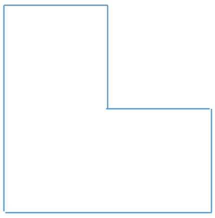
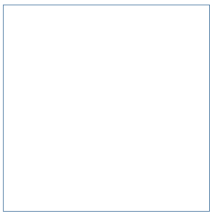
Chapter 1: Exercises

1) As a warm-up exercise (with a twist), divide the following shape that represents three quarters of a square into *four* identical shapes. Feel free to skip this exercise if you want to focus on statistical techniques immediately.



After you have done this, take a whole square, but this time divide it into *five* identical shapes.



2) Calculate the mean for the following numbers: 2339, 2089, 2056, 2276, 2233, 2056, 2241, 1995, 2043, 1976, 2062. These are the frequencies of verbs in fiction texts by British writers discussed in this section 1.2.

3) What is a model in scientific thinking?

4) Select the best-fitting geometrical model for the area of Great Britain (see Figure 1.20) that would help you investigate the area of the island.

a) rectangle □
b) circle ○
c) triangle △

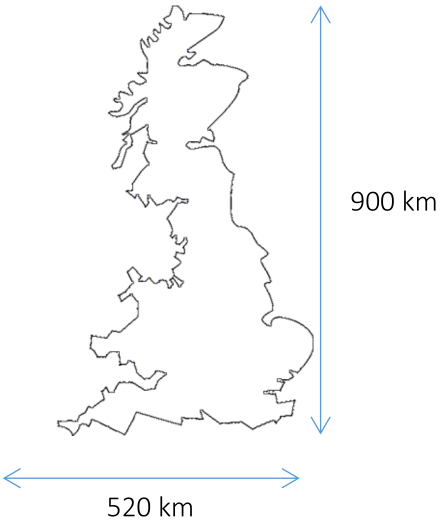


Figure 1.20 Great Britain: Main island

5) Use the model and the dimensions given in Figure 1.20 to calculate the area of Great Britain (the large island only).

6) Test your knowledge about the basic statistical terminology in a short quiz.

1. What is the difference between the *average* and the *mean*?
   a) There is no difference; these terms are synonyms.
   b) Mean is a type of average; so are the *median* and *trimmed mean.*
   c) Mean is usually larger than the average.

2. What is the *mean* of the following values: 5, 10, 15, 20, 25?
   a) 15.0
   b) 17.32
   c) 25.3

3. What is the *median* of the following values: 5, 10, 15, 20, 25?
   a) 10
   b) 15
   c) 20

4. What type of variable is the rank of words in a frequency list?
   a) nominal
   b) ordinal
   c) scale

5. What is dispersion in a corpus?
   a) The distribution of a (linguistic) variable in different parts of a corpus.
   b) Another term for standard deviation.
   c) Spread of a sample in a population.

6. If you plot a normally distributed set of data, what will the shape of the graph be?
   a) Flat
   b) J-curve
   c) Bell-shaped

7. What is a 95% confidence interval?
   a) Interval that shows that we can be 95% confident in the correctness of the result within this interval.
   b) The measure of objectivity of our findings.
   c) Interval constructed around a particular measure in a sample in such a way that the true value of the measure in the population will fall within this interval for 95% of samples.

8. What is a p-value?
   a) The probability that the null-hypothesis is true.
   b) The probability of seeing values at least as extreme as observed if the null-hypothesis were true.
   c) The probability of seeing a unicorn in Lancaster.

7) Imagine you have 500 texts (the population of interest), 250 written by a male and 250 written by a female author. However, you don't know which one is which. For the purposes of your study you want to select an unbiased sample of 40 texts that would represent the population. Use the Random number generator from the Lancaster Stats Tools online to create a list of 40 random numbers between 1 and 500 and note these down.

____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____

____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____

____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ ____

8) In the Answer section at the Companion website, you can find which texts were written by a male speaker and which by a female speaker. Check the answers in Exercise 7 and calculate the number of male and female speakers that were selected.

Male speakers in the sample:________________
Female speakers in the sample: ________________

Did you get an approximately equal representation of male and female speakers in the sample?

9) Later, you also find out that half of the 500 texts were written by a young speaker and another half by an older speaker (see the Answer section at the Companion website to find out which texts these are). Review the sample from Exercise 7 to find out if you have an approximately equal gender and age representation in your sample of 40 texts.

	male speaker	female speaker
young speaker		
older speaker		

10) Was random sampling a successful method? Why (not)?

11) What types of bias do we need to avoid in corpus design?

12) What type of research design (Whole corpus, Linguistic feature, Individual text/speaker) would you use in the following situations?

a) To find out if zero relativizer in relative clauses (e.g. *The second thing Ø I want to say is...*) is more frequent in spoken or written language.

b) To find out if hedges (e.g. *sort of, kind of*) are more common in male or female speech.

c) To find out the frequency of the definite article *the* in written English.

d) To investigate register-based variation of a large number of linguistic features including modals, discourse markers and private verbs.

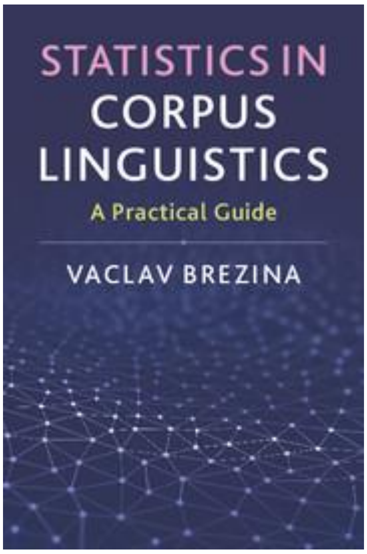
13) Can you spot six errors in the following dataset based on BE06, an approximately one-million-word corpus of written British English?

Word or expression	Frequency	Frequency per million
the	5,896	5,142.17
of	30,666	26,745.23
and	27,909	24,340.72
to	26,188	2,283.98
of the	6,887	6,006.47
and the	19,530	17,033.01
Words total	2,293,194	

14) Which visualization type (graph) would be appropriate in the following situations?

a) Describing the frequency distribution of a linguistic variable in one corpus.
b) Comparing the distribution of a linguistic variable in two corpora.
c) Finding the relationship between two linguistic variables.
d) Estimating if the differences between the frequencies of a linguistic variable can be generalised to the population.

15) Use the Graph tool from the Lancaster Stats Tools online and the data provided there to create graphs visualizing the main patterns in those datasets.



STATISTICS IN CORPUS LINGUISTICS

A Practical Guide

VACLAV BREZINA

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge: Cambridge University Press.

Do you use language corpora in your research or study, but find that you struggle with statistics? This practical introduction will equip you to understand the key principles of statistical thinking and apply these concepts to your own research, without the need for prior statistical knowledge. The book gives step-by-step guidance through the process of statistical analysis and provides multiple examples of how statistical techniques can be used to analyse and visualise linguistic data. It also includes a useful selection of discussion questions and exercises which you can use to check your understanding.

The book comes with a Companion website, which provides additional materials (answers to exercises, datasets, advanced materials, teaching slides etc.) and Lancaster Stats Tools online, a free click-and-analyse statistical tool for easy calculation of the statistical measures discussed in the book.