Kybernetika

Michal Černý

Binary segmentation and Bonferroni-type bounds

*Kybernetika*, Vol. 47 (2011), No. 1, 38--49

Persistent URL: http://dml.cz/dmlcz/141476

KYBERNETIKA — VOLUME 47 (2011), NUMBER 1, PAGES 38–49

# BINARY SEGMENTATION AND BONFERRONI–TYPE BOUNDS

Michal Černý

We introduce the function $Z(x;\xi,\nu) := \int_{-\infty}^{x} \varphi(t-\xi)\cdot\Phi(\nu t)\,\mathrm{d}t$, where $\varphi$ and $\Phi$ are the pdf and cdf of $N(0,1)$, respectively. We derive two recurrence formulas for the effective computation of its values. We show that with an algorithm for this function, we can efficiently compute the second-order terms of Bonferroni-type inequalities yielding the upper and lower bounds for the distribution of a max-type binary segmentation statistic in the case of small samples (where asymptotic results do not work), and in general for max-type random variables of a certain type. We show three applications of the method — (a) calculation of critical values of the segmentation statistic, (b) evaluation of its efficiency and (c) evaluation of an estimator of a point of change in the mean of time series.

*Keywords:* Bonferroni inequality, segmentation statistic, Z-function

*Classification:* 62E17, 05A20

## 1. INTRODUCTION

It is a traditional problem to determine the distribution of the random variable $T^{[n]} := \max_{1\le k\le n} T_k$ for a *fixed* $n \ge 2$, where $T_1,\ldots,T_n$ are dependent random variables. If $T_1,\ldots,T_n$ are independent, the problem is easy, while in the dependent case the distribution may be complicated and it is usually impossible to describe it by a 'nice' formula. Extremal theory often allows us to study the behavior of $T^{[n]}$ when $n\to\infty$ (under further assumptions on $T_1,\ldots,T_n$); however, asymptotic results are not always suitable if $n$ is small, which is the case we are interested in.

A natural tool to approximate the distribution of $T^{[n]}$ is the Bonferroni inequality. In this text we introduce a method which allows us to use second-order inequalities of the Bonferroni type for a special class of max-type random variables. As a prominent example of this class, we study the so-called binary segmentation statistic, a max-type statistic designed for testing the hypothesis that there is no change in the mean of time series against the hypothesis that the change exists (in Section 4 we shall be more precise). This statistic is important in quality control, in analysis of financial data and in econometrics. Other max-type random variables, for which our method is also applicable, occur variously, for example in the analysis of interval regression models.

We present three applications of the method:

- we show how to derive critical values for the segmentation statistic, which are easily computable and less conservative than the asymptotic values and the traditional approximations obtained by the first-order Bonferroni inequality;
- we show how to estimate the efficiency of the statistic;
- we derive a bound on an estimator of the point of change.

The main tool in the analysis is the $Z$ function introduced in the next section.

## 2. THE $Z$ FUNCTION

Let $\varphi(x) = (2\pi)^{-1/2}\exp(-\frac{1}{2}x^2)$ and $\Phi(x) = \int_{-\infty}^{x}\varphi(t)\,\mathrm{d}t$. We show some properties of the function

$$Z(x;\xi,\nu) = \int_{-\infty}^{x}\varphi(t-\xi)\cdot\Phi(\nu t)\,\mathrm{d}t. \tag{1}$$

We assume that values of $\Phi(x)$ are easily computable with suitable software.

**Numerical integration.** The $Z$ function has the following useful property: given computation precision, it is sufficient to integrate numerically over an interval of *fixed length*, regardless of the values of $x, \xi, \nu$. Indeed, if we numerically evaluate

$$\int_{(-\infty,x)\cap(\xi-\Delta,\xi+\Delta)}\varphi(t-\xi)\cdot\Phi(\nu t)\,\mathrm{d}t,$$

then the error of computation caused by the truncation of the integration range is bounded by $2\Phi(-\Delta)$ which is, say for $\Delta = 5$, sufficiently small.

We can also truncate $\Phi(\nu x)$ in a similar way: assume $\nu > 0$ and replace $\Phi(\nu x)$ with the function

$$\Phi^*(\nu x) = \begin{cases} 0 & \text{for } x < -\frac{\Delta}{\nu}, \\ \Phi(\nu x) & \text{for } x \in [-\frac{\Delta}{\nu}, \frac{\Delta}{\nu}], \\ 1 & \text{for } x > \frac{\Delta}{\nu}; \end{cases}$$

now the total error from truncation of $\varphi$ and $\Phi$ is bounded by

$$2\Phi(-\Delta) + 2\Delta\Phi(-\Delta) = 2\Phi(-\Delta)(1+\Delta).$$

The assumption $\nu > 0$ is without loss of generality, as for negative values of $\nu$ it is possible to use the identity

