level**

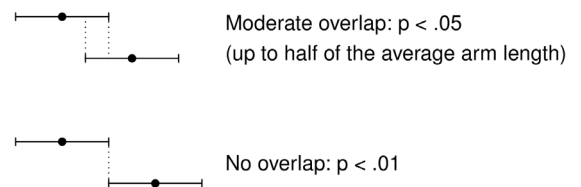
By convention, confidence intervals usually report a 95% level of confidence. If we require a higher level of confidence, the interval will be wider. The relationship between level of confidence and CI width is shown in the plot on the right for a relative frequency (pmw). The relationship has the shape of a trumpet. Higher levels of confidence come at an increasingly higher price in terms of precision/width. The 68% confidence interval corresponds to the standard error. If you see a plot showing +/- 1 standard error, you will know that you can interpret this interval as a 68% confidence interval.



▪ **Separate confidence intervals for groups: Interpretation of overlap**

It is best to analyze and communicate **targeted comparisons** by calculating and showing a measure that reflects this specific comparison. This may be a frequency ratio, a proportion difference, or a difference between means, accompanied by a 95% CI. Confidence intervals were created for this purpose, i.e. to show the precision of a single sample statistic, taking into account sampling variation. Sometimes, however, we are only provided with 95% CIs for two group individually. A common misconception when comparing two such confidence intervals is that the difference between two groups is not statistically significant if the error bars overlap. This is not true. In fact, if the overlap is moderate (up to half of the average arm length), the p -value for the comparison would be $p < .05$. For CIs that do not overlap, it would be $p < .01$.

This rule is restricted to cases when there are only 2 groups involved. If we look at several CIs simultaneously (e.g. for 3 or more groups), we should not be tempted to compare them and make statements about statistical significance. Each CI answers one specific question and thus serves one and only one purpose: to indicate the precision of a point estimate (sample statistic).



4. The new statistics in practice: Effect sizes and confidence intervals

4.1. The new statistics vs. significance testing

We briefly mentioned the practice of null hypothesis significance testing (NHST) above. This approach to the analysis and presentation of research results has been heavily criticized in the past. From a CI-perspective, NHST only asks whether – based on our sample – X (most often 0) is a plausible value for the population parameter. In NHST you perform a test that outputs a p -value. A p -value smaller than .05 is interpreted as signifying a **statistically significant** result.

Consider the comparison of two groups. If the p -value is below .05, the difference between the groups is claimed to be “statistically significant”. The problem with the p -value is that it combines two types of information: the magnitude of the difference found and the precision of this estimate. We know that larger samples yield higher precision, so a small p -value might reflect a large sample, not necessarily a large difference. We surely would like to know, because otherwise the adjective *significant* is misleading.

The “new statistics” take a different approach to the interpretation of research results. We first look at the difference and then construct a CI to see how precise our estimate of this difference is. The focus is where it ought to be: on the observed difference. As mentioned above, the result of a NHST is implied by a 95% CI.

A more general name for statistics such as the frequency ratio, the difference of proportions, the difference between means and the correlation coefficient, is **effect size**. The term covers a wide range of statistical measures and has been defined as “a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest” (Kelley and Preacher 2012: 140).

4.2. Why the new statistics are better

- The new statistics are in fact not new – most of the methods they apply have been around as long as NHST. What *would* be new is if they were used more widely by researchers.
- The new statistics speak **a language we understand**, and which our audience understands. While some types of effect sizes may seem cryptic to student researchers, the ones outlined above (frequency ratio, difference of proportions, mean difference) are easy to understand and interpret. Some practice is probably needed with the correlation coefficient.
- The idea behind a confidence interval is much **more intuitive** and easier to understand than the idea behind NHST and statistical significance. All you need to know is (i) how to interpret the statistical measure you are focusing on and (ii) how to interpret the confidence interval.
- Since the focus is on the effect size at every stage of the analysis, there is **less danger of misinterpreting a result** that is significant in the statistical sense as one that is meaningful in the practical sense.
- Effect sizes and confidence intervals give us much **more information** than a p -value. They tell us about the magnitude of an effect and its precision, not just whether it exists or not.
- Effect sizes and confidence intervals also make it easier to **compare results across studies** and to summarize the state of research on a topic. We can compare the effect we have found to those obtained in other studies. This is likely to yield a more substantial discussion of results.
- The **synthesis of results in the literature** is much more informative when focusing on effect sizes with confidence intervals than by merely counting votes (statistically significant vs. not statistically significant). A graphical presentation of the effect sizes obtained in previous studies will lead to more informed comparisons, raise new questions and provide a solid basis for the interpretation and contextualization of the results of your study.
- This is in fact the point at which **meta-analysis** starts. Meta-analysis aims at combining the evidence reported in different studies into a more precise estimate of the phenomenon of interest. Meta-analysis usually uses dot plots to provide a graphical summary of the existing evidence. Such plots are called forest plots.

