Statistics and Probability Letters 97 (2015) 192–196

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A sequential hypothesis test based on a generalized Azuma inequality☆

Daniël Reijsbergen [a,*], Werner Scheinhardt [b], Pieter-Tjerk de Boer [b]

[a] Laboratory for Foundations of Computer Science, University of Edinburgh, Scotland, United Kingdom

[b] Center for Telematics & Information Technology, University of Twente, Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 17 July 2014
Received in revised form 21 November 2014
Accepted 21 November 2014
Available online 4 December 2014

Keywords:
Inference
Sequential hypothesis test
Azuma–Hoeffding bound
Martingales

ABSTRACT

We present a new power-one sequential hypothesis test based on a bound for the probability that a bounded zero-mean martingale ever crosses a curve of the form $a(n+k)^b$. The proof of the bound is of independent interest.

1. Introduction

Consider a Bernoulli random variable X with unknown parameter p, i.e., $\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$. We wish to determine whether p is larger or smaller than some value p_0 by drawing i.i.d. samples $X_1, X_2, \ldots$ of X in a sequential manner (see e.g. Mukhopadhyay and de Silva, 2009), only stopping when we have sufficient evidence to accept either of the two following alternative hypotheses,

$$H_{+1} : p > p_0, \qquad H_{-1} : p < p_0. \tag{1}$$

Additionally, we have the null hypothesis $H_0 : p = p_0$, which *cannot* be shown to be correct, as its negation $p \neq p_0$ cannot be disproved statistically: no matter how many samples we draw and no matter how much evidence we see for $p = p_0$, there will always be some small ϵ such that we cannot reject the claim that $p = p_0 + \epsilon$.

We like to emphasize that the goal is to set up a test that will in principle *always* draw a '*correct*' conclusion, no matter how small $|p - p_0|$ is. The meaning of 'correct' here is in the usual sense, namely that the probability of a wrong conclusion is provably below some α, where $1 - \alpha$ is the confidence level. Thus, when $|p - p_0|$ is small, we continue the sequential test to the point where it (eventually) becomes clear whether we should accept H_{+1} or H_{-1}, rather than at some point losing interest and simply concluding that $p \approx p_0$.

☆ Part of this work has been done when the first author was at the University of Twente. The work has been supported by the Netherlands Organisation for Scientific Research (NWO), project number 612.064.812, and by the EU project QUANTICOL, 600708. We would also like to thank Jasper Goseling for helpful discussions, and the two anonymous referees for their helpful comments.

* Corresponding author.
E-mail addresses: dreijsbe@inf.ed.ac.uk (D. Reijsbergen), w.r.w.scheinhardt@utwente.nl (W. Scheinhardt), p.t.deboer@utwente.nl (P.-T. de Boer).

http://dx.doi.org/10.1016/j.spl.2014.11.018
0167-7152/

The test statistic we will use here is a random process Z_n given by

$$Z_0 = 0, \qquad Z_n \triangleq \sum_{i=1}^{n} X_i - np_0, \quad n = 1, 2, \ldots,$$

with a drift $p - p_0$ being zero, positive, or negative, depending on which hypothesis holds true. Clearly, when the experiment shows, e.g., Z_n drifting away to $+\infty$, then this is strong evidence for H_{+1}. The idea is then to fix some positive increasing *threshold function* f_n that allows us to accept H_{+1} as soon as $Z_n \geq f_n$, or to accept H_{-1} as soon as $Z_n \leq -f_n$. We continue sampling (increasing n by drawing new X_i) as long as $-f_n < Z_n < f_n$. Of course one could also choose different shapes for the upper- and lower thresholds, but a symmetric choice seems the most natural, considering the zero drift of Z_n under H_0.

Whatever function f_n is chosen, we will need to bound the probabilities of crossing them, given that some hypothesis is (or is not) valid. Let A_i, $i \in \{-1, +1\}$, be the event that we reject H_0 in favor of H_i. Then we impose the following conditions on the two errors of the first type (in (2)) and on the two errors of the second type (in (3)):

$$\mathbb{P}(A_{+1} \mid \neg H_{+1}) \leq \alpha_1, \qquad \mathbb{P}(A_{-1} \mid \neg H_{-1}) \leq \alpha_2 \tag{2}$$

$$\mathbb{P}(\neg A_{+1} \mid H_{+1}) \leq \beta_1, \qquad \mathbb{P}(\neg A_{-1} \mid H_{-1}) \leq \beta_2. \tag{3}$$