$$Z(x;\xi,\nu) = \Phi(x-\xi) - Z(x;\xi,-\nu).$$

**Expansions.** It is possible to write $Z(x;\xi,\nu)$ in the form

$$Z(x;\xi,\nu) = \frac{1}{\sqrt{2\pi}}\sum_{i=0}^{\infty}\left(-\frac{1}{2}\right)^i\frac{1}{i!}\underbrace{\int_{-\infty}^{x}(t-\xi)^{2i}\cdot\Phi(\nu t)\,\mathrm{d}t}_{=:L_{2i}(x)} \tag{2}$$

and

$$Z(x;\xi,\nu) = \frac{1}{2}\Phi(x-\xi) + \frac{\nu}{\sqrt{2\pi}}\sum_{i=0}^{\infty}\left(-\frac{\nu^2}{2}\right)^i \cdot \frac{1}{(2i+1)\cdot i!}\underbrace{\int_{-\infty}^{x} t^{2i+1}\cdot\varphi(t-\xi)\,\mathrm{d}t}_{=:K_{2i+1}(x)}. \tag{3}$$

These expressions follow from the Taylor series

$$\varphi(x) = (2\pi)^{-1/2}\sum_{i=0}^{\infty}\frac{1}{i!}\left(-\frac{1}{2}x^2\right)^i;$$

in case (2) applied to the term $\varphi(t-\xi)$ and in case (3) applied to the term $\Phi(\nu t)$. For computation of the initial segments of the series, the following recurrences are useful.

**Proposition 2.1.**

$$\begin{aligned}
L_0(x) &= x\Phi(\nu x) + \frac{1}{\nu}\varphi(\nu x),\\
L_1(x) &= \Big(\frac{x}{2} - \xi\Big)\cdot L_0(x) - \frac{1}{2\nu^2}\Phi(\nu x),\\
L_i(x) &= \frac{1}{i+1}\cdot\Big((x-\xi)^i L_0(x) - i\xi L_{i-1}(x) + \frac{i(i-1)}{\nu^2}L_{i-2}(x)\\
&\qquad - \frac{i}{\nu^2}(x-\xi)^{i-1}\Phi(\nu x)\Big) \quad \text{for } i\geq 2;\\
K_0(x) &= \Phi(x-\xi),\\
K_1(x) &= \xi K_0(x) - \varphi(x-\xi),\\
K_i(x) &= \xi K_{i-1}(x) - x^{i-1}\varphi(x-\xi) + (i-1)K_{i-2}(x) \quad \text{for } i\geq 2.
\end{aligned}$$

P r o o f . All of the expressions are derived from (2) and (3) by integration. Let us, for example, look at the equation for $L_i(x)$. Using the integration per-partes, from (2) we get

$$L_i(x) = (x-\xi)^i L_0(x) - i\int_{-\infty}^{x}(t-\xi)^{i-1}L_0(t)\,\mathrm{d}t.$$

An easy manipulation gives

$$L_i(x) = (x-\xi)^i L_0(x) - i(L_i(x) + \xi L_{i-1}(x)) - \frac{i}{\nu^2}((x-\xi)^{i-1}\Phi(\nu x) - (i-1)L_{i-2}(x)).$$

The expression for $L_i(x)$ follows. □

Observe that evaluation of each of the recurrences (2) and (3) requires only one computation of $\Phi(\cdot)$.

The following lemma shows an important property of the $Z$ function.

**Lemma 2.2.** Let $a, b, c, d, e \in \mathbb{R}$, $c \neq ad$. Then

$$\int_{-\infty}^{\infty}\int_{-\infty}^{dA+e}\int_{-\infty}^{aA+cC+b} \varphi(B)\varphi(C)\varphi(A)\,\mathrm{d}B\,\mathrm{d}C\mathrm{d}A$$
$$= Z\Big(\frac{b}{\sqrt{1+a^2+c^2}} + e\frac{\sqrt{1+a^2+c^2}}{c-ad};\ e\frac{\sqrt{1+a^2+c^2}}{c-ad},\ \frac{c-ad}{\sqrt{1+(a+cd)^2+d^2}}\Big). \quad (4)$$

P r o o f. The left-hand side of (4) may be written down as

$$\frac{1}{(2\pi)^{3/2}}\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{\infty}\exp\Big(-\frac{1}{2}\cdot[A^2+(B+aA+c(C+dA+e)+b)^2 + (C+dA+e)^2]\Big)\mathrm{d}A\mathrm{d}C\mathrm{d}B.$$