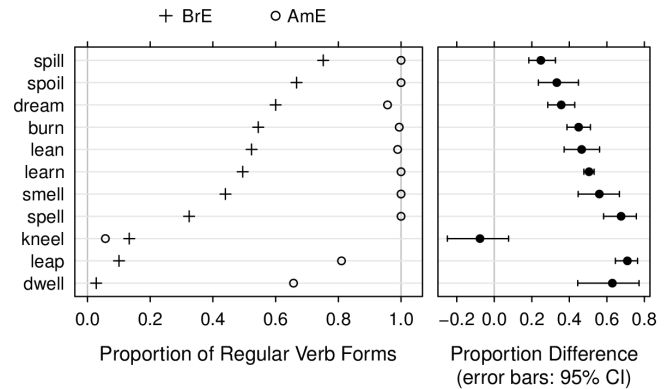
4.3. Application to corpus data

Dot plots are very suitable for showing error bars. Appending a panel to directly show differences or ratios appears to be a fruitful method for the analysis and presentation of corpus data. This method was proposed by Sönning (2014), but he does not know whether he was the first to do so. In any case, it deserves wide usage in corpus linguistics and can now be easily applied using the Excel module. 95% CIs can be added to the appended panel to show how precise the estimate of the differences/ratios are.

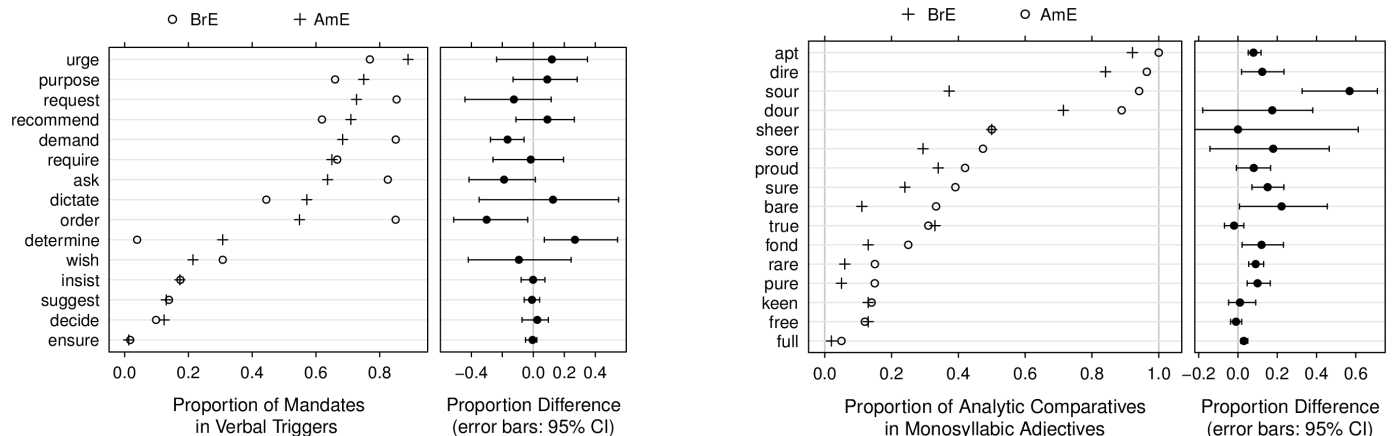
Binary data: Proportions/percentages

The plot on the right shows the proportion of regular simple past forms for 11 verbs in two varieties: British and American English (data from Levin 2009). The left panel shows the proportion for each variety, the right panel plots the difference between the proportions. The error bars show 95% confidence intervals.

