NEW STATISTICAL BOX-TEST AND ITS POWER

Igor Semaev and Mehdi M. Hassanzadeh

The Selmer Center,
Department of Informatics, University of Bergen
PB 7803, N-5020 Bergen, Norway,
e-mails: igor@ii.uib.no and
Mehdi.Hassanzadeh@ii.uib.no.

7 July 2011

Abstract. In this paper, statistical testing of N multinomial probabilities is studied and a new box-test, called *Quadratic Box-Test*, is introduced. The statistics of the new test has χ^2_s limit distribution as N and the number of trials n tend to infinity, where s is a parameter. The well-known empty-box test is a particular case for $s = 1$. The proposal is quite different from Pearson's goodness-of-fit test, which requires fixed N while the number of trials is growing, and linear box-tests. We prove that under some conditions on tested distribution the new test's power tends to 1. That defines a wide region of non-uniform multinomial probabilities distinguishable from the uniform. For moderate N an efficient algorithm to compute the exact values of the first kind error probability is devised.

Keywords: Statistical Testing, Chi-square Goodness-of-fit Test, Allocation Problem, Empty-Box Test, Linear Box-Test, Quadratic Box-Test, Probability of Errors.

1 INTRODUCTION

The security of most cryptographic systems depends upon a random sequence. For example, the secret key in block ciphers and stream ciphers, the primes p, q in RSA encryption and digital signature schemes, the nonce in most authentication protocols. As "a true random sequence" is a theoretical abstraction, its producing is not possible. Therefore a *pseudorandom sequence*, often generated by a deterministic algorithm, is used in cryptography instead. Ideally, it should be indistinguishable from a true random sequence within available computer power. Various statistical tests can be applied to check this.

In this paper, a new statistical test, named *Quadratic Box-Test*, is presented. It can be used for randomness evaluation and distinguishing attacks in cryptography. The main idea of our approach is to compare the distribution of repeated patterns in the tested data with a true random data. In Section 2, a theoretical background and related work are presented. In Section 3 the new test is introduced and in Section 4, which is the main part of our contribution, we will prove that its power tends to 1 when N tends to infinity. The first kind error probability of the test for low and moderate N is computed in Section 5, where a relatively efficient algorithm is devised. In Section 6, an application to functions with finite number of outputs is discussed. We will conclude in Section 7.

2 THEORETICAL BACKGROUND OF BOX-TEST

The problem of computing the box-test is related to the classical shot problem. Let n particles be allocated into N boxes, where the k-th box appears with the probability a_k and $a = (a_1, \ldots, a_N)$. Let $\mu_r(a)$ denote the number of boxes with exactly r particles. In Theorem 2.1.1 of [6] it was proved that in case $a = h$, where $h = (\frac{1}{N}, \ldots, \frac{1}{N})$, we have:

$$\mathbf{E}\mu_r(h) = Np_r + O(1),$$
$$\mathbf{Cov}(\mu_r(h), \mu_t(h)) = N\sigma_{rt} + O(1),$$

where $\alpha = \frac{n}{N}$, $p_r = \frac{\alpha^r}{r!}e^{-\alpha}$, and σ_{rt} are entries of the limit covariance matrix $\mathbf{B}$. They are defined by

$$\begin{aligned}\sigma_{rr} &= p_r\,(1 - p_r - p_r\frac{(\alpha - r)^2}{\alpha}), \qquad (1)\\ \sigma_{rt} &= -p_r p_t\,(1 + \frac{(\alpha - r)(\alpha - t)}{\alpha}).\end{aligned}$$

Generally, for the box probabilities $a = (a_1, \ldots, a_N)$ we have:

$$
\begin{aligned}
p_{rk} &= \frac{(\alpha N a_k)^r}{r!} e^{-\alpha N a_k}, \qquad p_r(a) = \frac{1}{N}\sum_{k=1}^{N} p_{rk}, \\
\sigma_{rr}(a) = & \quad \frac{1}{N}\sum_{k=1}^{N} p_{rk} - \frac{1}{N}\sum_{k=1}^{N} p_{rk}^2 - \frac{1}{\alpha}\left[\frac{1}{N}\sum_{k=1}^{N} p_{rk}(\alpha N a_k - r)\right]^2, \quad (2) \\
\sigma_{rt}(a) = & - \frac{1}{N}\sum_{k=1}^{N} p_{rk}p_{tk} \\
& - \frac{1}{\alpha}\left[\frac{1}{N}\sum_{k=1}^{N} p_{rk}(\alpha N a_k - r)\right]\left[\frac{1}{N}\sum_{k=1}^{N} p_{tk}(\alpha N a_k - t)\right].
\end{aligned}
$$

where $\sigma_{rt}(a)$ are entries of a matrix $\mathbf{A}$. Theorem 3.1.5 in [6] states that if N tends to infinity and $Na_k \leq C$ for a constant C, and $\alpha_0 \leq \alpha \leq \alpha_1$, then

$$
\begin{aligned}
\mathbf{E}\mu_r(a) &= Np_r(a) + O(1), \\
\mathbf{Cov}(\mu_r(a), \mu_t(a)) &= N\sigma_{rt}(a) + O(1).
\end{aligned}
$$