There exist quadratic functions $Q_1(A)$, $Q_2(B,C)$ and a constant $\upsilon$ such that the integral may be transformed into the form

$$\upsilon\cdot\int_{-\infty}^{0}\int_{-\infty}^{0}\exp(-\tfrac{1}{2}Q_2(B,C))\int_{-\infty}^{\infty}\exp(-\tfrac{1}{2}Q_1(A))\,\mathrm{d}A\mathrm{d}C\mathrm{d}B.$$

The inner integral equals $\sqrt{2\pi}$. Then there are quadratic functions $Q_3(B)$ and $Q_4(C)$ such that the integral may be written as

$$\eta\int_{-\infty}^{0}\exp(-\tfrac{1}{2}Q_3(B))\int_{-\infty}^{0}\exp(-\tfrac{1}{2}Q_4(C))\,\mathrm{d}C\mathrm{d}B$$

with some constant $\eta$. Linear substitutions transform it into the $Z$-form (1). □

## 3. SECOND–ORDER BONFERRONI–TYPE INEQUALITIES

Bonferroni-type inequalities are inequalities obtained as initial segments of the inclusion-exclusion principle

$$\Pr\Big[\bigcup_{k=1}^{n}A_k\Big] = \sum_{k=1}^{n}\Pr[A_k] - \sum_{k_1<k_2}\Pr\left[A_{k_1}\cap A_{k_2}\right] + \sum_{k_1<k_2<k_3}\Pr[A_{k_1}\cap A_{k_2}\cap A_{k_3}] - \cdots + (-1)^{n+1}\Pr\Big[\bigcap_{k=1}^{n}A_k\Big],$$

where $A_1, \ldots, A_n$ are events. The first-order Bonferroni inequality is

$$\Pr[\bigcup_{k=1}^{n}A_k] \leq \sum_{k=1}^{n}\Pr[A_k] \quad (5)$$

and second-order inequalities are inequalities involving second-order terms of the type $\Pr[A_{k_1} \cap A_{k_2}]$. There is a rich combinatorial theory on such inequalities, see for instance [7, 8, 10, 11]. Two important examples are

$$\Pr\Big[\bigcup_{k=1}^{n} A_k\Big] \geq \sum_{k=1}^{n} \Pr[A_k] - \sum_{k_1<k_2} \Pr[A_{k_1} \cap A_{k_2}]$$

and

$$\Pr\Big[\bigcup_{k=1}^{n} A_k\Big] \leq \sum_{k=1}^{n} \Pr[A_k] - \sum_{k=1}^{n-2} \Pr[A_k \cap A_{k+1}]. \tag{6}$$

Recall that (6) has been successfully used in [14], yielding a big gain of precision. The class of second-order inequalities also includes inequalities derived from higher-order inequalities where higher-order terms are estimated by second-order terms as, for instance,

$$\begin{aligned}\Pr\Big[\bigcup_{k=1}^{n-1} A_k\Big] \leq \sum_{k=1}^{n-1} \Pr[A_k] - \sum_{k=1}^{n-2} \Pr[A_k \cap A_{k+1}] \\ - \sum_{k=1}^{n-2} \Pr[A_k \cap A_{n-1}] + \sum_{k=1}^{n-2} \Pr[A_k \cap A_{k+1} \cap A_{n-1}],\end{aligned}$$

where we estimate

$$\Pr[A_k \cap A_{k+1} \cap A_{n-1}] \leq \min\{A_k \cap A_{k+1}, A_k \cap A_{n-1}, A_{k+1} \cap A_{n-1}\}.$$

Such inequalities are useful in the derivation of bounds on max-type random variables where we need to estimate the maximum of dependent random variables. Let $T_1, \ldots, T_n$ be random variables and define $T^{[n]} := \max_{1\leq k\leq n} T_k$. Then

$$\begin{aligned}\Pr[T^{[n]} \leq x] &= \Pr[T_k \leq x \text{ for all } k \in \{1, \ldots, n\}] \\ &= 1 - \Pr[T_k > x \text{ for some } k \in \{1, \ldots, n\}] \\ &= 1 - \Pr\Big[\bigcup_{k=1}^{n} A_k(x)\Big],\end{aligned}$$

where $A_k(x)$ denotes the event "$T_k > x$". Now we can use the Bonferroni-type inequalities to get lower or upper bounds on $\Pr[\bigcup_{k=1}^{n} A_k(x)]$, and hence upper or lower bounds on $\Pr[T^{[n]} \leq x]$.