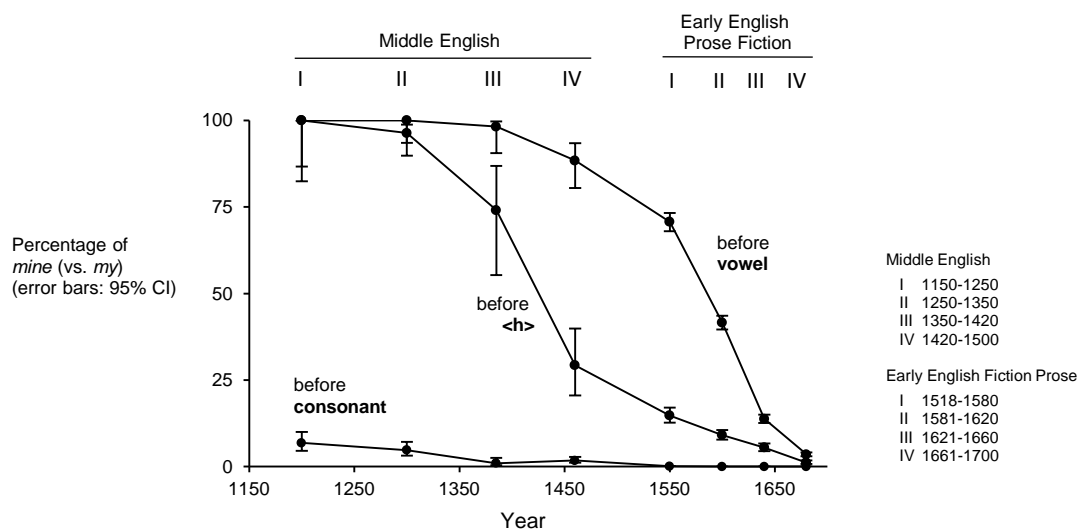
Such two-fold comparisons are typical for corpus data: we compare (i) **items**, here: 11 verbs and (ii) **groups**, here: 2 varieties.



Two further examples of this use of dot plots with binary variables shall be given here. The data are from Crawford (2009) and Mondorf (2009). The interested reader is referred to the original research articles.

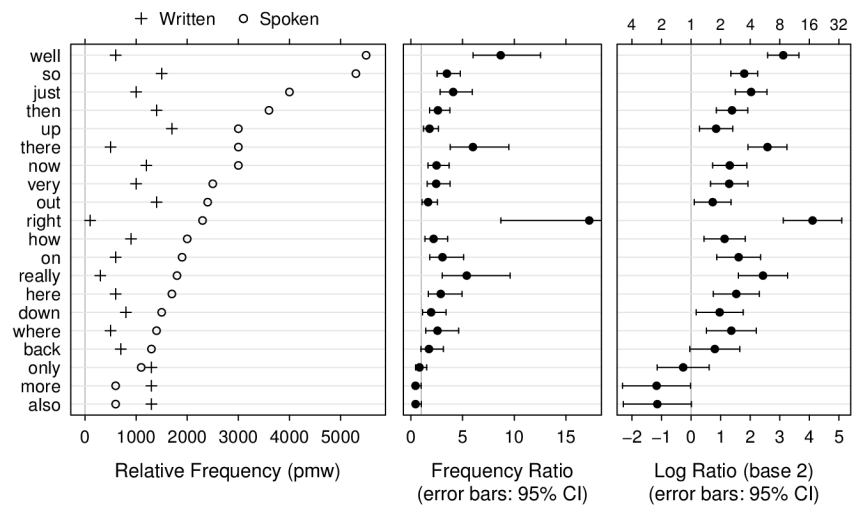


This example for binary data includes time as a variable (data from Schlüter 2009). Line plots are suited best for showing such diachronic trends. The underlying variable on the y-axis is binary (preceded by *my* vs. *mine*), and we are comparing three groups, here three different phonetic contexts (before C, V, or <h>).



Frequency data

Two-fold comparisons are also typical for corpus-based frequency data. The plot on the right uses the adverb example from above, comparing the frequency of the top 20 adverbs in spoken and written English. It is possible to **append more than one panel** thus expressing the comparison in different ways. For frequency data, we might append panels showing frequency differences and frequency ratios. In the plot below, the frequency ratio and the log ratio are shown.