Additionally, according to the Theorem 3.5.2 in [6], under the same conditions the multivariate random variable

$$
\nu(a) = \left(\frac{\mu_{r_1}(a) - \mathbf{E}\mu_{r_1}(a)}{\sqrt{N}}, \ldots, \frac{\mu_{r_s}(a) - \mathbf{E}\mu_{r_s}(a)}{\sqrt{N}}\right)
$$

asymptotically has multivariate normal distribution as N and n tend to infinity. We assume those conditions fulfilled throughout this article. The asymptotical normality of $\nu(a)$ may be used to check whether a multinomial sample was produced with prescribed box probabilities for large enough N. We are going to test the hypothesis $a = h$.

Any such test is naturally to call a *box-test*. For instance, a test based on the distribution of $\frac{\mu_0 - \mathbf{E}\mu_0}{\sqrt{N}}$ is called *empty box-test* and was introduced by David in [3]. It may have some advantage over Pearson's χ^2 goodness-of-fit test, which requires $\alpha = \frac{n}{N} \to \infty$ to approach limit distribution; see [5, 8].

2.1 Linear Box-Test

A *linear box-test*, which is a generalization of the empty-box test, was studied in [6]. It is defined by the dot-product $\nu(a)c$, where c is a constant vector of length s. Linear box-test statistic has asymptotically normal distribution too. The random vector

$$\left(\frac{\mu_{r_1}(a) - Np_{r_1}(a)}{\sqrt{N}}, \ldots, \frac{\mu_{r_s}(a) - Np_{r_s}(a)}{\sqrt{N}}\right)$$

has the same limit distribution as $\nu(a)$ and is denoted with the same character in this section. Similarly, we put

$$\eta(a) = \left(\frac{\mu_{r_1}(a) - Np_{r_1}(h)}{\sqrt{N}}, \ldots, \frac{\mu_{r_s}(a) - Np_{r_s}(h)}{\sqrt{N}}\right).$$

Let $c = (c_1, \ldots, c_s)$ be any real vector, whose entries do not depend on N. The random variable νc asymptotically as N tends to infinity has normal distribution with variance $c\mathbf{B}c$ and expectation 0, denoted $\mathbf{N}(0, \sqrt{c\mathbf{B}c})$. Let $0 < \epsilon < 1$ be a required significance level. From $\mathbf{N}(0, \sqrt{c\mathbf{B}c})$ distribution tables one finds D_ϵ such that

$$\mathbf{Pr}(|\mathbf{N}(0, \sqrt{c\mathbf{B}c})| \geq D_\epsilon) = \epsilon.$$

An allocation of n particles into N boxes is observed and statistic $\eta(a)c$ is computed. If $|\eta(a)c| \leq D_\epsilon$, then the hypothesis $a = h$ is accepted and otherwise rejected.

Example I: We take the statistic $\eta(a)c$ to depend only on μ_0 and μ_1 and put $c = (1, 1)$. Let $\alpha = 1$, then $p_0 = p_1 = e^{-1}$. Therefore,

$$\eta(a)c = \frac{\mu_0(a) + \mu_1(a) - 2Ne^{-1}}{\sqrt{N}}$$

and

$$\mathbf{B} = \frac{1}{e^2} \times \begin{pmatrix} e-2 & -1 \\ -1 & e-1 \end{pmatrix}.$$

Then $c\mathbf{B}c = \frac{2e-5}{e^2}$. The distribution of $\nu c = \eta(h)c$ becomes close to $\mathbf{N}(0, \sqrt{\frac{2e-5}{e^2}})$ as N grows. We put, for instance, $\epsilon = 0.1$ and find the quantile $D_\epsilon = 0.3998$.

Let $n = N = 20$ and the observed sequences of outcomes(boxes) is

$$19, 18, 5, 6, 17, 20, 14, 17, 3, 16, 20, 6, 3, 15, 7, 8, 7, 12, 14, 5. \qquad (3)$$

One finds $\mu_0 = 7$ as boxes numbered $1, 2, 4, 9, 10, 11, 13$ are absent, and $\mu_1 = 6$ as boxes numbered $8, 12, 15, 16, 18, 19$ appear just once, and $\mu_2 = 7$ as boxes $3, 5, 6, 7, 14, 17, 20$ appear twice. No box appears more than twice. So $\eta(a)c = -0.3835$ and as $|\eta(a)c| \leq 0.3998$ the hypothesis "multinomial distribution is uniform" is accepted with the first kind error probability at most 10%(in fact, the real value of the error probability is something different as N is fairly small here).

3 Quadratic Box-Test

In this section, our statistical test, called *Quadratic Box-Test*, is defined. It will be proved in Section 4 that under condition $N^{\frac{3}{2}}\sum_{k=1}^{N}(a_k - \frac{1}{N})^2 \to \infty$ for non-uniform distribution a, the power of quadratic box-test tends to 1 when the number of possible patterns, N, tends to infinity. That defines a set of non-uniform distributions a distinguishable by this test with probability tending to 1.