## 4. APPLICATIONS

**The binary segmentation statistic.** Let $y_1, \ldots, y_n$ be independent normal variables with common variance $\sigma^2$. Let us ask the question whether the data are homogenous in the sense that $E(y_1) = E(y_2) = \cdots = E(y_n)$, or whether there is a change in their mean. So, consider two hypotheses:

$$H: \quad y_i = \mu + \varepsilon_i \quad \text{for all } i = 1, \ldots, n,$$

and

$$A: \quad \exists\,\kappa \in \{1,\dots,n-1\},\ \exists\,\delta \neq 0: \quad y_i = \begin{cases} \mu+\varepsilon_i & \text{for } i=1,\dots,\kappa, \\ \mu+\delta+\varepsilon_i & \text{for } i=\kappa+1,\dots,n, \end{cases}$$

where $\mu, \delta$ and $\kappa$ are unknown parameters and $\varepsilon_1,\dots,\varepsilon_n$ are independent $N(0,\sigma^2)$ error terms. The *binary segmentation statistic*

$$T^{[n]} := \max_{1\le k\le n-1} \underbrace{\frac{1}{\sigma}\sqrt{\frac{n}{k(n-k)}}\sum_{i=1}^{k}(y_i-\overline{y})}_{=:T_k^{[n]}}, \tag{7}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, is (a form of) the max-likelihood statistic for testing the null hypothesis $H$ against the alternative $A$.

For derivation of the statistic, its applications and further discussion on the topic see [2, 4, 6, 9, 13, 14]. By extremal theory, the asymptotic distribution of $T^{[n]}$, when $n\to\infty$, is known; however, for small-sized samples, the exact distribution is very complicated. The asymptotic result will be stated later.

**Remark 1.** We assume that $\sigma^2$ is known. If $\sigma^2$ is unknown, then $\sigma$ in (7) has to be replaced by an estimate. In that case, the reduction to evaluation of the $Z$ function, described later in this section, is an open problem. However, our bounds are applicable if known upper/lower bounds on $\sigma^2$ are available.

We shall present the usage of the $Z$ function for derivation of bounds on the statistic $T^{[n]}$. However observe that our method is useful in a more general context, e. g. for max-type random variables of the type $\max_{1\le k\le n}\omega_k\sum_{j=1}^{k} y_j+\chi_k\sum_{j=k+1}^{n} y_j$ where $\omega_1,\dots,\omega_n;\chi_1,\dots,\chi_n$ are constants. Such max-type random variables occur for instance in the analysis of interval regression models (see [5, 12]).

**Example 1. Critical values for $T^{[n]}$.** When testing the hypothesis $H$ against $A$, we need to derive critical values for $T^{[n]}$ under $H$. Assume that $H$ holds, i. e. $y_i=\mu+\varepsilon_i$. Then $T_k^{[n]}\sim N(0,1)$: indeed, the fact $E(T_k^{[n]})=0$ is clear and

$$\operatorname{var}\left(\sum_{i=1}^{k}(y_i-\overline{y})\right) = \sum_{i=1}^{k}\operatorname{var}\left(\frac{n-k}{n}y_i\right) + \sum_{i=k+1}^{n}\operatorname{var}\left(\frac{k}{n}y_i\right) = \sigma^2\cdot\frac{k(n-k)}{n}.$$

The classical approach to the approximation of the distribution of (7) uses the first-order Bonferroni inequality (5). This approach yields the estimate

$$\Pr[T^{[n]}>x] \le (n-1)(1-\Phi(x)),$$

leading to very conservative $\alpha$-critical values

$$\pm\,\Phi^{-1}\Big(1-\frac{\alpha}{2(n-1)}\Big). \tag{8}$$

We shall show how to improve this bound with any type of the second-order Bonferroni-type inequality. We must derive an expression for

$$F_{k,l}^{[n]}(x) = \Pr\left[T_k^{[n]} > x \ \ \& \ \ T_l^{[n]} > x\right].$$

Assume that $1 \le k < l \le n-1$. By the definition of $T_k^{[n]}$ we can write

$$T_k^{[n]} = \alpha_1 A - \alpha_2 B - \alpha_3 C, \quad T_l^{[n]} = \beta_1 A + \beta_2 B - \beta_3 C,$$

where $A, B, C$ are $N(0,1)$ independent and

$$\alpha_1 = \sqrt{\frac{n-k}{n}}, \quad \alpha_2 = \sqrt{\frac{k(l-k)}{n(n-k)}}, \quad \alpha_3 = \sqrt{\frac{k(n-l)}{n(n-k)}}, \tag{9}$$

$$\beta_1 = \sqrt{\frac{k(n-l)}{nl}}, \quad \beta_2 = \sqrt{\frac{(l-k)(n-l)}{nl}}, \quad \beta_3 = \sqrt{\frac{l}{n}}. \tag{10}$$

We need to evaluate

$$\begin{aligned} F_{k,l}^{[n]}(x) &= \Pr[\tfrac{\alpha_2}{\alpha_1}B + \tfrac{\alpha_3}{\alpha_1}C < A - \tfrac{x}{\alpha_1} \ \ \& \ \ -\tfrac{\beta_2}{\beta_1}B + \tfrac{\beta_3}{\beta_1}C < A - \tfrac{x}{\beta_1}] \\ &= \int_{-\infty}^{\infty} \left[\iint_{\Omega_A} \varphi(B)\varphi(C)\,\mathrm{d}B\mathrm{d}C\right] \varphi(A)\,\mathrm{d}A, \end{aligned}$$

where

$$\Omega_A = \left\{[B,C] : \ \frac{\alpha_2}{\alpha_1}B + \frac{\alpha_3}{\alpha_1}C < A - \frac{x}{\alpha_1} \ \ \& \ \ -\frac{\beta_2}{\beta_1}B + \frac{\beta_3}{\beta_1}C < A - \frac{x}{\beta_1}\right\}.$$

The region $\Omega_A$ may be described as

$$\begin{aligned} \Omega_A = \Big\{[B,C] : \ C \in (-\infty, \ & \overbrace{\tfrac{\alpha_1\beta_2+\beta_1\alpha_2}{\alpha_3\beta_2+\beta_3\alpha_2}A - \tfrac{\alpha_2+\beta_2}{\alpha_3\beta_2+\beta_3\alpha_2}x}^{=:\overline{C}(A)}) \\ & \ \& \ \ B \in (\underbrace{-\tfrac{\beta_1}{\beta_2}A + \tfrac{\beta_3}{\beta_2}C + \tfrac{x}{\beta_2}}_{=:\underline{B}(A,C)}, \ \underbrace{\tfrac{\alpha_1}{\alpha_2}A - \tfrac{\alpha_3}{\alpha_2}C - \tfrac{x}{\alpha_2}}_{=:\overline{B}(A,C)})\Big\}. \end{aligned}$$

Thus we can write

$$\begin{aligned} F_{k,l}^{[n]}(x) = &\int_{-\infty}^{\infty}\int_{-\infty}^{\overline{C}(A)}\int_{-\infty}^{\overline{B}(A,C)} \varphi(B)\varphi(C)\varphi(A)\,\mathrm{d}B\mathrm{d}C\mathrm{d}A \\ &- \int_{-\infty}^{\infty}\int_{-\infty}^{\overline{C}(A)}\int_{-\infty}^{\underline{B}(A,C)} \varphi(B)\varphi(C)\varphi(A)\,\mathrm{d}B\mathrm{d}C\mathrm{d}A. \end{aligned}$$

Now we apply Lemma 2.2. Simplifying the resulting expressions, we get the following result.

**Proposition 4.1.** Let $1 \le k < l \le n-1$ and define

$$\zeta_{k,l}^{[n]} = \frac{\sqrt{(n-k)l}}{n} \cdot \sqrt{\frac{k}{n-l}}, \quad \eta_{k,l}^{[n]} = \frac{\sqrt{(n-k)l}}{n} \cdot \sqrt{\frac{n-l}{k}}. \tag{11}$$

Then

$$F_{k,l}^{[n]}(x) = Z\left(-x \cdot \left[\frac{k}{n} - \zeta_{k,l}^{[n]}\right];\ \ x \cdot \left[\zeta_{k,l}^{[n]} + \frac{n-k}{n}\right],\ \ -\sqrt{\frac{n}{l-k}}\sqrt{\frac{n-l}{k}}\right)$$
$$- Z\left(x \cdot \left[\frac{n-l}{n} - \eta_{k,l}^{[n]}\right];\ \ -x \cdot \left[\eta_{k,l}^{[n]} + \frac{l}{n}\right],\ \ \sqrt{\frac{n}{l-k}}\sqrt{\frac{k}{n-l}}\right).$$

Now we are able to compute the second-order terms in any second-order inequality.

Using the second-order inequality (6), we get the following improved (compared to the classical approach (5)) 5%-critical values for the segmentation statistic:

**Table 1.** Approximate 5% critical values for the segmentation statistic under $H$.

| $n$ | 20 | 30 | 40 | 60 | 80 | 100 | 198 | 3020 |
|---|---|---|---|---|---|---|---|---|
| asymptotic | 3.60 | 3.60 | 3.61 | 3.62 | 3.63 | 3.64 | 3.66 | 3.74 |
| using (5) | 3.01 | 3.13 | 3.22 | 3.34 | 3.42 | 3.48 | 3.66 | 4.30 |
| **using (6)** | **2.86** | **2.96** | **3.02** | **3.10** | **3.16** | **3.20** | **3.32** | **3.74** |
| simulated | 2.81 | 2.89 | 2.93 | 3.00 | 3.03 | 3.07 | 3.14 | 3.33 |

The asymptotic values have been derived from the following extremal-type theorem: *if* $y_1, \dots y_n$ *follow* $H$ (in fact, much less suffices), *then, with* $n \to \infty$,

$$\Pr\left[\sqrt{2\ln\ln n} \cdot T^{[n]} \le x + 2\ln\ln n + \tfrac{1}{2}(\ln\ln\ln n - \ln\pi)\right] \longrightarrow e^{-2e^{-x}}, \tag{12}$$

see [2]. Table 1 illustrates that for small values of $n$ appearing in practice, asymptotic results are too conservative. The value $n = 198$ is the lowest $n$ such that the critical value obtained with (5) exceeds the asymptotic critical value: for $n \ge 198$, the expression (8) is useless. The improved critical value "catches up" with the asymptotic critical value at $n = 3020$.