The **log ratio** is another measure for frequency comparisons. It has one disadvantage and two important advantages. The **disadvantage** is that most people (and audiences) are unfamiliar with the log scale. It helps to show the original (unlogged) ratios on the upper axis of the panel. This was done here. The panel on the right shows the \log_2 ratio (i.e. logarithms of the ratio with base 2). A log ratio of 2 corresponds to a ratio of $2^2 = 2 \times 2 = 4$. A log ratio of 4 corresponds to a ratio of $2^4 = 2 \times 2 \times 2 \times 2 = 16$. The ratios can be read off the top axis of the panel: *well* is around 8 times as frequent in spoken compared to written English. The **advantages** outweigh the disadvantages. Original ratios have a skewed distribution: the majority of scores is squashed into the left side of the plot, a few large ratios dominate the plot and thus mess up the **resolution**. This makes it is harder to compare the smaller ratios. Further, **ratios below 1** (here: adverbs that were more frequent in written English) are always crammed into the interval $[0;1]$ and thus overlooked. The choice of which frequency to divide by which has drastic consequences for the story the plot tells. This is clearly undesirable. Log ratios do not suffer from these drawbacks. Ratios in both directions receive equal weight and resolution is no longer a problem.

The log ratio has recently been proposed as a measure for use in keyness analyses to avoid sole reliance on test statistics derived from NHST, such as the Log-likelihood ratio test (Hardie 2014). This is clearly a welcome impulse. Keyness analyses would arguably also profit from the graphical methods presented here, as they combine descriptive and inferential information, with the visual presentation uncovering unexpected patterns and facilitating cross-item comparisons in terms of keyness. It then appear reasonable to apply a data-based ordering of the items used for keyness analysis, either according to value (if we are only interested in items that are more frequent in corpus A) or according to absolute value (if we are also interested in which items are underrepresented in corpus A). Different ways of ordering are likely to reveal different aspects of the data and should be a common strategy at the stage of data analysis.

5. Further reading

The information provided in this booklet should be enough to help you calculate CIs for the most frequently encountered simple research designs. The Excel module makes calculations and plotting much easier.

Thompson (2002), Schmidt (1996) - Short articles on why researchers should abandon NHST and how they would benefit from a change towards the use of effect sizes and confidence intervals.

Altman et al. (2000) - Simple explanations on how to calculate CIs for a wide range of statistics.

Cumming (2012) - Probably the best book on the ideas behind the new statistics. It is accompanied by Excel applications that allow you to explore ideas and concepts interactively. The focus is on quantitative data. This book inspired me in preparing this talk. Includes an introduction to meta-analysis.

Kline (2013) - The definite reference on the debate about statistics reform in the behavioral sciences. Includes modern applications such as robust statistical measures and an introduction to Bayesian statistics.

Grissom and Kim (2012), Ellis (2010) - Two great references on effect sizes. Grissom and Kim 2012 is much more detailed and technical.

6. Constructing confidence intervals

6.1. Frequency data

▪ Confidence interval for a (relative) frequency

The confidence interval is calculated using the raw frequency k . The confidence limits you calculate are then also in raw (absolute) counts. Afterwards, you can convert them to a relative frequency (such as per million words). The upper and lower limits of the 95% CI for a raw frequency are:

$$\text{Lower limit} = (1.96 + \sqrt{k + 0.02})^2$$

$$\text{Upper limit} = (1.96 + \sqrt{k + 0.96})^2$$

Source: Bégau et al. 2005
R: `function poisson.test()`

▪ Confidence interval for the ratio of two relative frequencies (rates)

Frequencies are typically compared using a ratio. We can compute a confidence interval for the ratio of two frequencies a and b in samples of size M and N , respectively. The upper and lower limits of the 95% CI for the ratio of the two relative frequencies (rates) are calculated as follows:

$$\text{Lower limit} = \frac{N}{2Mb^2} \times \left\{ 2ab + 1.96^2(a + b) - 1.96 \times \sqrt{(a + b)[4ab + 1.96^2(a + b)]} \right\}$$

$$\text{Upper limit} = \frac{N}{2Mb^2} \times \left\{ 2ab + 1.96^2(a + b) + 1.96 \times \sqrt{(a + b)[4ab + 1.96^2(a + b)]} \right\}$$

Source: Graham et al. 2003; Siev 1994
R: `function riskscoreci()` in the package PropCIs

6.2. Proportions (Binary variables)

▪ Confidence interval for a proportion (percentage)

For a binary variable, we can express the distribution of observations by giving the proportion (or percentage) of one of the two categories C_1 and C_2 . The choice of category will not influence the results. If 12 out of 20 observations were in category C_1 , we could give the proportion as 0.6 (or 60%). Statisticians use the generic terms “successes” (here: 12) and “failures” (here: 8). The following method calculates the 95% confidence interval for this proportion.

