RANK TESTS FOR INDEPENDENCE AND RANDOMNESS

Martin Schindler
Department of Statistics, Charles University, Prague, Czech Republic

Keywords: Rank, correlation, trend, independence, randomness, ties, locally most powerful rank tests.

Contents

Summary

This chapter concerns rank tests for independence of two random variables. A form of locally most powerful test against particular alternative is derived and the most often used variants of this test are mentioned, namely the test of van der Waerden type, Spearman rank correlation coefficient, the quadrant test and the Kendall rank correlation coefficient, a member of non-linear rank statistics. Tests of hypothesis of randomness against trend alternatives are then studied and compared to the tests for independence that are formally similar. These tests, as well as a modification of the Kruskal-Wallis statistic in the presence of ties, are applied also to contingency tables. Modifications of all test statistics in the presence of ties and their expectation and variance formulas are given. The tests are applied and compared to each other and to classical tests of independence on examples. All of these tests are distribution free and relatively easy to use, which makes them widely applicable.

1. Introduction

A frequent statistical problem is to decide whether there exists any relationship between two characteristics in a set of units. These units can be for example people, cities, firms etc. To every unit two characteristics are assigned: e.g. human height and weight, number of inhabitants and criminality of a city, average wage and size of a firm etc. To test a hypothesis of independence (i.e. a hypothesis that there is no correlation between these two characteristics) we randomly choose a sample of size N from the population of units and get the values of the characteristics from all of the N units. When investigating a dependence of human height and weight, one could assume underlying bivariate normal distribution, i.e. one could assume that the data $\{(X_i, Y_i)\}_{i=1}^{N}$, where X_i is the height and Y_i is the weight of the ith individual form a sample from a bivariate normal distribution with positive variances and correlation ρ. Under these assumptions the hypothesis of independence simplifies to hypothesis: $H_0 : \rho = 0$ against alternative $K : \rho \neq 0$. Optimal choice for this problem is a test based on a sample correlation coefficient

$$r = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}}, \tag{1}$$

where $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$. Then:

$$T = \frac{r}{\sqrt{1-r^2}}\sqrt{N-2} \sim t_{N-2} \qquad \text{if } N \geq 3. \tag{2}$$

We reject the hypothesis H_0 on the level of significance α, if $T \geq t_{N-2}(\alpha)$, where $t_{N-2}(\alpha)$ stands for the critical value of a t-distribution with $N-2$ degrees of freedom.

Unfortunately one cannot assume bivariate normal distribution in most cases. Then we assume only that the random sample $\{(X_i, Y_i)\}_{i=1}^{N}$ comes from an arbitrary bivariate continuous distribution $h(x, y)$. Under these assumptions the rank tests for independence are used. The most frequently used rank test is based on the Spearman correlation coefficient r_S. It is easily expressed by means of r (see Eq. (1)) by replacing $\{X_i\}_{i=1}^{N}$ with their ranks $\{R_i\}_{i=1}^{N}$ among all $\{X_i\}_{i=1}^{N}$ and similarly by replacing $\{Y_i\}_{i=1}^{N}$ with their ranks $\{Q_i\}_{i=1}^{N}$. Under the assumption of an underlying continuous distribution no ties are present in the sample with probability one and the ranks are uniquely defined. The Spearman correlation coefficient is defined as follows:

$$r_S = \frac{\sum_{i=1}^{N}(R_i - \frac{N+1}{2})(Q_i - \frac{N+1}{2})}{\sqrt{\sum_{i=1}^{N}(R_i - \frac{N+1}{2})^2}\sqrt{\sum_{i=1}^{N}(Q_i - \frac{N+1}{2})^2}} = 1 - \frac{6\sum_{i=1}^{N}(R_i - Q_i)^2}{N(N^2-1)}. \tag{3}$$

For r_S it holds that: $r_S \leq 1$, as well as for ρ. At the same time it can be seen from the right part of the Eq. (3) that if R_i and Q_i very close for every i, one can anticipate a positive correlation and r_S is close to one. Whereas if the ranks are very different, r_S is close to minus one. Thus the correlation coefficient r_S is a popular estimate of the strength of the dependence of two characteristics in a population and we can use r_S as a nonparametric estimate of the correlation coefficient of the distribution $h(x, y)$.

For illustration let us use a data concerning knowledge of students in science and history.

Example: Suppose that 10 students from a class were randomly chosen and were given a science test and history test with the following results.

Student	1	2	3	4	5	6	7	8	9	10
Science score	26	22	16	23	50	44	47	48	40	43
History score	51	22	19	34	55	35	31	52	53	49

The aim is to test the hypothesis of independence between the science and history scores. Let us replace test scores by their ranks separately in both samples. The following table results:

Student	1	2	3	4	5	6	7	8	9	10
Science rank	4	2	1	3	10	7	8	9	5	6
History rank	7	2	1	4	10	5	3	8	9	6

It can be seen that the science ranks more or less corresponds to the history ranks and using Eq. (3) results in: $r_S = 0.6606$. For example, from a table of the null distribution of r_S it is seen that under H_0: $P(r_S \geq 0.6606) = 0.044$, which is the significant probability as one rejects the null hypothesis H_0 against alternative of positive or negative dependence for large values of r_S. So we reject the hypothesis of independence on the 5% level of significance in favor of the alternative. Let us compare this result with the test based on sample correlation coefficient, that one can use only under the assumption of a bivariate normal distribution. From Eq. (1) one gets: $r = 0.6216$ and under H_0: $P(r \geq 0.6216) = 0.055$. So this procedure would not reject the null hypothesis on the 5% level of significance.

Tests for hypothesis of randomness (H_*) are closely related to tests for independence. they are formally the same, but the interpretation of results is different. They are used to investigate the effect of a treatment or a factor on an characteristic (response). Let us suppose that the treatment is applied at N different levels. Usually the levels of treatment are assigned randomly to N subjects, on which the characteristic of interest is observed, one observation is taken on each level. The factor is very often time. For instance, a characteristic (average summer temperature) is observed throughout several years and we want to know whether the temperature rises during the period of time (testing H_* against an alternative of an upward trend). Rejecting the hypothesis of randomness in favor of an alternative means that a change of a factor (treatment) causes a change of the response. The interpretation for tests of independence (H_0) is different though. On the example mentioned earlier in this section one can see that the levels of a factor (science test score) cannot be randomly assigned to the subjects (students), whereas a random sample from a population of interest must be taken and both of the characteristics of the units in the sample found out. The levels of both characteristics are then random. Therefore, by rejecting the hypothesis of independence (H_0) one cannot say that a high science score causes a high (or low) history score. One can only say that there is a relation between them, which can be either caused by acting of one of the characteristic on the other or a result of a common unknown influence on both of them.

Finally, tests for independence and randomness in contingency tables will be discussed. These tests are similar to the tests for randomness, where there is only a finite number of possible responses. Different tests are used against different alternative hypotheses.

-
-
-

Bibliography

Hájek J., Šidák Z. (1967): Theory of Rank Tests. Academia, Praha. [Comprehensive text on rank tests and their properties including the asymptotic distribution and optimality.]

Kendall M. G. (1962): Rank Correlation Methods. (3rd ed.) Griffin, London. [Excellent application oriented text about rank tests including tests of randomness.]

Biographical Sketch

Martin Schindler was born on January 26 1981, in Hranice, the Czech Republic. He studied mathematical statistics at Charles University in Prague and received a master’s degree in mathematical statistics in 2004 with a thesis on rank tests for independence. Presently, he continues to study for a doctoral degree in the area of probability and mathematical statistics.