The test was found during a study on cryptographic hash-functions. A good hash-function should have values indistinguishable from those produced with multinomial uniform probabilities. Hash-function values are naturally to consider as allocations into boxes labeled with its different values. According to NIST requirements, the total number of a hash function different values may be as big as 2^{512} [7]. Therefore, in order to apply a box-test the values are split into N regions of equal probability.

Suppose that an allocation of n particles into N boxes is observed and only the values $\mu_{r_1}, \ldots, \mu_{r_s}$ are computed. Let again

$$\eta(a) = \left(\frac{\mu_{r_1}(a) - Np_{r_1}(h)}{\sqrt{N}}, \ldots, \frac{\mu_{r_s}(a) - Np_{r_s}(h)}{\sqrt{N}}\right),$$

where a is the tested box distribution. The statistic of quadratic box-test is the quadratic form $\eta\mathbf{B}^{-1}\eta$ where $\mathbf{B}$ is the limit covariance matrix for $\nu = \nu(h)$ with entries σ_{rt} defined by (1).

Standard argument ([5], Section 15.10) shows that $\nu\mathbf{B}^{-1}\nu$ has asymptotically χ^2_s-distribution as N tends to infinity. From χ^2_s-distribution tables one finds C_ε such that $\mathbf{Pr}(\chi^2_s \geq C_\varepsilon) = \varepsilon$, where ε is the significance level probability. If $\eta\mathbf{B}^{-1}\eta \leq C_\epsilon$, then the hypothesis $a = h$ is accepted, otherwise rejected. When $s = 1$ and $r_1 = 0$ the quadratic test is equivalent to the empty-box test.

For $a = h$ we have $\eta\mathbf{B}^{-1}\eta = \nu\mathbf{B}^{-1}\nu$. By the limit Theorem, the test's first kind error probability $\mathbf{Pr}(\nu\mathbf{B}^{-1}\nu \geq C_\varepsilon) \to \varepsilon$ as $N \to \infty$. In Section **??** the exact values of $\mathbf{Pr}(\nu\mathbf{B}^{-1}\nu \geq C_\varepsilon)$ for some $\mu = (\mu_{r_1}, \ldots, \mu_{r_s})$, $s = 1, 2, 3, 4$ and low N are presented. Numerical results demonstrate that the convergence rate depends on r_i and may be slow. Therefore, a test only based on the limit probability might not be reliable for such N.

Example II: We want the statistic $\eta\mathbf{B}^{-1}\eta$ to depend only on μ_0 and μ_1. Let $\alpha = 1$ as Example I. One computes

$$\mathbf{B^{-1}} = \frac{e^2}{e^2 - 3e + 1} \times \begin{pmatrix} e-1 & 1 \\ 1 & e-2 \end{pmatrix}$$

and

$$\eta(a) = \left(\frac{\mu_0(a) - Ne^{-1}}{\sqrt{N}}, \frac{\mu_1(a) - Ne^{-1}}{\sqrt{N}} \right).$$

Therefore,

$$\begin{aligned} \eta\mathbf{B}^{-1}\eta &= \frac{e^2 N^{-1}}{(e^2 - 3e + 1)} \\ &\times \begin{pmatrix} \mu_0 - Ne^{-1} \\ \mu_1 - Ne^{-1} \end{pmatrix}^t \begin{pmatrix} e-1 & 1 \\ 1 & e-2 \end{pmatrix} \begin{pmatrix} \mu_0 - Ne^{-1} \\ \mu_1 - Ne^{-1} \end{pmatrix}. \end{aligned} \tag{4}$$

As N grows, the distribution of $\eta\mathbf{B}^{-1}\eta$ becomes close to χ_2^2 for $a = h$. We put $\epsilon = 0.1$ and find $C_\epsilon = 4.6051$. For the outcomes (3), where $n = N$, $\mu_0 = 7$ and $\mu_1 = 6$, we compute $\eta\mathbf{B}^{-1}\eta = 3.9664 < C_\epsilon$. Therefore the hypothesis "multinomial distribution is uniform" is accepted with the first kind error probability at most 10%. With the method described in Section 5 we compute that the real error probability is about 8%.

4 Power of the Quadratic Box-Test

In this section, we prove that our test is consistent when n and N tends to infinity for some non-uniform a. The second kind error probability is the probability to accept $a = h$, whereas this is wrong. It is defined by $\beta(a) = \mathbf{Pr}(\eta\mathbf{B}^{-1}\eta \leq C_\epsilon)$. We will prove $\beta(a)$ tends to zero for those a, or, in other words, the test's power $W_{n,N}(a)$ tends to 1 if $(n, N) \longrightarrow \infty$, as $W_{n,N}(a) = 1 - \beta(a)$.

When N tends to infinity, under the uniformity condition $a = h$, the distribution of $\eta\mathbf{B}^{-1}\eta$ tends to the distribution of χ_s^2 and its expectation tends to s, which is a constant. First, we prove that if the multinomial distribution a satisfies some restrictions, and in particular it is not uniform, then the expectation of $\eta\mathbf{B}^{-1}\eta$ tends to infinity. Then we will prove that $W_{n,N}(a) \to 1$ when $(n, N) \to \infty$. Let

$$\delta = \left(\frac{\mathbf{E}\mu_{r_1}(a) - \mathbf{E}\mu_{r_1}(h)}{\sqrt{N}}, \ldots, \frac{\mathbf{E}\mu_{r_s}(a) - \mathbf{E}\mu_{r_s}(h)}{\sqrt{N}} \right) \tag{5}$$

so that $\eta(a) = \nu(a) + \delta$.

Theorem 1. $\mathbf{E}(\eta\mathbf{B}^{-1}\eta) \to \infty$ *if and only if* $|\delta| \to \infty$.

Proof. We write $\eta = \nu + \delta$. Therefore,

$$\mathbf{E}(\eta\mathbf{B}^{-1}\eta) = \mathbf{E}(\nu\mathbf{B}^{-1}\nu) + 2\mathbf{E}(\delta\mathbf{B}^{-1}\nu) + \delta\mathbf{B}^{-1}\delta,$$

where $\mathbf{E}(\delta\mathbf{B}^{-1}\nu) = \delta\mathbf{B}^{-1}\mathbf{E}(\nu) = 0$. Then $\nu\mathbf{B}^{-1}\nu \geq 0$ as $\mathbf{B}^{-1}$ is positive definite. So if $\delta\mathbf{B}^{-1}\delta$ tends to infinity, then $\mathbf{E}(\eta\mathbf{B}^{-1}\eta)$ tends to infinity. The former is true if and only if $|\delta| \to \infty$.

We can write $\mathbf{E}(\nu\mathbf{B}^{-1}\nu) \leq c\,\mathbf{E}(|\nu|^2)$, where c is a constant dependent on $\mathbf{B}$. The latter is bounded by the maximal of

$$\mathbf{E}\left(\frac{\mu_{r_k}(a) - \mathbf{E}\mu_{r_k}(a)}{\sqrt{N}} \right)^2 = \frac{\mathbf{Cov}(\mu_{r_k}(a), \mu_{r_k}(a))}{N} \tag{6}$$

times a positive constant defined by $\mathbf{B}$. With (2), the value (6) is bounded in case $Na_k \leq C$ and $\alpha_0 \leq \alpha \leq \alpha_1$. So as N tends to infinity $\mathbf{E}(\nu\mathbf{B}^{-1}\nu)$ is bounded too. Therefore, $\mathbf{E}(\eta\mathbf{B}^{-1}\eta) \to \infty$ if and only if $\delta\mathbf{B}^{-1}\delta \to \infty$. That proves the Theorem. □

We say $Na_k \to 1$ if for any $\tau > 0$ there exists N_τ such that $|Na_k - 1| < \tau$ for all $N > N_\tau$ and $k = 1, \ldots, N$.

Theorem 2. *Assume* $Na_k \to 1$ *for each* k *as* N *tends to infinity. Then* $|\delta| = o(\sqrt{N})$. *If additionally* $(\alpha - r_i)^2 \neq r_i$ *for some* i*, then* $|\delta| \to \infty$ *if and only if* $N^{\frac{3}{2}} \sum_{k=1}^{N} (a_k - \frac{1}{N})^2 \to \infty$.

That defines the area of a, where Theorem 4 is valid. For instance, $a_k = \frac{1}{N} + \frac{\gamma_k}{N^{5/4}}$, where γ_k tends to infinity such that $\gamma_k = o(N^{1/4})$.

We now study conditions for $|\delta| \to \infty$. Consider two events:

$$\left| \frac{\mathbf{E}\mu_r(a) - \mathbf{E}\mu_r(h)}{\sqrt{N}} \right| \to \infty, \tag{7}$$

$$N^{\frac{3}{2}} \sum_{k=1}^{N} b_k^2 \to \infty, \tag{8}$$

where $a_k = \frac{1}{N} + b_k$. Theorem 3 implies that if $(\alpha - r_i)^2 \neq r_i$ for some i, then $|\delta| \to \infty$ if and only if (8).

Theorem 3. *Let $Na_k \to 1$ for each k as N tends to infinity.*

1. (7) *is* $o(\sqrt{N})$,
2. *If* (7) *is hold, then* (8) *is correct,*
3. *Assume* $(\alpha - r)^2 \neq r$, *then* (7) *is hold if and only if* (8) *is hold.*

