ROBUST 2008

RATIO STATISTICS

František VÁVRA, Tomáš PAVELKA, Blanka ŠEDIVÁ, Kateřina VOKÁČOVÁ, Patrice MAREK, Martina NEUMANOVÁ

DEPARTMENT OF MATHEMATICS, FACULTY OF APPLIED SCIENCES,

UNIVERSITY OF WEST BOHEMIA, UNIVERZITNI 8, 306 14 PILSEN, CZECH REPUBLIC

The results of independent experiments used for testing of some method or some phenomenon are often represented by samples $(r_1,s_1),(r_2,s_2),\ldots,(r_n,s_n)$, where $r_i,s_i \geq 0; \forall\, i = 1,\ldots,n$, $\sum_{i=1}^{n} r_i > 0$. The main goal of our contribution is to analyse random variable $\xi_n = \sum_{i=1}^{n} r_i \Big/ \left(\sum_{i=1}^{n} r_i + \sum_{i=1}^{n} s_i \right)$. The analysis concerns about probability description, tolerance bounds estimation and introduces its asymptotic characteristics as well. Furthermore are examined two particular cases of n. In the first case, n is assumed to be non-random nevertheless being changeable (e.g. comparison of different speech recognition methods). In the other case, n is assumed to be fixed (e.g. calculation of a state unemployment rate from all district unemployment rates).

MODEL

- $(r_1,s_1),(r_2,s_2),\ldots,(r_n,s_n)$ - represent the observations of independent experiments, where
$$r_i, s_i \geq 0; \forall\, i = 1,\ldots,n \text{ and } \sum_{i=1}^{n} r_i > 0.$$
- The mean value, variation and the covariance of each variable are known.
$$\forall\, i = 1,\ldots,n; E\{r_i\} = e_{ir}; E\{s_i\} = e_{is}; \sigma^2(r_i) = \sigma^2_{ir}; \sigma^2(s_i) = \sigma^2_{is} \text{ and}$$
$$\forall\, i = 1,\ldots,n; j = 1,\ldots,n; E\left\{\left(s_i - e_{js}\right)\left(r_i - e_{jr}\right)\right\}.$$
- The random variable of our interest is the criterion $\xi_n = \dfrac{\sum_{i=1}^{n} r_i}{\sum_{i=1}^{n} r_i + \sum_{i=1}^{n} s_i} = \dfrac{1}{1 + \dfrac{\sum_{i=1}^{n} s_i}{\sum_{i=1}^{n} r_i}} = \dfrac{1}{1+\eta_n}$,

  which represents the quality of a tested method or a rate of some phenomenon.

PROBABILITY DESCRIPTION

$$F_{\xi_n}(x) = P\left\{\frac{1}{1+\eta_n} < x\right\} = P\left\{\frac{1}{x} < 1+\eta_n\right\} = P\left\{\frac{1}{x} - 1 < \eta_n\right\} = 1 - P\left\{\eta_n < \frac{1}{x} - 1\right\} = 1 - F_{\eta_n}\left(\frac{1}{x} - 1\right).$$

For enough large n:

$$F_{\eta_n}(x) = P\{\eta_n < x\} = P\left\{\sum_{i=1}^{n} s_i - x\sum_{i=1}^{n} r_i < 0\right\} \approx$$
$$\approx \Phi\left(-\sqrt{n} \frac{\left(e_s - xe_r\right)}{\sqrt{\sigma_s^2 + x^2\sigma_r^2 - 2x\rho_{s,r}\sigma_s\sigma_r}} \right)$$

SUMMING UP

$$F_{\eta_n}(x) = P\{\eta_n < x\} \approx \Phi\left(-\sqrt{n} \frac{\left(e_s - xe_r\right)}{\sqrt{\sigma_s^2 + x^2\sigma_r^2 - 2x\rho_{s,r}\sigma_s\sigma_r}} \right)$$

$$F_{\xi_n}(x) \approx 1 - \Phi\left(-\sqrt{n} \frac{\left(e_s - (\frac{1}{x} - 1)e_r\right)}{\sqrt{\sigma_s^2 + \left(\frac{1}{x} - 1\right)^2\sigma_r^2 - 2\left(\frac{1}{x} - 1\right)\rho_{s,r}\sigma_s\sigma_r}} \right) \Leftrightarrow 0 < x < 1$$

POINT ESTIMATE OF MEASURE OF LOCATION

The most representative measure of the location for such type of distribution is the median, which is the solution of the equation $F_{\xi_n}(x_{med}) = 1/2$.

