Statistical Analysis of Corpus Data with R
— Exercise Sheet for Unit #3 —

In this exercise, you will use techniques of descriptive and inferential statistics for continuous data to analyze the frequencies of passive verb phrases across texts from the extended *Brown Family* of corpora (for details, see Xiao 2008, 395–397). These corpora are collections of edited written American English (AmE, published in the U.S.) and British English (BrE, published in the UK) from the 1930s, 1960s and 1990s. Here, we focus on the four central corpora: Brown (AmE, 1960), LOB (BrE, 1960), Frown (AmE, 1990) and FLOB (BrE, 1990).

Your goal is to determine whether there are significant differences between the frequency of passives in American and British English, and whether there is evidence for a decline between the 1960s and 1990s as claimed by Leech et al. (2009, 164).

The data set `PassiveBrownFam` included in the `SIGIL` package provides separate counts of passive and active verb phrases for each text in the extended Brown Family, as well as precomputed percentages of passive VPs. Check that you have access to this data set and familiarize yourself with its structure:

```
> library(SIGIL)
> nrow(PassiveBrownFam)    # check that data set is available
[1] 2499
> ?PassiveBrownFam         # read documentation carefully
```

Now answer the following questions and carry out the specified tasks:

1. How many texts are there in each corpus of the extended Brown Family? What is their distribution across different genres?
   - Display tables of text counts by corpus and by genre.
   - Is the genre distribution the same for each corpus?
   - *optional:* visualize the distributions with suitable bar plots
2. Use the `subset` command to obtain separate data sets for the four target corpora.
3. Summarize the distribution of passive frequencies (i.e. percentages of passive VPs) across texts in each corpus.
   - Compute the mean μ and standard deviation σ of passive frequencies.
   - Which other summary statistics might be useful?
4. Visualize the distribution of passive frequencies.
   - Plot a histogram for each corpus.
   - Add a contour line showing the estimated density function.
   - *optional:* Can you combine all four histograms and/or contours into a single plot?
5. Do the frequency counts follow a Gaussian distribution (at least roughly)?
   - Add suitable Gaussian density curves to the histogram plots.
   - Use quantile-quantile plots to assess normality of the distribution.

6. Use Student's *t*-test for a pairwise comparison of the four corpora.
    - Which pairs of corpora can meaningfully be compared?
    - Which version of the *t*-test is appropriate in this situation?
    - Specify the precise null hypothesis H_0 of this test.
    - Report and interpret the full results of the significance tests.
    - *optional:* Carry out non-parametric Mann-Whitney tests (using the `wilcox.test` function), and compare the two sets of results.

7. Discuss the results of your analysis.
    - Is there a significant difference between American and British English?
    - Do your data corroborate Leech et al.'s (2009) claim of a decline in passive use?
    - What linguistic conclusions can you draw from your observations?

8. Do passive frequencies vary between different genres?
    - Use the "formula" interface of `boxplot` to display side-by-side boxplots of passive frequencies in the different genres.
    - Which significance test can be used to assess whether there are genuine differences between genres? What is its precise null hypothesis H_0?
    - Why would it be problematic to carry out pairwise comparisons for all genre combinations with Student's *t*-test?
    - *optional:* Use a suitable adjustment or other procedure to determine which genres are significantly different from each other.

9. How meaningful are frequency comparisons between corpora?
    - Discuss the implications of significant genre differences for the corpus-level comparisons you have carried out above.

References

Leech, Geoffrey; Hundt, Marianne; Mair, Christian; Smith, Nicholas (2009). *Change in Contemporary English: A Grammatical Study*. Studies in English Language. Cambridge University Press, Cambridge.

Xiao, Richard (2008). Well-known and influential corpora. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, chapter 20, pages 383–457. Mouton de Gruyter, Berlin.