Proof. $Na_k \to 1$ if and only if $x_k = Nb_k \to 0$. We put $f(x) = (1 + x)^r e^{-\alpha x}$ and with (2) compute

$$\begin{aligned}
\frac{\mathbf{E}\mu_r(a) - \mathbf{E}\mu_r(h)}{\sqrt{N}} &= \frac{\alpha^r e^{-\alpha}}{r!\sqrt{N}} \sum_{k=1}^{N} (f(x_k) - f(0)) + O(\frac{1}{\sqrt{N}}) \\
&= \frac{\alpha^r e^{-\alpha}}{r!\sqrt{N}} \sum_{k=1}^{N} \left((r-\alpha)\, x_k + f''(\theta_k\, x_k)\, \frac{x_k^2}{2} \right) \\
&\quad + O(\frac{1}{\sqrt{N}}) \\
&= \frac{\alpha^r e^{-\alpha}}{r!\sqrt{N}} \sum_{k=1}^{N} f''(\theta_k\, x_k)\, \frac{x_k^2}{2} + O(\frac{1}{\sqrt{N}})
\end{aligned}$$

where $0 \leq \theta_k \leq 1$ and because $\sum_{k=1}^{N} x_k = 0$. There exist two constants c_1 and c_2 such that $c_1 \leq f''(x) \leq c_2$ for all small enough x. Therefore,

$$\begin{aligned}
\frac{\alpha^r e^{-\alpha} c_1}{2\, r!} \left(N^{\frac{3}{2}} \sum_{k=1}^{N} b_k^2 \right) &\leq \frac{\mathbf{E}\mu_r(a) - \mathbf{E}\mu_r(h)}{\sqrt{N}} + O(\frac{1}{\sqrt{N}}) \\
&\leq \frac{\alpha^r e^{-\alpha} c_2}{2\, r!} \left(N^{\frac{3}{2}} \sum_{k=1}^{N} b_k^2 \right) \qquad (9)
\end{aligned}$$

That implies the first and second statements. We compute $f''(0) = (\alpha - r)^2 - r$. If $(\alpha - r)^2 \neq r$, then c_1 and c_2 may be taken both positive or both negative. That implies the last statement. $\square$

Theorem 2 is a corollary of Theorem 3. Now, we want to proof that the power of our test goes to 1 when $(n, N) \to \infty$ and it is done in Theorem 4.

Theorem 4. *Let $|\delta| \to \infty$ as N tends to infinity, then $\beta(a) = O(|\delta|^{-2}) \to 0$, therefore $W_{n,N}(a) \to 1$.*

Proof. First, we estimate the variance of $\eta\mathbf{B}^{-1}\eta$ and then prove the statement with the Chebyshev inequality. We use the notation in Theorem 1, where $\eta(a) = \nu(a) + \delta$, so

$$\eta\mathbf{B}^{-1}\eta = \nu\mathbf{B^{-1}}\nu + 2\delta\mathbf{B^{-1}}\nu + \delta\mathbf{B^{-1}}\delta.$$

Then $\mathbf{Var}(\eta\mathbf{B}^{-1}\eta) = \mathbf{Var}(U_1 + U_2)$, where $U_1 = \nu\mathbf{B^{-1}}\nu$ and $U_2 = 2\delta\mathbf{B^{-1}}\nu$ as $\delta\mathbf{B^{-1}}\delta$ is not a random variable. Therefore,

$$\mathbf{Var}(\eta\mathbf{B}^{-1}\eta) = \mathbf{Var}(U_1) + \mathbf{Var}(U_2) + 2\mathbf{Cov}(U_1, U_2).$$

The variance of $U_1 = \nu\mathbf{B^{-1}}\nu$ is bounded as the coordinates of $\nu(a)$ are asymptotically normal with zero means and bounded covariance matrix $\mathbf{A}$ defined by (2). The latter follows as $Na_k \leq C$. Then

$$\mathbf{Var}(U_2) = 4\delta\mathbf{B^{-1}}\mathbf{A}\mathbf{B}^{-1}\delta = O(|\delta|^2).$$

We also have

$$|\mathbf{Cov}(U_1, U_2)| \leq \sqrt{\mathbf{Var}(U_1)\mathbf{Var}(U_2)} = O(|\delta|).$$

All this implies $\mathbf{Var}(\eta\mathbf{B}^{-1}\eta) = O(|\delta|^2)$. By the Chebyshev inequality, we get

$$\begin{aligned}
\beta(a) &= \mathbf{Pr}\left(\eta\mathbf{B}^{-1}\eta \leq C_\epsilon\right) \\
&\leq \mathbf{Pr}\left(|\,\eta\mathbf{B}^{-1}\eta - \mathbf{E}(\eta\mathbf{B}^{-1}\eta)\,| \geq \mathbf{E}(\eta\mathbf{B}^{-1}\eta) - C_\epsilon\right) \\
&\leq \frac{\mathbf{Var}(\eta\mathbf{B}^{-1}\eta)}{(\mathbf{E}(\eta\mathbf{B}^{-1}\eta) - C_\epsilon)^2} = O\left(\frac{1}{|\delta|^2}\right) \to 0
\end{aligned}$$

as in Theorem's condition $\mathbf{E}(\eta\mathbf{B}^{-1}\eta) \geq c\,|\delta|^2$ for a positive constant c; see the proof of Theorem 1. By assumption $|\delta|$ tends to infinity and C_ϵ is a constant. That proves the Theorem. □

5 First Kind Error Probability for Low N

In this section, the first error probability for low N is studied. For enough large N, this type of error tends to ε which is the significance level, but for low N it is different and calculated in this section. In case of $a = h$ for low and moderate N, the statistic $Q_s = \eta \mathbf{B}^{-1} \eta = \nu \mathbf{B}^{-1} \nu$ is a function of $\mu = (\mu_{r_1}, \ldots, \mu_{r_s})$.

