Statistical Methods Usage of Descriptive Statistics in Corpus Linguistic

Valeriy Didusov and Zoia Kochueva

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

valeradidusiov@gmail.com, kochueva@kochuev.com

The purpose of this study: the development of software to determine the criteria of Pearson's chi-squared on the basis of both scientific and literary texts. In the first part of the study it is conducted the analysis of contemporary corpus linguistics. Statistical studies and their views of corpus linguistics are also described. Held his own description of the body of fiction and non-fiction texts. In the second part of the paper it is focused on statistical methods in corpus linguistics, namely the descriptive statistics. The description of the operation of the developed software.

Research in various linguistic areas has the subject a text or collection of texts and imply at first selection of the material, and then analysis and processing of large amounts of text to identify some language patterns. Traditional methods of linguistic analysis of the text can perform all of these tasks, but their low efficiency makes the usage of methods of computer analysis of the text more frequent. This reduces the work of linguist and considerably increase the amount of processing data and avoid inaccuracies and errors in calculations. A computer text analysis enables the establishment of speech patterns based not on theoretical but empirical data [1].

A statistical hypothesis is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets. The comparison is deemed statistically significant if the relationship between the data sets would be an unlikely realization of the null hypothesis according to a threshold probability – the significance level. Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance. The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by identifying two conceptual types of errors (type 1 & type 2) [2].

The major purpose of hypothesis testing is to choose between two competing hypotheses about the value of a population parameter. For example, one hypothesis might claim that the wages of men and women are equal, while the alternative might claim that men make more than women.

The hypothesis actually to be tested is usually given the symbol H_0, and is commonly referred to as the null hypothesis. As is explained more below, the null hypothesis is assumed to be true unless there is strong evidence to the contrary –

similar to how a person is assumed to be innocent until proven guilty. The other hypothesis, which is assumed to be true when the null hypothesis is false, is referred to as the alternative hypothesis, and is often symbolized by H_A or H_1. Both the null and alternative hypothesis should be stated before any statistical test of significance is conducted. In other words, you technically are not supposed to do the data analysis first and then decide on the hypotheses afterwards [3].

Algorithm:

1. State the null hypothesis and the alternate hypothesis.
2. Select the appropriate test statistic and level of significance.
3. State the decision rules.

The decision rules state the conditions under which the null hypothesis will be accepted or rejected. The critical value for the test-statistic is determined by the level of significance. The critical value is the value that divides the non-reject region from the reject region.

4. Compute the appropriate test statistic and make the decision.

Compare the computed test statistic with critical value. If the computed value is within the rejection region(s), we reject the null hypothesis; otherwise, we do not reject the null hypothesis.

5. Interpret the decision.

Based on the decision in Step 4, we state a conclusion in the context of the original problem [4].

In order to carry out some statistical research and apply statistical methods in linguistics you must have a representative corpus that meets the task. In this study it was used the corpus "Scientific and fiction" in order to study deeper statistical hypotheses testing methods in linguistics. Used corpus is a sample that contains only some of the necessary material in contrast to the general population. Since the usage of whole general population is not possible it was decided to have a representative sample of all elements of the population and objects that appear frequently in the general population, often manifested in it. The purpose of this corpus is to create the most informative sampling on this topic for further statistical analysis.

As the test material it was used corpus containing a collection of scientific and literary texts. Testing statistical hypotheses were carried out in accordance with established Pearson's chi-squared criteria. Knowing such values as number of sentences, nouns, conjunctions help to test statistical hypotheses. The analysis allowed us to determine the null hypothesis. As a result of the program, the user can see the results of treatment. The main indicator is the value of chi-squared criteria.

References

1. Lehmann E. Testing Statistical Hypotheses / E. Lehmann, L. Romano, P. Joseph. – New York: Springer, 2005.
2. Gosall K. Doctor's Guide to Critical Appraisal / K. Gosall, N. Kaur, S. Gurpal. – Knutsford: PasTest, 2012.
3. Schervish M. Theory of Statistics / M. Schervish. – New York: Springer, 1996.
4. Wikipedia: Statistical hypothesis testing: [Electronic source]. – Access mode: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing