ROBUSTNESS OF TWO-LEVEL TESTING PROCEDURES UNDER DISTORTIONS OF FIRST LEVEL STATISTICS

A.L. Kostevich, I.S. Nikitina
National Research Center for Applied Problems of Mathematics and Informatics
Belarusian State University, Minsk, BELARUS
e-mail: kostevich@bsu.by

Abstract

We investigate robustness of some two-level testing procedures under distortions induced by using an asymptotic distribution of first level statistics instead of an exact one. We demonstrate that ignoring the distortions results in unreliable conclusions and we propose robustness conditions for the two-level procedures.

1 Introduction. Mathematical model

When studying random number generators, analysing cryptographic algorithms, performing statistical simulation, investigating genetic data researcher often needs to apply a number of statistical tests to a large amount of data. The widely used technique in this case is to use two-level procedures for multiple hypotheses testing [2]. However using an asymptotic distribution of first level statistics instead of an exact one (see, e.g. [4]) may result in an unreliable conclusion of the procedure.

Consider testing the null hypothesis $\mathcal{H}_0$ against the alternative $\mathcal{H}_1$ on K independent samples $X^{(1)}, \dots, X^{(K)}$, each of size n. On the first level the two-level testing procedure applies a test C_1 to the samples and computes test statistics $S_1^{(1)}, \dots, S_1^{(K)}$ and corresponding P-values $P^{(1)}, \dots, P^{(K)}$. If $S_1^{(i)}$ has a continuous distribution with a c.d.f. F and $\mathcal{H}_0$ holds then $P^{(1)}, \dots, P^{(K)}$ are independent and have uniform distribution $U[0,1]$. The second level of performing the two-level procedure is to compare the empirical distribution of $\{P^{(i)}\}$ with the uniform distribution via a goodness-of-fit test C_2 such the Kolmogorov test, chi-square test, etc. So the two-level procedure is

$$Proc(\{P^{(i)}\}, F_{U[0,1]}, \alpha) = \begin{cases} \text{decide } \mathcal{H}_0, & \text{if } S_{Proc}(\{P^{(i)}\}, F_{U[0,1]}) < \Delta(\alpha), \\ \text{decide } \mathcal{H}_1, & \text{otherwise,} \end{cases} \tag{1}$$

where $S_{Proc} = S_{Proc}(\{P^{(i)}\}, F_{U[0,1]})$ is a statistic of the test C_2, $F_{U[0,1]}$ is the theoretical c.d.f. of P-values, α is a fixed significance level, $\Delta(\alpha)$ is a threshold of the test C_2.

When calculating P-values one usually uses a limiting c.d.f. F instead of the exact c.d.f. F_n of the statistic $S_1^{(i)}$ ($F_n \to F$ as $n \to \infty$). Therefore the calculated P-values are distorted and their c.d.f. is in a neighbourhood of the uniform distribution c.d.f. $F_{U[0,1]}$. We assume that this neighbourhood is the Levy neighbourhood of the form

$$\mathcal{P}_\varepsilon(F_0) = \{F | (\forall t) F_0(t-\varepsilon) - \varepsilon \le F(t) \le F_0(t+\varepsilon) + \varepsilon\}, \quad \varepsilon > 0.$$

We denote the distorted P-values by $\{P^{(i)}_{\varepsilon(n)}\}$ and their c.d.f. by $F_{\varepsilon(n)}$:

$$F_{\varepsilon(n)} \in \mathcal{P}_{\varepsilon(n)}(F_{U[0,1]}), \quad \varepsilon(n) \to 0, \quad n \to \infty. \tag{2}$$

On the second level of the procedure (1) one should use either the "plain" statistic $S_{Proc} = S_{Proc}(\{P^{(i)}\}, F_{U[0,1]})$ or the "distorted" statistic $\tilde{S}_{Proc} = S_{Proc}(\{P^{(i)}_{\varepsilon(n)}\}, F_{\varepsilon(n)})$. If the distortions (2) take place and $F_{\varepsilon(n)}$ is unknown then both the statistics are unavailable. So one has to use the procedure (1) with the statistic $S_{\varepsilon(n),Proc} = S_{Proc}(\{P^{(i)}_{\varepsilon(n)}\}, F_{U[0,1]})$:

$$Proc(\{P^{(i)}_{\varepsilon(n)}\}, F_{U[0,1]}, \alpha) = \begin{cases} \text{decide } \mathcal{H}_0, & \text{if } S_{Proc}(\{P^{(i)}_{\varepsilon(n)}\}, F_{U[0,1]}) < \Delta(\alpha), \\ \text{decide } \mathcal{H}_1, & \text{otherwise.} \end{cases} \tag{3}$$

Define the Type I Error probability of the procedure (3) as

$$a_{\varepsilon(n),K} = \Pr[]Proc(\{P^{(i)}_{\varepsilon(n)}\}, F_{U[0,1]}, \alpha) = \mathcal{H}_1 \mid \mathcal{H}_0 \text{ is true.} \tag{4}$$

Since we use the c.d.f. $F_{U[0,1]}$ instead of $F_{\varepsilon(n)}$ the Type I Error probability may not be equal to α. Clearly, it tends to 1 as $K \to \infty$ if n is fixed and the test C_2 is consistent. Thus the procedure (3) may result in an unreliable conclusion.

We investigate robustness of this procedure under distortions (2), more precisely, we find conditions for convergence $a_{\varepsilon(n),K} \to \alpha$ as $n, K \to \infty$ for some widely used two-level procedures.

2 Robustness of two-level procedures

2.1 Using the Kolmogorov test as C_2

Let us investigate robustness of the two-level procedure (3) based on the Kolmogorov test C_2 under the assumption of the distortions (2). Since both the "plain" statistic S_{Kolm} and the "distorted" statistic $\tilde{S}_{Kolm}$ are unavailable,

$$S_{Kolm} = \sqrt{K} \sup_{0 \le x \le 1} \left| 1/K \sum_{i=1}^{K} \mathbb{I}\{P^{(i)} \le x\} - x \right|,$$

$$\tilde{S}_{Kolm} = \sqrt{K} \sup_{0 \le x \le 1} \left| 1/K \sum_{i=1}^{K} \mathbb{I}\{P^{(i)}_{\varepsilon(n)} \le x\} - F_{\varepsilon(n)}(x) \right|,$$

one has to use the statistic $S_{Kolm,\varepsilon(n)} = \sqrt{K} \sup_{0 \le x \le 1} \left| 1/K \sum_{i=1}^{K} \mathbb{I}\{P^{(i)}_{\varepsilon(n)} \le x\} - x \right|$. Let us investigate properties of the procedure (3) based on the statistic $S_{Kolm,\varepsilon(n)}$.

Statement 1 ([1]). *Under the distortions* (2) *if n is fixed and $F_{\varepsilon(n)}(x) \ne F_{U[0,1]}(x)$, then the Type I Error probability of the procedure* (3) *tends to 1: $a_{\varepsilon(n),K} \to 1$, $K \to \infty$.*

The next theorem gives the condition for convergence of the Type I Error probability (4) to the desired significance level α.

Theorem 1. *Under $\mathcal{H}_0$ if $\sqrt{K}\varepsilon(n) \to 0$ as $n, K \to \infty$, then $|S_{Kolm,\varepsilon(n)} - \tilde{S}_{Kolm}| \xrightarrow{a.s.} 0$ and $a_{\varepsilon(n),K} \to \alpha$ as $n, K \to \infty$.*

2.2 Using the chi-square test as C_2

Consider now the procedure (3) based on the chi-square test. When conducting this test one divides the interval $[0;1]$ into M sub-intervals and calculates the numbers $\nu_j(\{P^{(i)}\})$ of $P^{(i)}$'s in the j-th sub-interval for $j = \overline{1,M}$. Let $p_j^{(0)}$ (and $p_j^{(n)}$) be the probability of the event that $P^{(i)}$ ($P_{\varepsilon(n)}^{(i)}$ accordingly) lies within the j-th sub-interval.

Lemma 1. *Under $\mathcal{H}_0$ if c.d.f. $F_{\varepsilon(n)}$ of $\{P_{\varepsilon(n)}^{(i)}\}$ belongs to $\mathcal{P}_{\varepsilon(n)}$, then there exist $\varepsilon_j(n) \in \mathbb{R}$, $j = \overline{1,M}$, $\varepsilon_j(n) \to 0$ as $n \to \infty$, such that $p_j^{(n)} = p_j^{(0)}(1+\varepsilon_j(n))$.*

Under the assumption (2) one should use the "distorted" statistic $\tilde{S}_{\chi^2}$ instead of the "plain" statistic S_{χ^2},

$$\tilde{S}_{\chi^2} = \sum_{j=1}^{M}(\nu_j(\{P_{\varepsilon(n)}^{(i)}\}) - Kp_j^{(n)})^2/(Kp_j^{(n)}), \quad S_{\chi^2} = \sum_{j=1}^{M}(\nu_j(\{P^{(i)}\}) - Kp_j^{(0)})^2/(Kp_j^{(0)}).$$

As this statistic is unavailable too, one has to use the statistic

$$S_{\chi^2,\varepsilon(n)} = \sum_{j=1}^{M}(\nu_j(\{P_{\varepsilon(n)}^{(i)}\}) - Kp_j^{(0)})^2/(Kp_j^{(0)}).$$

Investigate now the properties of the procedure (3) based on the statistic $S_{\chi^2,\varepsilon(n)}$.

Statement 2 ([1]). *Under the distortions (2) if $(p_1^{(n)},\dots,p_M^{(n)})' \neq (p_1^{(0)},\dots,p_M^{(0)})'$ and n is fixed, then the Type I Error probability (4) tends to 1: $a_{\varepsilon(n),K} \to 1$, $K \to \infty$.*

The next theorem gives the condition for convergence of the Type I Error probability (4) to the desired significance level α.

Theorem 2. *Let $|\varepsilon_j(n)| \le t(n), j = \overline{1,M}$ and $t(n) \to 0$, $n \to \infty$. Under $\mathcal{H}_0$ if $\sqrt{K}t(n) \to 0$, $n \to \infty$, then $\mathsf{E}\left\{(S_{\chi^2,\varepsilon(n)} - \tilde{S}_{\chi^2})^2\right\} \to 0$ and $a_{\varepsilon(n),K} \to \alpha$ as $n, K \to \infty$.*

2.3 Using the aggregated data test as C_2

Investigate now robustness of the two-level procedure (3) based on an aggregated data test [3]. The statistic of the test C_2 is based on the proportion of accepted hypotheses on the significance level α_c: $\nu_0(\{P^{(i)}\}) = \frac{1}{K}\sum_{i=1}^{K}\mathbb{I}\{P^{(i)} \ge \alpha_c\}$ and has the form $S_B = \sqrt{K}\frac{1-\alpha_c-\nu_0(\{P^{(i)}\})}{\sqrt{\alpha_c(1-\alpha_c)}}$. Under the assumption of the distortions (2) one should use the "distorted" statistic $\tilde{S}_B = \sqrt{K}\frac{1-F_{\varepsilon(n)}(\alpha_c)-\nu_0(\{P_{\varepsilon(n)}^{(i)}\})}{\sqrt{F_{\varepsilon(n)}(\alpha_c)(1-F_{\varepsilon(n)}(\alpha_c))}}$ instead of the "plain" statistic. As this statistic is unavailable too, one has to use the statistic $S_{B,\varepsilon(n)} = \sqrt{K}\frac{1-\alpha_c-\nu_0(\{P_{\varepsilon(n)}^{(i)}\})}{\sqrt{\alpha_c(1-\alpha_c)}}$.

Let us investigate the properties of the procedure (3) based on the statistic $S_{B,\varepsilon(n)}$.