The goal is to compute $\mathbf{Pr}(Q_s \geq C_\varepsilon)$, where C_ε is the χ^2_s-quantile of level ε. This is the first error probability of test. The probability $\mathbf{Pr}(Q_1(\mu_0) \geq C_\varepsilon)$ is computed with a simplified method, where the values $\mathbf{Pr}(\mu_0 = k)$ are found by the recurrent relation (5) in Chapter 1 of [6]. As above we denote

$$\nu = \left(\frac{\mu_{r_1}(h) - Np_{r_1}}{\sqrt{N}}, \ldots, \frac{\mu_{r_s}(h) - Np_{r_s}}{\sqrt{N}} \right).$$

For $C = C_\varepsilon$ we are to compute the probability

$$\mathbf{Pr}(Q_s \leq C_\varepsilon) = \mathbf{Pr}\left(\nu \mathbf{B}^{-1} \nu \leq C\right) = \sum_K \mathbf{Pr}(\mu = K). \tag{10}$$

Over all integer s-vectors K with zero or positive entries such that

$$(\frac{K - Np}{\sqrt{N}}) \mathbf{B}^{-1} (\frac{K - Np}{\sqrt{N}}) \leq C,$$

where $p = (p_{r_1}, \ldots, p_{r_s})$. Let $\mu(n, N) = (\mu_{r_1}(n, N), \ldots, \mu_{r_s}(n, N))$, then by formula (35) in Chapter 2 of [6],

$$\begin{aligned} \mathbf{Pr}[\mu(n, N) = K] &= \mathbf{Pr}\left[\mu(n - (k, r), N - k) = 0\right] \times \\ &\times \frac{N^{[k]}\, n^{[(k,r)]}}{\prod_{i=1}^{s} k_i!(r_i!)^{k_i}} \times \frac{(1 - \frac{k}{N})^{n-(k,r)}}{N^{(k,r)}}, \end{aligned} \tag{11}$$

where $k = k_1 + \ldots + k_s$, $x^{[k]} = x(x-1)\ldots(x-k+1)$ and $(k, r) = k_1 r_1 + \ldots + k_s r_s$. The probability $\mathbf{Pr}[\mu(n - (k, r), N - k) = 0]$ is computed with the recurrent relation:

$$\begin{aligned} \mathbf{Pr}[\mu(n, N) = 0] &= \mathbf{Pr}[\mu(n - t, N - 1) = 0] \times \\ &\times \mathbf{Pr}[\mu(t, 1) = 0] \times \sum_{t=0}^{n} \binom{n}{t} \frac{(N-1)^{n-t}}{N^n}, \end{aligned} \tag{12}$$

where the initial values are

$$\mathbf{Pr}[\mu(n, 1) = 0)] = 0, \quad n \in \{r_1, \ldots, r_s\},$$
$$\mathbf{Pr}[\mu(n, 1) = 0)] = 1, \quad n \notin \{r_1, \ldots, r_s\}.$$

Sometimes it is better to use a more general recurrence. Let $1 \le N_1 < N$, then

$$\begin{aligned} \mathbf{Pr}[\mu(n,N)=0] &= \sum_{t=0}^{n} \binom{n}{t} \left(\frac{N_1}{N}\right)^t \left(1-\frac{N_1}{N}\right)^{n-t} \\ &\times \mathbf{Pr}[\mu(t,N_1)=0] \times \mathbf{Pr}[\mu(n-t,N-N_1)=0]. \end{aligned}$$

Cauchy-Schwarz inequality implies $x\mathbf{B}^{-1}x \ge b_{jj}^{-1}|x_j|^2$, where $\mathbf{B} = (b_{ij})$. From the inequality $x\mathbf{B}^{-1}x \le C$ we get

$$|x_j| \le \sqrt{Cb_{jj}}. \tag{13}$$

Therefore the values k_i used in computing by (10) are restricted by

$$\left| \frac{k_j - Np_{r_j}}{\sqrt{N}} \right| \le \sqrt{C\, b_{jj}}. \tag{14}$$

and may be searched.

We however explain a better approach now. As $\mathbf{B}^{-1}$ is symmetric positive definite, the decomposition $\mathbf{B}^{-1} = UU^T$ is possible, where U is an upper triangular square matrix. Algorithm 1 can be used to compute U such that $V = UU^T$.

Algorithm 1 Compute the upper triangular real matrix $U_{s\times s}$

Input: Real symmetric positive definite $s \times s$ matrix V

1. Compute

$$v_{ij} \leftarrow v_{ij} - \frac{v_{is}v_{js}}{v_{ss}}, \quad \text{and} \quad v_{is} \leftarrow \frac{v_{is}}{\sqrt{v_{ss}}},$$

   for $i = 1,\ldots,s$ and $j = 1,\ldots,s-1$. So that $v_{sj} = 0$ for $j = 1,\ldots,s-1$.

2. First $s-1$ rows and first $s-1$ columns of V make a symmetric positive definite matrix. Put $s \leftarrow s-1$ and apply step 1 to that matrix.