SUMMING UP

$$e_s = \left(\frac{1}{x_{med}} - 1\right)e_r \Rightarrow x_{med} = \frac{e_r}{e_r + e_s}$$

TOLERANCE INTERVAL

The goal is to find the percentile range that represents a specified proportionality of a “population”. $\alpha = \alpha_1 + \alpha_2; 0 < \alpha, \alpha_1, \alpha_2 < 1$ are chosen to satisfy

$$P\left(x_U < \xi_n\right) = \alpha_1 \Leftrightarrow P\left(\xi_n < x_U\right) = 1 - \alpha_1 \text{ and } P\left(\xi_n < x_L\right) = \alpha_2.$$

Determination of the lower bound x_L

$$1 - \Phi\left(-\sqrt{n} \frac{\left(e_s - (\frac{1}{x_L} - 1)e_r\right)}{\sqrt{\sigma_s^2 + \left(\frac{1}{x_L} - 1\right)^2\sigma_r^2 - 2\left(\frac{1}{x_L} - 1\right)\rho_{s,r}\sigma_s\sigma_r}} \right) = \alpha_2$$

If the notation $z = \frac{1}{x_L} - 1$ is used, and the equation is transformed to $Az^2 + 2zB + C = 0$, where

$$A = \left(a\sigma_r^2 - e_r^2\right); B = \left(e_s e_r - \rho_{s,r} a \sigma_r \sigma_s\right); C = \left(a\sigma_s^2 - e_s^2\right) \text{ and } \frac{\left[\Phi^{-1}\left(1-\alpha_2\right)\right]^2}{n} = a.$$

The roots are $z_{1,2} = \dfrac{-B \pm \sqrt{B^2 - AC}}{A}$ The identification of the valid solution is arranged by checking whether the particular root satisfies the following equation $\dfrac{\Phi^{-1}\left(1-\alpha_2\right)}{\sqrt{n}}\sqrt{\sigma_s^2 + z^2\sigma_r^2 - 2z\rho_{s,r}\sigma_s\sigma_r} + \left(e_s - ze_r\right) = 0$. The only one root is valid. Let’s denote the solution z_*. Using the transformation $z_* = \frac{1}{x_L} - 1$ the solution $x_L = \frac{1}{1+z_*}$ is obtained.

Determination of the upper bound x_U

The procedure of the calculation is almost identical to previous calculation of x_L. The only differences are $\dfrac{\left[\Phi^{-1}\left(\alpha_1\right)\right]^2}{n} = a$ and the roots’ testing equation is

$$\frac{\Phi^{-1}\left(\alpha_1\right)}{\sqrt{n}}\sqrt{\sigma_s^2 + z^2\sigma_r^2 - 2z\rho_{s,r}\sigma_s\sigma_r} + \left(e_s - ze_r\right) = 0.$$

ACCEPTABILITY OF ASYMPTOTICAL REPRESENTATION

The verification of acceptable asymptotic characteristics of the distribution function tests the following two aspects:

1. the condition of **the correctness of the approximation -** the function $F_{\xi_n}(x)$ should be no-decreasing function in the interval $\langle 0,1 \rangle$.
2. the condition of the $\beta = \beta_0 + \beta_1$ **acceptability** - the value of $F_{\xi_n}(x)$ for x in “zero value” should be non negative and smaller than known small positive number β_0. The value of $F_{\xi_n}(x)$ for x in “one” shouldn’t exceed the one and should be greater than $1 - \beta_1$. Where β_1 is positive, sufficiently small.

- **The necessary and sufficient condition of correctness** of the approximation is $z = \dfrac{\sigma_s}{\sigma_r}\dfrac{\rho_{s,r}e_s\sigma_r - e_r\sigma_s}{e_s\sigma_r - \rho_{s,r}e_r\sigma_s} \leq 0$, where $z = \frac{1}{x} - 1$,

Note: The mentioned correctness condition is not dependent on n - the number of observations.

- **The condition of the β-acceptability**

$$n \geq \max\left\{ \frac{\sigma_s^2}{e_s^2}\left(\Phi^{-1}\left(\beta_1\right)\right)^2 ; \frac{\sigma_r^2}{e_r^2}\left(\Phi^{-1}\left(1-\beta_0\right)\right)^2 \right\}$$

Note: The condition of the **β-**acceptability is dependent on the observation number n.