Symbols:	r	observed number of successes (here: 12)
	q	observed proportion of failures (here: $8/20 = 0.4$)
	n	total number of observations (here: 20)

First calculate three quantities: $A = 2r + 1.96^2$ $B = 1.96\sqrt{1.96^2 + 4rq}$ $C = 2(n + 1.96^2)$

The 95% confidence interval is then: $\text{Lower limit} = \frac{(A-B)}{C}$ $\text{Upper limit} = \frac{(A+B)}{C}$

Source: Newcombe and Gardner 2000
R: `function scoreci()` in the package PropCIs

▪ **Confidence interval for the difference of proportions (or percentages)**

If you compare two groups and the dependent variable is binary, a simple and easy-to-understand method is to look at the difference between the proportions. Again, the choice of which category to compare does not influence the results. You can calculate a 95% CI for the difference of proportions as follows. First you find the upper and lower confidence interval limits of the proportion in each group using the method described above. It does not matter which group is group 1 or group 2 as long as you are consistent. This yields:

	l_1	lower limit for group 1	u_1	upper limit for group 1
	l_2	lower limit for group 2	u_2	upper limit for group 2
Symbols:	p_1	proportion of “successes” in group 1		
	p_2	proportion of “successes” in group 2		
	D	difference of proportions ($p_1 - p_2$)		

The confidence interval for the difference of proportions is then:

$$Upper\ limit = D + \sqrt{(p_2 - l_2)^2 + (u_1 - p_1)^2}$$

$$Lower\ limit = D - \sqrt{(p_1 - l_1)^2 + (u_2 - p_2)^2}$$

Source: Newcombe and Gardner 2000
R: function `diffscoreci()` in the package `PropCIs`

6.3. Quantitative data

Overview

Data	Central tendency		Correlation	
	Single group	Comparing groups	Single group	Comparing groups
Normal	Mean + CI	Difference between means + CI	Pearson correlation coefficient + CI	Difference between Pearson correlation coefficients + CI
Non-normal	Median + CI Trimmed mean + CI	Difference between trimmed means + CI	Spearman correlation coefficient + CI	Difference between Spearman correlation coefficients + CI

▪ **Confidence interval for the mean**

To calculate the 95% confidence interval for the mean we first need to calculate the standard error. It is best to let Excel calculate the standard deviation (Function: STABW.N) for us. Here is the formula for the standard error:

$$SE = \frac{SD}{\sqrt{n}}$$

The 95% confidence interval is then given by the following formulas.

$$Lower\ limit = Mean - (t \times SE)$$

$$Upper\ limit = Mean + (t \times SE)$$

The value t depends on the sample size n , more precisely on df which here is $(n - 1)$. You can take an approximate value from the following table:

df	6	8	10	12	14	16	18	20	25	30	35	40	50	>60
t	2.45	2.31	2.23	2.18	2.15	2.12	2.10	2.09	2.06	2.04	2.03	2.02	2.01	2

R: function `t.test()`

▪ **Confidence interval for the median**

It is fairly easy to calculate a confidence interval for the median. Calculate two values A and B:

$$A = \frac{n}{2} - \left(1.96 \times \frac{\sqrt{n}}{2}\right) \qquad B = 1 + \frac{n}{2} + \left(1.96 \times \frac{\sqrt{n}}{2}\right)$$

where n is the sample size. Round A and B to the nearest whole number. Order the observations (values) in your sample. The A^{th} and B^{th} observations form the 95% CI for the median.

▪ **Confidence interval for the 20% trimmed mean (Tukey-McLaughlin method)**

Trimmed means are not used very frequently (yet). However, they have a number of advantages compared to e.g. medians (for us, a practical advantage is the fact that it is easier to calculate a CI for the difference between two trimmed means than it is to calculate a CI for the difference between two medians).

We first need to **calculate a trimmed mean**. It is common practice to use 20% trimming. Order the scores from lowest to highest. Remove the highest 20% and lowest 20% of the scores as follows: Multiply your sample size by 0.2 and round the result *down* to the nearest whole number A . Remove the A highest and A lowest scores and calculate the mean of the remaining 60% of the scores.