3. Repeat steps above s times. Return V.

Algorithm 1 is in fact reducing the quadratic form xVx. After $\mathbf{B}^{-1}$ was decomposed, we get $x\mathbf{B}^{-1}x = (xU)(xU)^T$. So the inequality $x\mathbf{B}^{-1}x \leq C$ is equivalent to

$$(u_{11}x_1)^2 + (u_{12}x_1 + u_{22}x_2)^2 + \ldots + (u_{1s}x_1 + \ldots + u_{ss}x_s)^2 \leq C$$

and therefore to the inequality system

$$|x_1| \quad \leq \quad \frac{\sqrt{C}}{u_{11}}, \tag{15}$$

$$|x_2 + \frac{u_{12}}{u_{22}}x_1| \leq \frac{\sqrt{C - (u_{11}x_1)^2}}{u_{22}}, \tag{16}$$

$$\ldots,$$

$$\begin{aligned} |x_s \quad &+ \quad \frac{u_{1s}}{u_{ss}}x_1 + \ldots + \frac{u_{1s-1}}{u_{ss}}x_{s-1}| \\ &\leq \quad \frac{\sqrt{C - (u_{11}x_1)^2 - \ldots - (u_{1s-1}x_1 + \ldots + u_{s-1s-1}x_{s-1})^2}}{u_{ss}}. \end{aligned}$$

That gives a clue how to solve $x\mathbf{B}^{-1}x \leq C$ for $x_j = \frac{k_j - Np_{r_j}}{\sqrt{N}}$ and integer k_j efficiently.

Algorithm 2 efficiently computes the first error probability for low N. This algorithm is used to calculate the exact value of first error probability in case of $a = h$ for low and moderate N. Table 1 is calculated for $s = 3$ and $\mu = (\mu_2, \mu_4, \mu_5)$. Tables 2-5 are calculated for $s = 1, 2, 3, 4$ and $\mu = (\mu_0, ..., \mu_{s-1})$ respectively. We take $\varepsilon = 0.01$ and 0.05. So that ε is the limiting value for the probability as N grows to infinity. However even for relatively large N this is not true.

Table 1: $\mathbf{Pr}(Q_3(\mu_2, \mu_4, \mu_5) \geq C_\varepsilon)$

ε, N	2^4	2^5	2^6	2^7	2^8	2^9
0.05	0.0449	0.0907	0.0376	0.0561	0.0510	0.0522
0.01	0.0412	0.0163	0.0181	0.0134	0.0142	0.0120

Algorithm 2 Compute the first kind error probability for low N

Input: N, ϵ or significant level, s and $(r_1, r_2, ..., r_s)$.

1. Pre-compute probabilities $\mathbf{Pr}(\mu(n_2, N_2) = 0)$ for all $n_2 \le n_1$ and $N_2 \le N_1$ with (12), which simplifies to

$$\begin{aligned}\mathbf{Pr}(\mu(n,N) &= 0) \\ &= \sum_{t=s}^{n-sN+s} \binom{n}{t} \frac{(N-1)^{n-t}}{N^n} \mathbf{Pr}(\mu(n-t, N-1) = 0)\end{aligned}$$

in case $\mu = (\mu_0, \dots, \mu_{s-1})$. The values n_1, N_1 are defined below.

2. To compute with (11), we have $n - (k, r) \le n_1$ and $N - k \le N_1$, where we can put

$$n_1 = \lfloor n - \sum_{j=1}^{s} r_j (Np_{r_j} - \sqrt{C\, b_{jj} N}) \rfloor$$

and

$$N_1 = \lfloor N - \sum_{j=1}^{s} (Np_{r_j} - \sqrt{C\, b_{jj} N}) \rfloor$$

as $k_j = Np_{r_j} + \delta_j \sqrt{N}$ and $|\delta_j| \le \sqrt{C\, b_{jj}}$ by (14).

3. One runs over all $K = (k_1, \dots, k_s)$ such that

$$(\frac{K - Np}{\sqrt{N}}) \mathbf{B}^{-1} (\frac{K - Np}{\sqrt{N}}) \le C. \tag{17}$$

So k_1 is taken such that $x_1 = \frac{k_1 - Np_{r_1}}{\sqrt{N}}$ satisfies (15), that is k_i belongs to some interval. Upon fixed k_1, integer k_2 is taken such that $x_2 = \frac{k_2 - Np_{r_2}}{\sqrt{N}}$ satisfies (16), that is from some interval, and so on. If the interval for k_j is empty or exhausted, the algorithm backtracks and takes another k_{j-1}. Any K produced is a solution to (17). The search space is further reduced with the restrictions:

$$k = \sum_{i=1}^{s} k_i \le N,$$

$$(k, r) = \sum_{i=1}^{s} k_i r_i \le n.$$

In case $\mu = (\mu_0, \dots, \mu_{s-1})$ we have additional restriction $n - (k, r) \ge s(N - k)$. Relevant probabilities $\mathbf{Pr}(\mu(n, N) = K)$ are computed by (11) with the pre-computed $\mathbf{Pr}(\mu(n_2, N_2) = 0)$ and summed to $\mathbf{Pr}\left(\nu \mathbf{B}^{-1} \nu \le C\right)$ according to (10).