**Remark 2.** Further results are available; for instance, by [3], the critical values are also useful if $y_i$'s follow the $AR(1)$ process with parameter $\varrho \in (-1, 1)$. In that case, we shall use $T_k^{[n]}$ in the form $\frac{1-\varrho}{\sigma}\sqrt{\frac{n}{k(n-k)}}\sum_{i=1}^{k}(y_i - \overline{y})$ instead of (7) and the critical values derived for $T^{[n]}$ remain preserved.

**Remark 3.** Table 1 compares several methods for estimation of critical values. In order the comparison be complete, it is necessary to mention one more important method for derivation of critical values: the permutation principle. Let $\Pi_n$ be the set of all permutations of $\{1,\dots,n\}$. For $\pi \in \Pi_n$ denote

$$T_\pi^{[n]}(y_1,\dots,y_n) = \max_{1\le k\le n-1} \frac{1}{\sigma}\sqrt{\frac{n}{k(n-k)}}\sum_{i=1}^{k}(y_{\pi(i)} - \overline{y}).$$

If $y_1,\dots,y_n$ are fixed and $\pi \in \Pi_n$ is random, then $T_\pi^{[n]}(y_1,\dots,y_n)$ is a discrete random variable. The function

$$P_{y_1,\dots,y_n}^{[n]}(x) = \frac{1}{n!}\cdot\left|\{\pi\in\Pi_n:\ T_\pi^{[n]}(y_1,\dots,y_n)\le x\}\right|$$

is called *the permutation distribution function* (conditioned by given $y_1,\dots,y_n$).

It is interesting that if $H$ holds, then in the limit $n\to\infty$ the random variable $T_\pi^{[n]}(y_1,\dots,y_n)$ has the same asymptotic behavior as $T^{[n]}$. Indeed, the following remarkable theorem holds [1]: *if $y_1,\dots,y_n$ follow $H$* (in fact, much less suffices), *then, with $n\to\infty$,*

$$\begin{aligned}&\Pr_{\pi\in\Pi_n}\Big[\sqrt{2\ln\ln n}\cdot T_\pi^{[n]}(y_1,\dots,y_n)\\ \le\ & x + 2\ln\ln n + \tfrac{1}{2}(\ln\ln\ln n - \ln\pi)\ \mid\ y_1,\dots,y_n\Big] \longrightarrow e^{-2e^{-x}}\end{aligned}$$

*almost surely.* This theorem suggests that for fixed $n$ and $y_1,\dots,y_n$, the permutation distribution function $P_{y_1,\dots,y_n}^{[n]}(x)$ could be a good approximation of the true distribution of $T^{[n]}$. In practice, given $y_1,\dots,y_n$ and $x$, it is not computationally feasible to evaluate $P_{y_1,\dots,y_n}^{[n]}(x)$ exactly. However, simulations show that if we want to derive usual quantiles, taking about $20,000$ permutations at random provides a reasonable approximation of $P_{y_1,\dots,y_n}^{[n]}(x)$.

For each $n\in\{20,30,40,60,80,100\}$ we simulated the values of $y_1,\dots,y_n$ under $H$ for $5,000$ times. Denote them $y_1^i,\dots,y_n^i$, $i=1,\dots,5000$. Then, for every $n$ and $i=1,\dots,5000$, we calculated $P_{approx}^{[n,i]}$, the approximation of $P_{y_1^i,\dots,y_n^i}^{[n]}$ using $20,000$ random permutations from $\Pi_n$. Then we derived the empirical 5%-critical values $p_{n,i}$ from $P_{approx}^{[n,i]}$. Table 2 summarizes the average values (i. e. the values $q_n := \frac{1}{5000}\sum_{i=1}^{5000} p_{n,i}$).