We will consider these probabilities in more detail later; see the proof of Corollary 2. Notice that using (2) and (3) the probability of a wrong conclusion under H_0 is bounded as $\mathbb{P}(A_{-1} \cup A_{+1} \mid H_0) \leq \alpha_1 + \alpha_2$. We do not impose a stricter bound on this, assuming that in reality we never have $p = p_0$ exactly.[1]

The shape of f_n should be sublinear to ensure that the process will hit one of the thresholds with probability 1 (unless $p = p_0$). If this is not the case, (2) may hold, but (3) will not hold for small $p - p_0$; in other words, a conclusion may never be reached, especially when p is close to p_0. When f_n behaves like a square root, we know from the literature on Wiener processes (Shepp, 1967) that even under H_0 one of the thresholds will be crossed after finite time with probability 1; as a consequence, if p is not equal but very close to p_0, both outcomes are almost equally likely, so no bounds on the error probability can be guaranteed.

Striking a good balance between these two extremes is often done by letting f_n asymptotically behave as $\sqrt{n \log(n)}$, see Darling and Robbins (1967) and Darling and Robbins (1968). We will refer to this type of scheme as the *Darling–Robbins scheme*, after the authors of these papers. In the current paper we will pursue another choice, which has not received much attention to the best of our knowledge, namely taking f_n as a power of n, between $\frac{2}{3}$ and 1. It may seem at first sight that this adds 'more than necessary' to the square root, as n^b grows faster than $\sqrt{n \log(n)}$ for any $b > \frac{2}{3}$, but because the derived bound for the n^b case is sharper, it turns out that the new test may require fewer samples depending on its parametrization; see Section 4.

To prove that our choice for f_n indeed satisfies (2) and (3), we need to bound the probability under H_0 that Z_N is above f_N (or below $-f_N$), where N is a random value—namely the smallest n for which $Z_n \notin (-f_n, f_n)$. For non-sequential tests, this only needs to be done for some fixed N, namely the sample size, and the probabilities involved are tail probabilities, to which we can apply, e.g., the Central Limit Theorem or the Chernoff–Hoeffding bound. Another, related bound is the Azuma–Hoeffding inequality, which gives a bound on tail probabilities of *martingales with bounded differences*.[2] Since our process Z_n is a martingale with bounded differences, the Azuma–Hoeffding inequality could be used to establish a fixed sample size test, even if the Chernoff–Hoeffding bound may be tighter. This possibility was already mentioned in Krafft and Schmitz (1969), and was recently investigated in Sason (2011a,b).

Actually, the *same* bound as the one for the Azuma–Hoeffding inequality can also be used for the probability that Z_n is above f_n (or below $-f_n$) for *some* $n \geq N$, rather than just Z_N being above f_n (or below $-f_n$) for fixed N. This was shown by Ross in Section 6.5 of Ross (1996) for the case of linear f_n, where he calls this the *generalized Azuma inequality*. Inspired by this, we apply the same reasoning as Ross (1996) to thresholds of the form n^b, again using (super)martingales. In fact, the proof depends on a lemma that is of independent interest as it may be useful for proving generalized Azuma inequalities for different f_n. The result of the bound is a hypothesis test that can be more powerful than the Darling–Robbins scheme, when the parameters are chosen appropriately.

The outline of the paper is as follows. In Section 2 we give the main result and the application to our hypothesis testing scheme. The proof of the main result is based on two lemmas that we present and prove in Section 3. In Section 4 we provide insight into how the parameters of the test should be chosen, and compare the new test empirically to the Darling–Robbins scheme.

2. Main result

The main result of the paper can be stated as follows.

[1] If $p = p_0$ is possible, one can choose α_1 and α_2 smaller, at the expense of (many) more samples needed. E.g., for a 95% confidence test one can choose $\alpha_1 = \alpha_2 = 0.025$, rather than $\alpha_1 = \alpha_2 = 0.05$.

[2] A supermartingale is a stochastic process Y_n for which $\mathbb{E}(Y_n | Y_{n-1}, \ldots, n_0) \leq Y_{n-1}, n > 0$; when this holds with equality, Y_n is called a martingale; when $\mathbb{P}(|Y_n - Y_{n-1}| < c) = 1$ for some finite c and all $n > 0$, the (super)martingale is said to have bounded differences.

Theorem 1 (*A Generalized Azuma Inequality*). *Let* $Z_n = \sum_{i=1}^{n} X_i - np_0$, *where the* X_i, $i \in \mathbb{N}$, *are i.i.d. random variables with support* $\subset [0, 1]$ *and* $\mathbb{E}(X_i) = p_0$. *Let* $f_n = a(n+k)^b$ *with* $k > 0, a > 0$ *and* $b \in (\frac{2}{3}, 1)$. *Also, let* $N = \min\{n > 0 : Z_n \notin (-f_n, f_n)\}$, *or* $N = \infty$ *if* $Z_n \in (-f_n, f_n)$ *for all* $n > 0$. *Then*

$$\mathbb{P}(N < \infty \text{ and } Z_N \geq f_N) \leq \exp(-8(3b-2)a^2k^{2b-1}), \tag{4}$$

$$\mathbb{P}(N < \infty \text{ and } Z_N \leq -f_N) \leq \exp(-8(3b-2)a^2k^{2b-1}). \tag{5}$$

Proof. Statement (4) follows immediately by applying Lemma 3 in Section 3 to the $\tilde{Z}_n$ process of Lemma 4. Statement (5) follows by applying (4) to the process $-Z_n = \sum_{i=1}^{n}(1 - X_i) - n(1 - p_0)$, which has the same structure as Z_n. □

Corollary 2. *The sequential test as described in the Introduction, with* $f_n = a(n+k)^b$, *satisfies* (2) *and* (3) *with* $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = \alpha$ *if we choose the parameters a, b, and k such that* $8(3b-2)a^2k^{2b-1} \geq -\log(\alpha)$.

Proof. For any $p \leq p_0$, the first probability in (2) is upper-bounded by $\mathbb{P}(A_{+1} \mid H_0)$, which is exactly the probability in (4); similarly the second probability in (2) is upper-bounded by (5). Next define A_0 as $\neg A_{-1} \cap \neg A_{+1}$, the event that we do not reject H_0, i.e., that $-f_n < Z_n < f_n$ for all $n > 0$. This event has probability zero, unless $p = p_0$. Thus, we can write the first probability in (3) as $\mathbb{P}(A_0 \mid H_{+1}) + \mathbb{P}(A_{-1} \mid H_{+1}) = \mathbb{P}(A_{-1} \mid H_{+1})$, and then upper-bound this, for any value of $p > p_0$, by $\mathbb{P}(A_{-1} \mid H_0)$, which is (5); similarly the second probability in (3) is upper-bounded by (4). □

First we present and prove the two lemmas that lie at the foundation of the proof of Theorem 1. Note that the first lemma in general holds for any increasing shape of f_n; this suggests it may be used to generalize Theorem 1 to other choices for f_n.

3. Two basic lemmas

In Section 6.5 of Ross (1996), the proof of the original 'generalized Azuma inequality' is based on a result that determines an upper bound for the probability that a martingale with bounded differences ever crosses some line $a(n + k)$, with $a, k > 0$; see Proposition 6.5.1 of Ross (1996). In this section we provide an analogous result for crossing a line of the form $f_n = a(n + k)^b$, with $a, k > 0$ and $b \in (\frac{2}{3}, 1)$. The proof is similar, based on the supermartingale property of $W_n = e^{c_n(Z_n - f_n)}$ for some sequence c_n (rather than some constant c as in Ross, 1996). One complication is that we need a lower bound on Z_n, which we achieve by letting Z_n stop as soon as it crosses $-f_n$; this does not affect the usefulness of the result.

Lemma 3. *Let* Y_n *be a zero-mean martingale with bounded differences. Also, let* $f_n > 0$ *be any positive increasing threshold function. If we can find a positive sequence* c_n *such that* $W_n = e^{c_n(Y_n - f_n)}$ *is a supermartingale, then*

$$\mathbb{P}(\exists n \geq 0 : Y_n \geq f_n) \leq \exp(-f_0c_0). \tag{6}$$

Proof. Assuming that W_n is a supermartingale, we define the bounded stopping time $N(m) = \min\{n : Y_n \geq f_n \text{ or } n = m\}$ to find that

$$\mathbb{P}(Y_{N(m)} \geq f_{N(m)}) = \mathbb{P}(W_{N(m)} \geq 1) \leq \mathbb{E}(W_{N(m)}) \leq \mathbb{E}(W_0) = \exp(-f_0c_0),$$

where the first inequality is the Markov inequality, and the second is due to the Supermartingale Stopping Theorem (Theorem 6.4.1 of Ross, 1996) and the fact that $N(m)$ is bounded. The result then follows by letting $m \to \infty$. □