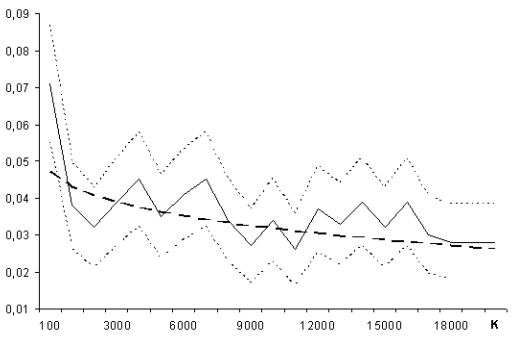
Theorem 3. *Let n is fixed. Under the distortions (2) and $K \to \infty$ the Type I Error probability of the procedure (3) tends to 0, if $F_{\varepsilon(n)}(\alpha_c) < \alpha_c$, and tends to 1, if $F_{\varepsilon(n)}(\alpha_c) > \alpha_c$.*

The next theorem gives the condition for convergence of the Type I Error probability to the desired significance level α.

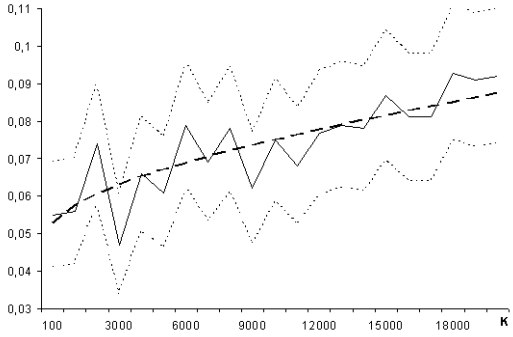
Theorem 4. *Under* $\mathcal{H}_0$ *if* $\sqrt{K}\varepsilon(n) \xrightarrow[n,K\to\infty]{} 0$, *then* $|S_{B,\varepsilon(n)} - \tilde{S}_B| \xrightarrow{a.s.} 0$ *and* $a_{\varepsilon(n),K} \to \alpha$ *as* $n, K \to \infty$. *Moreover,* $a_{\varepsilon(n),K} - \alpha = O(\sqrt{K}\varepsilon(n))$.

3 Simulation study

We consider the procedure from subsection 2.3 which uses the monobit test as the test C_1. The statistic of the monobit test is the number of ones in a sample. Under $\mathcal{H}_0$ the statistic has the binomial distribution, but in practice one uses the approximation by the normal distribution. The figures below present the estimates of the Type I Error probability $a_{\varepsilon(n),K}$ (solid line) with the 95-% confidence interval (dotted line) and the theoretical values of $a_{\varepsilon(n),K}$ (dashed line) for $\alpha_c = 0.0455$ (in this case $F_{\varepsilon(n)}(\alpha_c) - \alpha_c = -0.00042 < 0$) and $\alpha_c = 0.049$ ($F_{\varepsilon(n)}(\alpha_c) - \alpha_c = 0.00043 > 0$), $n = 65536$, $\varepsilon(n) \approx 0.0015$, $\alpha = 0.05$. One can see that $a_{\varepsilon(n),K}$ tends to 0 or 1 subject to the sign of $F_{\varepsilon(n)}(\alpha_c) - \alpha_c$. As consistent with the results of Theorem 4 $a_{\varepsilon(n),K} \approx \alpha = 0.05$ if $\sqrt{K}\varepsilon(n)$ is nearly 0.



$\alpha_c = 0.0455$



$\alpha_c = 0.049$

References

[1] Kendall M. G., Stuart A. (1973). *The advanced theory of statistics*, vol. II. Griffin, London.

[2] L'Ecuyer P. (1998). *Random Number Generators*. Handbook of Simulation, Wiley.

[3] NIST Special Publication 800-22. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications, 2000.

[4] Selivanov B.I. (2006). On the exact computation of decomposable statistics distributions for polynomial scheme. *Discrete Mathematics*, v. **18**, No. 3, pp. 85–94. (In Russian)