Rank Tests, Generating Functions and Computer Algebra

Chris Smit
University of Pretoria, Department of Statistics
Pretoria, South Africa
csmit@hakuna.up.ac.za

1. Introduction

Calculation of the non-nulldistribution of rank tests (to find the power of a test) has always been extremely complicated and only a very limited selection of power value calculations for small samples has been known in literature. The solution to the problem of finding the power has usually been to either use the (normal) approximation of the test statistic for large samples, or to find small sample approximations through simulation (Monte Carlo studies). The purpose of this paper is to show how the **exact** distribution of the Wilcoxon signed rank test statistic can be found from the probability generating function (pgf) both under the null and the alternative hypothesis and even in the bivariate case, using the computer algebra package *Mathematica*.

2. Brief literature overview of the signed rank test's distribution

Klotz (1963) found power values for the signed rank test for small samples from normal distributions using numerical calculations on a mainframe computer. Mitic (1996) used computer algebra (*Mathematica*) and a combinatorial approach to derive a general procedure for finding the null distribution of the Wilcoxon signed rank test. Lourens (1999) developed a general *Mathematica* notebook to find the exact power of both Wilcoxon's tests (one - and two-sample cases) for practically any continuous distribution, complying with the usual regularity conditions on density functions. Due to insufficient computer memory and the time-consuming calculation process, her results were limited to very small samples (typically for $n \le 5$ or 6). Van de Wiel, Bucchianico & van der Laan (1999) were the first to use the probability generating function combined with computer algebra to find the **null distribution** of the two Wilcoxon tests, as well as some other distribution-free tests.

3. The exact distribution of Wilcoxon's one-sample test

Let $X_1, X_2, ..., X_n$ be a random sample from a symmetric continuous distribution with cumulative distribution function *F(x)* and median h. Consider testing of $H_0 : h = h_0$ against $H_1 : h > h_0$. Define $D_i = X_i - h_0$, i=1,2, ..,n and let R_i^+ denote the rank of $|D_i|$ when all the |*D*|'s are ranked in order of increasing magnitude. Finally, let $Z_{(i)} = 1$, if the *i*th smallest |*D*| is associated with a positive *D*, and 0, otherwise. Under H_0, p_i=0.5 and under H_1, p_i>0.5. The following form of the test statistic is used: $U^+ = \sum_{i=1}^{n} i\, Z_{(i)}$. The pgf of U^+ is given by $f_{U^+}(t) = \prod_{i=1}^{n} \left(1 - p_i + p_i t^i\right)$. It is known (see, for example, Gibbons (1971)) that $p_i = n\binom{n-1}{i-1}\int_0^\infty [F_D(u) - F_D(-u)]^{i-1}[1 - F_D(u) + F_D(-u)]^{n-i} f_D(u)du$, where F_D and f_D are respectively the cumulative distribution function and density function of *D*. In this paper it is illustrated how *Mathematica* can be used to find exact values for p_i (under H_1) and how to derive the distribution of the one-sample Wilcoxon test statistic under H_0 and H_1 (assuming various parent distributions) from $f_{U^+}(t)$.

4. The bivariate signed rank test

$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is a random sample from a bivariate continuous distribution (symmetric in each component) with mean vector $m = (m_1, m_2)$ and unknown covariance matrix Σ. Consider testing of $H_0 : m = m_0 = (m_{10}, m_{20})$ against a right-sided alternative. Let $D_{1i} = x_i - m_{10}$ and $D_{2i} = y_i - m_{20}$ and let R_{1i}^+ and R_{2i}^+ be the absolute ranks of the D_{1i} and D_{2i} respectively, with the R_{1i}^+ being observed in natural order, 1,2, .,n ,without loss in generality. Denote by U^+ and V^+ respectively the signed rank test statistics for the two components in the same notation as in the univariate case. The bivariate signed rank test of Bennett (1964) is based on a test statistic which is approximately distributed as a chi-square variable with 2 degrees of freedom.

The pgf of (U^+, V^+) , under the null hypothesis and given the observed sample, is $G_{U^+,V^+}(u,v) = (\frac{1}{4})^n \prod_{i=1}^{n} \{(1+u^i)(1+v^{r_i}) + \hat{q}(1-u^i)(1-v^{r_i})\}$, with $\hat{q}$ a measure of the correlation between X and Y, defined by Bennett. Developing $G_{U^+,V^+}(u,v)$ for the observed sample gives the conditional bivariate probability distribution of (U^+, V^+) . From this distribution we may calculate the p-value: $P(U^+ \geq u_{obs}, V^+ \geq v_{obs} \mid H_0, \text{given sample})$, where u_{obs} and v_{obs} are the observed values of the two components. We demonstrate how this exact p-value can be obtained through *Mathematica*.

REFERENCES

Bennett, B.M. (1964). A bivariate signed rank test. Journal of the Royal Statistical Society Series B 26, 457-561.

Gibbons, J.D. (1971). Nonparametric statistical inference.McGraw-Hill Kogakusha LTD, Tokyo

Klotz, J. (1963). Small sample power and efficiency for the one-sample Wilcoxon and normal scores tests. Annals of Mathematical Statistics 34, 624-632.

Lourens, A. (1999). Finding the power of distribution-free tests using a computer algebra system. Unpublished PhD thesis, University of Pretoria.

Van de Wiel, M.A., Di Bucchianico, A. and van der Laan, P. (1999). Symbolic computation and exact distributions of nonparametric tests statistics. The Statistician 48, 507-516.

RESUME

Le calcul de la répartition des tests de rang a toujours été extrêmement difficile et on ne trouve dans la document ation spécialisée qu' un choix très limité des calculs de valeur de puissance pour petits échantillons. Le but de cette communication est de montrer comment la répartition exacte de la statistique du test des signes de Wilcoxon (test des signes pour observations appariées) peut être trouvée à partir de la fonction génératrice de probabilités à la fois sous l' hypothèse nulle et l' hypothèse alternative et même dans le cas bivarié en se servant du progiciel algébrique *Mathematica*.