Lemma 4. *Let* Z_n *and* f_n *be as in Theorem 1, and let* $\tilde{Z}_n$ *be the same process as* Z_n *but stopped*[3] *at* $-f_n$ *and* f_n. *Also let* $c_n = 8(3b-2)f_n/(n+k) = 8a(3b-2)(n+k)^{b-1}$. *Then the process* $W_n \triangleq e^{c_n(\tilde{Z}_n - f_n)}$ *is a supermartingale.*

Proof. The process W_n is a supermartingale if and only if $\mathbb{E}(W_n|\mathcal{F}_{n-1}) \leq W_{n-1}$ with $\mathcal{F}_n \triangleq W_n, \ldots, W_1$. Clearly this property holds for all $n > N$ since then $W_n = W_{n-1}$. In the sequel we assume $n \leq N$ and write Z_n instead of $\tilde{Z}_n$, keeping in mind that $-f_{n-1} < Z_{n-1} < f_{n-1}$. First we write, for $n \leq N$,

$$\mathbb{E}(W_n|\mathcal{F}_{n-1}) = e^{c_n(f_{n-1} - f_n)} \cdot e^{c_n(Z_{n-1} - f_{n-1})} \cdot \mathbb{E}\left(e^{c_n(X_n - p_0)}|\mathcal{F}_{n-1}\right).$$

The middle factor is precisely $W_{n-1}^{\frac{c_n}{c_{n-1}}}$, while the expectation in the third factor is actually independent of $\mathcal{F}_{n-1}$. So for W_n to be a supermartingale we must have

$$e^{c_n(f_{n-1} - f_n)} \cdot \mathbb{E}(e^{c_n(X_n - p_0)}) \leq W_{n-1}^{1 - \frac{c_n}{c_{n-1}}}. \tag{7}$$

[3] I.e., if $\exists n$ such that $Z_n \notin (-f_n, f_n)$ then, with N being the smallest such n, let $\tilde{Z}_n = Z_n$ for $n < N$ and let $\tilde{Z}_n = -f_N$ or $\tilde{Z}_n = f_N$ for $n \geq N$, depending on whether $Z_N \leq -f_N$ or $Z_N \geq f_N$.

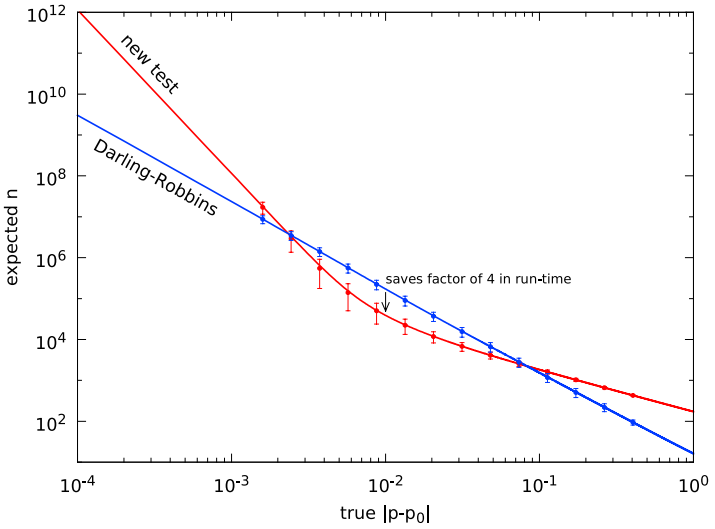

Fig. 1. Comparison between the new test and Darling–Robbins in terms of their expected sample size as a function of the true value $|p - p_0|$. The test parameters were optimized for a guess $\gamma = 0.01$. The expected sample sizes are determined via two different approaches. Solid lines represent rough, numerical approximations, obtained as the intersection of $n \cdot |p - p_0|$ and f_n. On the other hand, dots represent results from computer simulation runs, with ± 1 standard deviation intervals around them.