Then you need to calculate what is called the Winsorized variance. Excel can do this for you, but you first need to prepare the scores – you need to “Winsorize” them. Basically, you do the same as before when you calculated a trimmed mean. However, you don’t remove the highest 20% and lowest 20% of the scores, but you replace them with the highest or lowest score that remains after trimming, respectively. Then you use the Excel function VAR.P to calculate the variance of these Winsorized scores s_{win}^2 . Take the square root to obtain the Winsorized standard deviation s_{win} . Then you calculate the standard error of the trimmed mean as follows:

$$SE_{TM} = \frac{s_{win}}{0.6\sqrt{n}}$$

The 95% confidence interval for the 20% trimmed mean is then given by the following formulas.

$$Lower\ limit = M_{tr} - SE_{TM}(t \times n_{tr})$$

$$Upper\ limit = M_{tr} + SE_{TM}(t \times n_{tr})$$

where n_{tr} is the number of scores that remain after trimming and the value for t depends on the sample size (more precisely, df , here: $n - 1$), which you can look up in the table above.

Method: Tukey-McLaughlin method (Tukey and McLaughlin 1963)
R: function `trimci()` in the package `WRS`

▪ **Confidence interval for the difference between two means**

Two groups can be compared by looking at the difference between the means. The following method for calculating a 95% CI for the difference between means not only relies on normally distributed measurements but also on equal dispersion/variability in each group. Compare the standard deviations of the two groups and boxplots to decide whether this assumption is met. If the dispersion of scores in the two samples does not differ drastically you can go ahead using this method. Otherwise use the method for trimmed means described below. First calculate a pooled standard deviation:

Symbols:	n_1	sample size group 1	M_1	mean group 1
	n_2	sample size group 2	M_2	mean group 2
	s_1	standard deviation group 1		
	s_2	standard deviation group 2		

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Then you calculate a standard error for the difference between the two means:

$$SE_{Diff} = s_{pooled} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The 95% confidence interval for the difference between the means $M_{Diff} (M_1 - M_2)$:

$$Lower\ limit = M_{Diff} - (t \times SE_{Diff})$$

$$Upper\ limit = M_{Diff} + (t \times SE_{Diff})$$

The value t depends on the sample sizes. Refer to the table above, where approximate values are given. In this case df is determined by $n_1 + n_2 - 2$.

R: function `t.test()`

▪ Confidence interval for the difference between trimmed means (Yuen-Welch procedure)

First you need to calculate the Winsorized variance for both groups as explained above. Then you calculate an error variance w for each group:

$$w_1 = \frac{s_{Win}^2(n_1-1)}{n_{tr1}(n_{tr1}-1)} \quad w_2 = \frac{s_{Win}^2(n_2-1)}{n_{tr2}(n_{tr2}-1)}$$

where s_{Win}^2 is the Winsorized variance, n is the sample size in the respective group, n_{tr} is the number of scores that remain after trimming. Then calculate the standard error for the difference between the trimmed means ($M_{tr1} - M_{tr2}$):

$$SE_{Diff_{tr}} = \sqrt{w_1 + w_2}$$

The 95% confidence interval is then:

$$Lower\ limit = M_{tr1} - M_{tr2} - SE_{Diff_{tr}} \times t$$

$$Upper\ limit = M_{tr1} - M_{tr2} + SE_{Diff_{tr}} \times t$$

The value t depends on the sample size, more precisely df . The calculation of the correct df is a bit more complicated for this method:

$$df = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_{tr1} - 1} + \frac{w_2^2}{n_{tr2} - 1}}$$

Look up the t value in the table above and plug it into the formula.

Method: Yuen-Welch procedure (Yuen 1974)
R: function `yuen()` in the package `WRS`

▪ **Confidence interval for the Pearson correlation coefficient**

You can calculate the Pearson correlation coefficient in Excel with the function KORR. To obtain a 95% CI for the correlation coefficient you first obtain a value Z:

$$Z = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$$

Then you calculate two values F and G:

$$F = Z - \frac{1.96}{\sqrt{n-3}} \qquad G = Z + \frac{1.96}{\sqrt{n-3}}$$

The confidence interval limits are:

$$\text{Lower limit} = \frac{e^{2F} - 1}{e^{2F} + 1}$$

$$\text{Lower limit} = \frac{e^{2G} - 1}{e^{2G} + 1}$$

▪ **Confidence interval for the Spearman correlation coefficient**

Spearman's correlation coefficient is the Pearson correlation coefficient of the ranked scores in both groups. You first rank the scores in each group separately and then use the function KORR to calculate the correlation coefficient of the ranks in Excel. A confidence interval for Spearman's correlation coefficient is then calculated the same way as described above for Pearson's correlation coefficient.