Table 2: $\mathbf{Pr}(Q_1(\mu_0) \geq C_\varepsilon)$

ε, N	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
0.05	0.0460	0.0505	0.0440	0.0496	0.0565	0.0472	0.0508
0.01	0.0044	0.0114	0.0155	0.0114	0.0093	0.0107	0.0106

Table 3: $\mathbf{Pr}(Q_2(\mu_0, \mu_1) \geq C_\varepsilon)$

ε, N	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
0.05	0.0354	0.0450	0.0503	0.0476	0.0493	0.0498	0.0499
0.01	0.0066	0.0069	0.0095	0.0093	0.0094	0.0101	0.0099

Table 4: $\mathbf{Pr}(Q_3(\mu_0, \mu_1, \mu_2) \geq C_\varepsilon)$

ε, N	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
0.05	0.0306	0.0373	0.0468	0.0491	0.0489	0.0490	0.0494
0.01	0.0099	0.0138	0.0121	0.0111	0.0106	0.0102	0.0101

Table 5: $\mathbf{Pr}(Q_4(\mu_0, \mu_1, \mu_2, \mu_3) \geq C_\varepsilon)$

ε, N	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}
0.05	0.0564	0.0403	0.0621	0.0579	0.0527	0.0515	0.0507
0.01	0.0200	0.0229	0.0178	0.0171	0.0153	0.0134	0.0120

6 Statistical Analysis

Let F be a function with N values. For instance, F may be produced from a hash function H, where the output was restricted to $\log_2 N$ bits. Let $x_1, \ldots, x_n$ be the sequence of inputs and $y_1, \ldots, y_n$ be the sequence of related outputs: $y_i = F(x_i)$. The function is considered good if for any $x_1, \ldots, x_n$ without repetitions the sequence $y_1, \ldots, y_n$ is indistinguishable from a multinomial uniform distribution sample. Let a statistical test with significance level ε be used. For instance, quadratic box-test with $\varepsilon = 0.05$. In fact one should use exact probabilities from Section 5. Assume m experiments, where the output were $y_{i1}, \ldots, y_{in}$, $i = 1, .., m$ and they are produced for different input strings $x_{i1}, \ldots, x_{in}$. That is a binomial scheme, where a success is the uniformity hypothesis rejection for one output string $y_{i1}, \ldots, y_{in}$. The success probability is ε. One counts the number S_m of strings , where the uniformity hypothesis was rejected. Let $q = \frac{S_m}{m}$. Under uniformity condition, $\mathbf{Pr}(\frac{S_m}{m} = q) \leq e^{-2(q-\varepsilon)^2 m}$ by Chernoff's inequality. Therefore, one says: The uniformity hypothesis was rejected with error probability at most $e^{-2(q-\varepsilon)^2 m}$.

Example. Let $\varepsilon = 0.05$ and $q = 0.07$, and $m = 100000$. Then F is rejected with error probability at most 1.81×10^{-35}.

Remark that one can also use the exact value

$$\mathbf{Pr}(S_m = qm) = \binom{m}{qm} \varepsilon^{qm} (1-\varepsilon)^{m-qm}.$$

7 Conclusion

In this paper, we propose a new statistical test, called *Quadratic Box-Test*, of N multinomial probabilities a. For some non-uniform a the power of the test tends to 1 when the number of trials n and N tend to infinity. In other words, our test is consistent for large N and those a. Also we present an efficient algorithm to compute the exact first error probability and calculate it for low and moderate N. Finally, testing discrete functions is discussed.

References

[1] E. Filiol, "A new statistical testing for symmetric ciphers and hash functions". In V. Varadharajan and Y. Mu, editors, International Conference on Information, Communications and Signal Processing, volume 2119 of Lecture Notes in Computer Science, pages 21-35. Springer-Verlag, 2001.

[2] H. Englund, T. Johansson, and M. S. Turan, "A Framework for Chosen IV Statistical Analysis of Stream Ciphers", In INDOCRYPT 2007. See also Tools for Cryptoanalysis 2007.

[3] F.N. David, *Two combinatorial tests whether a sample has come from a given population,* Biometrika, vol. 37(1950), 97–110.

[4] S.W. Golomb, "Shift Register Sequences", Revised Edition, Aegean Park Press, 1982, Chapter 3.

[5] M.G. Kendall, A. Stuart, *The Advanced Theory of Statistics,* vol. 2, Ch. Griffin & Company Limited, London.

[6] V.F. Kolchin, B.A. Sevast'ynov, V.P. Chistyakov, *Random Allocations*, V.H. Winston & Sons, Washington, D.C., 1978.

[7] National Institute of Standards and Technology of the United States; `http://csrc.nist.gov/groups/ST/hash/documents/FR_Notice_Nov07.pdf`

[8] S.S. Wilks, *Mathematical Statistics*, J.Wiley & Sons, N.Y.,1962.