APPLICATION

Fixed number n

The measurement of the rate of unemployment of the whole country

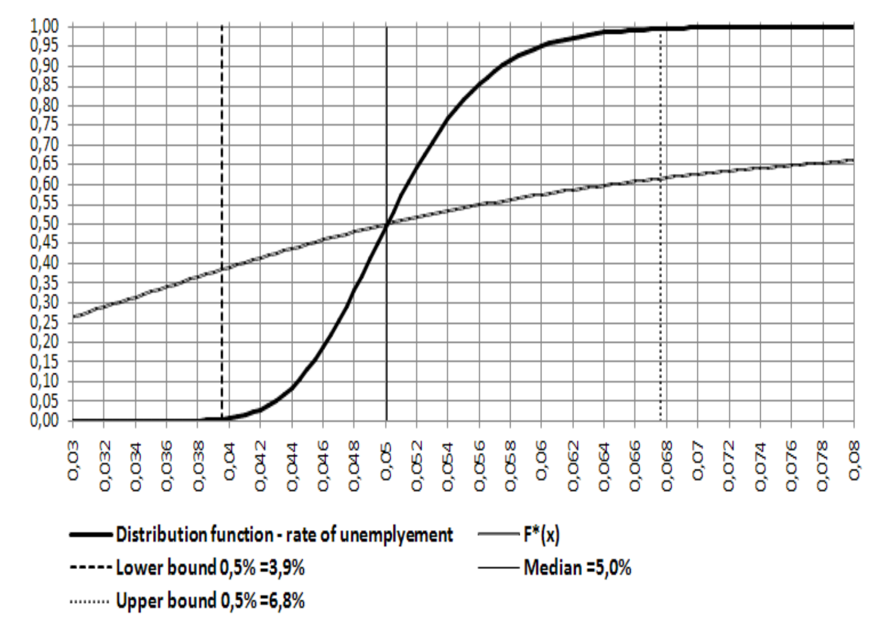
- r_i is the number of registered unemployed in i-th district,
- s_i is the number of employed in i-th district,
- n is the number of districts

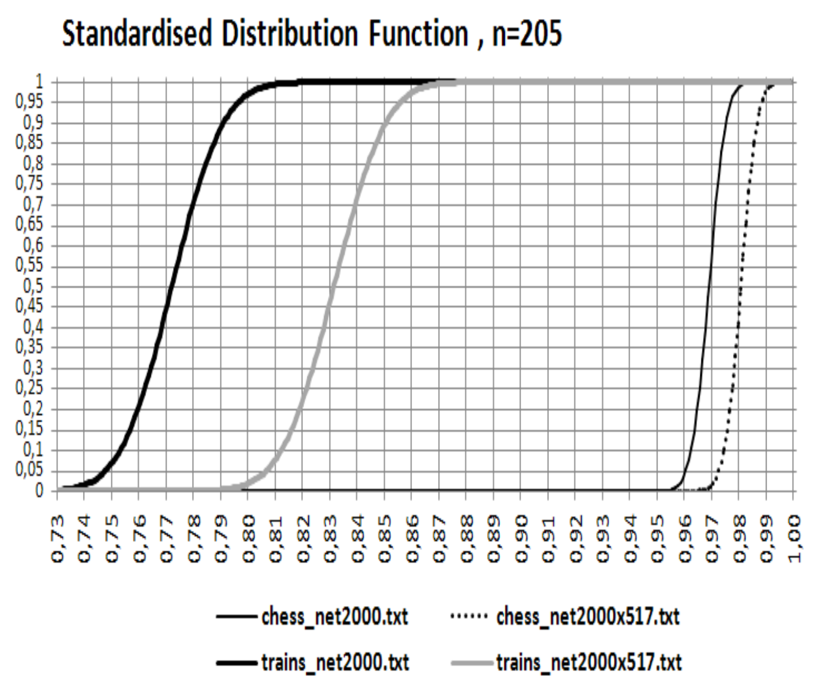
Changeable number n

The testing of speech recognition methods, where

- r_i is the number of recognised words in i-th sentence,
- s_i is the number of incorrectly recognised words in i-th sentence,
- N_i is the length of the i-th sentence in number of words,
- n is the number of input sentences of the method.

The comparison is possible just with the adjusted functions with the same number of the observations n. The concept of the β-acceptability offers to choose the smallest possible value n, valid for all the corpuses.





CONCLUSION

The mentioned methods can be applied in the cases where the measurements are in the form of a ratio criterion. The ratio criterion takes the values from the interval $\langle 0,1 \rangle$ for the purposes of our contribution. The extension of this interval by “one” could be arranged just by a technical procedure. Nevertheless it is necessary to solve few problems with the correctness and acceptability of the asymptotic approximation. Further research can be focussed on the stochastic comparison mentioned in the second analysed case with changeable n.

This work was supported by MŠMT – the name of the project: Research of continues and discreet mathematical structures and development of new methods. (MSM4977751301)