▪ **Confidence interval for the difference between two correlation coefficients**

If you compare two groups, or the results of your study with those obtained in another you can simply calculate the difference between the correlation coefficients. First you calculate a 95% confidence interval for each correlation coefficient separately using the instructions above. This yields:

l_1	lower limit for group 1	u_1	upper limit for group 1
l_2	lower limit for group 2	u_2	upper limit for group 2

Symbols:	r_1	correlation coefficient in group 1
	r_2	correlation coefficient in group 2
	D	difference between the correlation coefficients ($r_1 - r_2$)

The 95% confidence interval for the difference D ($r_1 - r_2$) is then:

$$\text{Lower limit} = D - \sqrt{(r_1 - l_1)^2 + (u_2 - r_2)^2}$$

$$\text{Upper limit} = D + \sqrt{(u_1 - r_1)^2 + (r_2 - l_2)^2}$$

References

- Altman, Douglas G., David Machin, Trevor N. Bryant and Martin J Gardner. 2000. *Statistics with confidence*. London: British Medical Journal books.
- Bégaud, Bernard, Karin Martin, Abdelilah Abouelfath, Pascale Tubert-Ritter, Nicholas Moore and Yola Moride. 2005. An easy to use method to approximate Poisson confidence limits. *European Journal of Epidemiology* 20: 213-216.
- Biber, Douglas and Bethany Gray. 2013. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41: 104-134.
- Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Crawford, William J. 2009. The mandative subjunctive. *One language, two grammars?*, edited by Günter Rohdenburg and Julia Schlüter. Cambridge: CUP. 257-276.
- Cumming, Geoff. 2012. *Understanding the new statistics*. New York: Routledge.
- Ellis, Paul D. 2010. *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York, NY: Cambridge University Press.
- Graham, P.L., K. Mengersen and A.P. Norton. Confidence limits for the ratio of two rates based on likelihood scores: non-iterative method. *Statistics in Medicine* 22: 2071-2083.
- Grissom, Robert J. and John J. Kim. 2012. *Effect sizes for research. Univariate and multivariate applications*. New York: Routledge.
- Hardie, Andrew. 2014. Log Ratio – an informal introduction. <http://cass.lancs.ac.uk/?p=1133>
- Kelley, Ken and Kristopher J. Preacher. 2012. On effect size. *Psychological Methods* 17, 137-152.
- Kline, Rex B. 2013. *Beyond significance testing*. Washington, DC: American Psychological Association.
- Levin, Magnus. 2009. The formation of the preterite and the past participle. *One language, two grammars?*, edited by Günter Rohdenburg and Julia Schlüter. Cambridge: CUP. 60-85.
- Mondorf, Britta. 2009. Synthetic and analytic comparatives. *One language, two grammars?*, edited by Günter Rohdenburg and Julia Schlüter. Cambridge: CUP. 86-107.
- Newcombe, Robert G. and Douglas G. Altman. 2000. Proportions and their differences. *Statistics with confidence*, edited by Douglas G. Altman, David Machin, Trevor N Bryant and Martin J Gardner. London: British Medical Journal books. 45-56.
- Schlüter, Julia. 2009. Coonsonant or ‘vowel’? A diachroni study of the status of initial <h> from early Middle English to nineteenth-century English. *Phonological weakness in English: From Old to Present-Day English*, edited by Donka Minkova. New Yor: Palgrave Macmillan. 168-196.
- Schmidt, Frank L. 1996. Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods* 1, 115-129.
- Siev, David. 1994. Estimating vaccine efficacy in prospective studies. *Preventive Veterinary Medicine* 20: 279–296.
- Sönning, Lukas. 2014. The dot plot: A fine tool for data visualization. Paper presented at *Advances in Visual Methods for Linguistics 2014*. Tübingen.
- Thompson, Bruce. 2002. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher* 31, 25-31.
- Zou, Guang Y. 2007. Toward using confidence intervals to compare correlations. *Psychological Methods* 12, 399-413.