By Lemmas 6.3.1 and 6.3.2 of Ross (1996) we have that

$$\mathbb{E}(e^{c_n(X_n - p_0)}) \leq (1 - p_0)e^{-p_0 c_n} + p_0 e^{(1-p_0)c_n} \leq e^{c_n^2/8},$$

so that, taking logarithms, (7) is implied by

$$c_n(f_n - f_{n-1}) \geq \frac{1}{8}c_n^2 + (c_{n-1} - c_n)(f_{n-1} - Z_{n-1}).$$

Using $Z_{n-1} > -f_{n-1}$, which is due to $n \leq N$ and the stopping assumption, and then dividing by $c_n f_n$, it remains to show that

$$1 + \frac{f_{n-1}}{f_n} \geq \frac{1}{8}\frac{c_n}{f_n} + 2\frac{c_{n-1}}{c_n}\frac{f_{n-1}}{f_n}. \tag{8}$$

Our particular choice of c_n, which has not been used in the proof so far, stems from the fact that by this choice (8) can be rewritten to $1 + z^b \geq (3b-2)(1-z) + 2z^{2b-1}$, where we write $z \triangleq \frac{n+k-1}{n+k}$. Define $g_b(z) \triangleq 2z^{2b-1} - z^b - 1 + (3b-2)(1-z)$, then we need $g_b(z) \leq 0$ for $z \in (0, 1)$ which, since $g_b(1) = 0$, is implied by $g_b'(z) \leq 0$ on $(0, 1)$. The latter is indeed the case, since

$$g_b'(z) = (4b - 2)(z^{b-1} - 1)\left(z^{b-1} - \frac{2 - 3b}{4b - 2}\right)$$

and clearly $\frac{2-3b}{4b-2} < 0$ when $b \in (\frac{2}{3}, 1)$. This proves (8), hence the result follows. □

4. Application and performance

In order to use a test based on Corollary 2, values need to be given to a, k, b and α. As we argue in Reijsbergen et al. (in press), these parameters have a large impact on the sample size, but there is no universally best choice. The optimal choice depends on $|p - p_0|$, which is unknown *a priori*. However, if the investigator can come up with a *guess* γ for $|p - p_0|$, then the optimal values for a and k can be approximated by numerically minimizing the value n at which $n \cdot |p - p_0|$ hits the boundary f_n. Such a guess can be based on e.g. exploratory simulation runs or strong similarity with another, already solved problem. This is similar to Bayesian statistics, although we emphasize that only the expected sample size depends on γ, but not the error probability guarantees.

As we argue in Reijsbergen et al. (in press), $b = \frac{3}{4}$ strikes a good balance between sensitivity to γ and good performance when γ is (almost) correct. Given this choice (and α), we can compare the new test with the power-one test based on Darling–Robbins with $f_n = \sqrt{\bar{a}(n+1)\log(n+\bar{k})}$, where $\bar{a}$ and $\bar{k}$ are also optimized based on γ. In general, the new test is better when the guess is almost correct, while the Darling–Robbins test is less sensitive to γ. This is illustrated in Fig. 1. For example, if $p_0 = 0.5$ and the guess is 0.51, then the new test will do better if $p \in [0.5024, 0.5827]$ (or if

$p \in [0.4173, 0.4976]$), otherwise Darling–Robbins does better. Near the center of this interval, the new test is better by a factor of more than 4 in sample size.

As a final comment we mention that the test in this paper can also be applied more generally to situations where the X_i are not Bernoulli distributed. Future work in this direction will include the context of importance sampling, provided that the increments X_i (which then correspond to likelihood ratios) can be bounded.

References

Darling, D.A., Robbins, H., 1967. Iterated logarithm inequalities. Proc. Natl. Acad. Sci. USA 57 (5), 1188–1192.
Darling, D.A., Robbins, H., 1968. Some nonparametric sequential tests with power one. Proc. Natl. Acad. Sci. USA 61 (3), 804–809.
Krafft, O., Schmitz, N., 1969. A note on Hoeffding's inequality. J. Amer. Statist. Assoc. 907–912.
Mukhopadhyay, N., de Silva, B.M., 2009. Sequential Methods and their Applications. CRC Press.
Reijsbergen, D., de Boer, P.T., Scheinhardt, W., Haverkort, B.R., 2014. On hypothesis testing for statistical model checking. Int. J. Softw. Tools Technol. Trans. (STTT) http://dx.doi.org/10.1007/s10009-014-0350-1. in press.
Ross, S.M., 1996. Stochastic Processes. John Wiley & Sons.
Sason, I., 2011a. Moderate deviations analysis of binary hypothesis testing. arXiv:1111.1995.
Sason, I., 2011b. On refined versions of the Azuma–Hoeffding inequality with applications in information theory. arXiv:1111.1977.
Shepp, L.A., 1967. A first passage problem for the Wiener process. Ann. Math. Statist. 38 (6), 1912–1914.