**Table 2.** The permutation method — simulations.

| $n$ | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|
| $q_n$ | 2.58 | 2.75 | 2.83 | 2.92 | 2.96 | 3.03 |

Table 2 shows that for small-sized samples, the permutation method on average underestimates the true critical value significantly (see the simulated values in the last row of Table 1) and hence is likely to "detect" changepoints excessively. Moreover, the deviation from the true critical value might be quite high (for example, in

the simulation it was $\min_i p_{20,i} = 1.39$ and $\max_i p_{20,i} = 4.39$). We can conclude that though the permutation method is a popular and often used data-driven method, it does not provide a very exact approximation to the *true* critical values under $H$. As we can see in Table 1, the method (6) is more exact (and also computationally easier).

**Example 2** — efficiency of the statistic. Assume that $A$ holds and let $\kappa \in \{1, \dots, n-1\}$ and $\delta \neq 0$ be fixed. We want to quantify how successful the test is in detection of the existing change. In other words, we want to estimate $\Pr[T^{[n]} > x_\alpha]$, where $x_\alpha$ is an $\alpha$-critical value.

If $\kappa \le k$, then

$$T_k^{[n]} = -\kappa\delta\sqrt{\frac{n-k}{kn}} + \underbrace{\sqrt{\frac{n}{k(n-k)}}\left(\sum_{i=1}^{\kappa} y_i + \sum_{i=\kappa+1}^{k}(y_i-\delta) - \frac{k}{n}\left(\sum_{i=1}^{\kappa} y_i + \sum_{i=\kappa+1}^{n}(y_i-\delta)\right)\right)}_{\sim N(0,1)},$$

and similarly, if $\kappa > k$, then $T_k^{[n]} = -(n-\kappa)\delta\sqrt{\frac{k}{n(n-k)}} + N(0,1)$, so we get an expression for the first-order Bonferroni term

$$\sum_{k=1}^{n-1}\Pr[T_k \ge x_\alpha] = n-1-\sum_{k=1}^{\kappa}\Phi\Big(x_\alpha + \kappa\delta\sqrt{\frac{n-k}{kn}}\Big) - \sum_{k=\kappa+1}^{n-1}\Phi\Big(x_\alpha + (n-\kappa)\delta\sqrt{\frac{k}{n(n-k)}}\Big).$$

Now let $1 \le k < l \le n-1$ and let us evaluate the second-order terms

$$G_{k,l}^{[n]}(x_\alpha) := \Pr[T_k \ge x_\alpha \ \ \& \ \ T_l \ge x_\alpha].$$

The idea of decomposition of $T_k$ and $T_l$ as linear functions of independent $N(0,1)$ variables $A$, $B$, $C$, shown in Example 1, also applies under the alternative $A$:

$$T_k^{[n]} = \theta_k^{[n]} + \alpha_1 A - \alpha_2 B - \alpha_3 C, \quad T_l^{[n]} = \lambda_l^{[n]} + \beta_1 A + \beta_2 B - \beta_3 C, \tag{13}$$

where $A, B, C$ are $N(0,1)$ independent, $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2$ and $\beta_3$ are given by (9) and (10) and

$$\theta_k^{[n]} = \begin{cases} -\delta\kappa\sqrt{\frac{n-k}{nk}} & \text{if } k < \kappa, \\ -\delta(n-\kappa)\sqrt{\frac{k}{n(n-k)}} & \text{if } k \ge \kappa, \end{cases}$$

$$\lambda_l^{[n]} = \begin{cases} -\delta(n-\kappa)\sqrt{\frac{l}{n(n-l)}} & \text{if } l < \kappa, \\ -\delta\kappa\sqrt{\frac{n-l}{nl}} & \text{if } l \ge \kappa. \end{cases}$$

Simplifying the resulting expressions, we get the following analogy of Proposition 4.1, valid under the hypothesis $A$.

**Proposition 4.2.**

$$G_{k,l}^{[n]}(x) = Z\Big(\frac{k}{n}(\theta_k^{[n]} - x) + \zeta_{k,l}^{[n]}(x - \lambda_l^{[n]});\ \ \frac{n-k}{n}(x - \theta_k^{[n]}) + \zeta_{k,l}^{[n]}(x - \lambda_l^{[n]}), \\ - \sqrt{\frac{n}{l-k}}\sqrt{\frac{n-l}{k}}\Big) \\ - Z\Big(\frac{n-l}{n}(x - \lambda_l^{[n]}) + \eta_{k,l}^{[n]}(\theta_k^{[n]} - x);\ \ \frac{l}{n}(\lambda_l^{[n]} - x) + \eta_{k,l}^{[n]}(\theta_k^{[n]} - x), \\ \sqrt{\frac{n}{l-k}}\sqrt{\frac{k}{n-l}}\Big),$$

where $\zeta_{k,l}^{[n]}$ and $\eta_{k,l}^{[n]}$ are given by (11). □

Now we can compute any second-order Bonferroni bound.

**Example 3** — estimator of the location of a change. The usual estimator of the location of an existing change (i. e. if $H$ has been rejected) is

$$\widehat{\kappa}^{[n]} := \mathrm{argmax}_{1\le k\le n-1}\ T_k^{[n]}.$$

Its asymptotic distribution is known, see [2]. We show an estimate on its successfulness in correct detection of the point of change. Observe that

$$\begin{aligned}\Pr[\widehat{\kappa}^{[n]} = k] &= \Pr[T_k^{[n]} - T_l^{[n]} > 0 \text{ for all } l \neq k] \\ &= 1 - \Pr[T_k^{[n]} - T_l^{[n]} \le 0 \text{ for some } l \neq k] \\ &\ge 1 - \sum_{l\neq k} \Pr[T_k^{[n]} - T_l^{[n]} \le 0].\end{aligned}$$

With (13) we have

$$T_k^{[n]} - T_l^{[n]} = \begin{cases} (\theta_k^{[n]} - \lambda_l^{[n]}) + (\alpha_1 - \beta_1)A - (\alpha_2 - \beta_2)B - (\alpha_3 + \beta_3)C & \text{if } k < l, \\ (\lambda_l^{[n]} - \theta_k^{[n]}) - (\alpha_1 - \beta_1)A + (\alpha_2 - \beta_2)B + (\alpha_3 + \beta_3)C & \text{if } k > l, \end{cases}$$

and we can easily evaluate $\Pr[T_l^{[n]} - T_k^{[n]} \le 0]$. For instance, if $\sigma^2 = 1$, $n = 10$, $\delta = -3$ and $\kappa = 8$, then $\Pr[\widehat{\kappa}^{[n]} = \kappa] \ge 0.7$. Although it is clear that in this set-up the changepoint is very significant, it is not obvious that $\widehat{\kappa}^{[n]}$ *identifies the true value* $\kappa$ *exactly* with high probability. Another example: if $\sigma^2 = 1$, $n = 30$, $\kappa = 15$ and $\delta = -3$, then the probability that $\widehat{\kappa}^{[n]}$ misses $\kappa$ is smaller than 0.2.

ACKNOWLEDGEMENT

This work was supported by the Internal Grant Agency of University of Economics, Prague.

(Received March 1, 2010)

REFERENCES

[1] J. Antoch and M. Hušková: Permutation tests in change point analysis. Statist. Probab. Lett. *53* (2001), 37–46.

[2] J. Antoch, M. Hušková and D. Jarušková: Off-line statistical process control. In: Multivariate Total Quality Control, Physica-Verlag, Heidelberg 2002, 1–86.

[3] J. Antoch, M. Hušková, and Z. Prášková: Effect of dependence on statistics for determination of change. J. Statist. Plan. Infer. *60* (1997), 291–310.

[4] J. Antoch and D. Jarušková: Testing a homogeneity of stochastic processes. Kybernetika *43* (2007), 415–430.

[5] M. Černý and M. Hladík: The regression tolerance quotient in data analysis. In: Proc. 28th Internat. Conf. on Mathematical Methods in Economics 2010 (M. Houda and J. Friebelová, eds.), University of South Bohemia, České Budějovice *1* (2010), 98–104.

[6] X. Chen: Inference in a simple change-point problem. Scientia Sinica *31* (1988), 654–667.

[7] K. Dohmen: Improved inclusion-exclusion identities and Bonferroni inequalities with applications to reliability analysis of coherent systems.
Internet: `citeseer.ist.psu.edu/550566.html`.

[8] K. Dohmen and P. Tittman: Inequalities of Bonferroni-Galambos type with applications to the Tutte polynomial and the chromatic polynomial. J. Inequal. in Pure and Appl. Math. *5* (2004), art. 64.

[9] E. Gombay and L. Horváth: Approximations for the time of change and the power function in change-point models. J. Statist. Plan. Infer. *52* (1996), 43–66.

[10] J. Galambos: Bonferroni inequalities. Ann. Prob. *5* (1997), 577–581.

[11] J. Galambos and I. Simonelli: Bonferroni-type Inequalities with Applications. Springer Verlag, Berlin 1996.

[12] M. Hladík and M. Černý: New approach to interval linear regression. In: 24th Mini-EURO Conference On Continuous Optimization and Information-Based Technologies in The Financial Sector MEC EurOPT 2010, Selected Papers (R. Kasımbeyli et al., eds.), Technika, Vilnius (2010), 167–171.

[13] A. Sen and M. Srivastava: On tests for detecting change in mean. Ann. Statist. *3* (1975), 98–108.

[14] K. Worsley: Testing for a two-phase multiple regression. Technometrics *25* (1983), 35–42.

*Michal Černý, Department of Econometrics, University of Economics Prague, Winston Churchill Square 4, 130 67 Praha 3. Czech Republic.*

*e-mail: cernym@vse.cz*