ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On the sample mean after a group sequential trial☆

Ben Berckmoes [a,*], Anna Ivanova [b], Geert Molenberghs [b,c]

[a] Universiteit Antwerpen, Departement Wiskunde-Informatica, Middelheimlaan 1, 2020 Antwerpen, Belgium
[b] KU Leuven,
[c] Universiteit Hasselt,

ARTICLE INFO

Article history:
Received 8 November 2017
Received in revised form 23 March 2018
Accepted 28 March 2018
Available online xxxx

Keywords:
Bias
Confidence interval
Group sequential trial
Likelihood theory
Mean squared error
Sample mean

ABSTRACT

A popular setting in medical statistics is a group sequential trial with independent and identically distributed normal outcomes, in which interim analyses of the sum of the outcomes are performed. Based on a prescribed stopping rule, one decides after each interim analysis whether the trial is stopped or continued. Consequently, the actual length of the study is a random variable. It is reported in the literature that the interim analyses may cause bias if one uses the ordinary sample mean to estimate the location parameter. For a generic stopping rule, which contains many classical stopping rules as a special case, explicit formulas for the expected length of the trial, the bias, and the mean squared error (MSE) are provided. It is deduced that, for a fixed number of interim analyses, the bias and the MSE converge to zero if the first interim analysis is performed not too early. In addition, optimal rates for this convergence are provided. Furthermore, under a regularity condition, asymptotic normality in total variation distance for the sample mean is established. A conclusion for naive confidence intervals based on the sample mean is derived. It is also shown how the developed theory naturally fits in the broader framework of likelihood theory in a group sequential trial setting. A simulation study underpins the theoretical findings.

1. Introduction

Throughout the paper, $X_1, X_2, \ldots$ will be a fixed sequence of independent and identically distributed random variables with law $N(\mu, \sigma^2)$, and $\psi_1, \psi_2, \ldots$ a fixed sequence of Borel measurable maps of $\mathbb{R}$ into $[0, 1]$.

For natural numbers $L \in \mathbb{N}_0$ and $0 < m_1 < m_2 < \cdots < m_L < n$, we consider a random sample size N with the following properties:

(a) N can take the values $m_1, m_2, \ldots, m_L, n$,
(b) $\forall i \in \{1, \ldots, L\} : \{N = m_i\}$ is independent of $X_{m_i+1}, X_{m_i+2}, \ldots$,
(c) $\forall i \in \{1, \ldots, L\} : \mathbb{P}\left[N = m_i \mid X_1, \ldots, X_{m_i}\right] = \psi_{m_i}(K_{m_i}) \prod_{j=1}^{i-1} \left[1 - \psi_{m_j}(K_{m_j})\right]$, where we denote $K_m = \sum_{i=1}^{m} X_i$ and the empty product is 1.

The above setting serves as a paradigm for a group sequential trial of random length N with outcomes $X_1, X_2, \ldots$ At each m_i, an interim analysis of the sum K_{m_i} of the outcomes is performed and, based on the generic stopping rule (c), one decides whether the trial is stopped, i.e. $N = m_i$, or continued, i.e. $N > m_i$.

☆ A supplementary file with data accompanies the paper.

* Corresponding author.
E-mail address: ben.berckmoes@uantwerpen.be (B. Berckmoes).

https://doi.org/10.1016/j.csda.2018.03.016
0167-9473/

Note that the product in (c) is merely the usual decomposition of the conditional probability to reach a certain sample size and the product of conditional probabilities of continuing at smaller sample sizes, given that the trial is ongoing. This is similar to decompositions encountered in longitudinal or time-series transition models, and dropout models in longitudinal studies. It follows the law of total probability.

More precisely, at the ith interim analysis only the values of the full sums $K_{m_1}, \ldots, K_{m_i}$ have been analyzed. Therefore,

$$\begin{aligned}&\mathbb{P}\left[N = m_i \mid X_1, \ldots, X_{m_i}\right]\\ &\quad = \mathbb{P}\left[N = m_i \mid K_{m_1}, \ldots, K_{m_i}\right]\\ &\quad = \mathbb{P}\left[N = m_i, N \neq m_{i-1}, \ldots, N \neq m_1 \mid K_{m_1}, \ldots, K_{m_i}\right],\end{aligned}$$

which, by the law of total probability,

$$\begin{aligned}&= \mathbb{P}\left[N = m_i \mid N \neq m_{i-1}, \ldots, N \neq m_1, K_{m_1}, \ldots, K_{m_i}\right]\\ &\quad \prod_{j=1}^{i-1} \mathbb{P}\left[N \neq m_j \mid N \neq m_{j-1}, \ldots, N \neq m_1, K_{m_1}, \ldots, K_{m_i}\right],\end{aligned}$$

which, because, given that $N \neq m_{j-1}, \ldots, N \neq m_1$, the event $\{N = m_j\}$ only depends on the analysis of the full sum K_{m_j},

$$= \mathbb{P}\left[N = m_i \mid N \neq m_{i-1}, \ldots, N \neq m_1, K_{m_i}\right] \prod_{j=1}^{i-1} \mathbb{P}\left[N \neq m_j \mid N \neq m_{j-1}, \ldots, N \neq m_1, K_{m_j}\right],$$

which is exactly the decomposition in (c).

We wish to highlight that the above model contains very useful settings that are extensively studied in the literature. To illustrate this, we let, for each m,

$$\psi_m(x) = 1_{\{|\cdot| \geq C_m\}}(x) = \begin{cases} 1 & \text{if} \quad |x| \geq C_m \\ 0 & \text{if} \quad |x| < C_m \end{cases},$$

with $C_m \in \mathbb{R}_0^+$ a constant. For these choices of ψ_m, expression (c) is turned into

$$\begin{aligned}&\mathbb{P}[N = m_i \mid X_1, \ldots, X_{m_i}]\\ &\quad = 1_{\{|\cdot| \geq C_{m_i}\}}(K_{m_i}) \prod_{j=1}^{i-1} \left[1 - 1_{\{|\cdot| \geq C_{m_j}\}}(K_{m_j})\right]\\ &\quad = \begin{cases} 1 & \text{if} \quad \left|K_{m_i}\right| \geq C_{m_i} \text{ and } \forall j \in \{1, \ldots, i-1\} : \left|K_{m_j}\right| < C_{m_j} \\ 0 & \text{otherwise} \end{cases}.\end{aligned}$$

So this corresponds to a trial which is stopped either at the first m_i for which $\left|K_{m_i}\right| \geq C_{m_i}$, or at n. If, for a fixed constant $C \in \mathbb{R}_0^+$, $C_m = \sigma C \sqrt{m}$, this setting corresponds to the Pocock boundaries, studied in e.g. Siegmund (1978) and Chang (1989), and, if, for a fixed constant $C \in \mathbb{R}_0^+$, $C_m = C$, this setting corresponds to the O'Brien–Fleming boundaries, studied in e.g. Woodroofe (1992). More generally, taking $\psi_m = 1_{\mathcal{S}_m}$ with $\mathcal{S}_m \subset \mathbb{R}$ a Borel measurable set, leads to the setting studied in e.g. Emerson and Fleming (1990) and Liu and Hall (1999). Finally, taking $\psi_m(x) = \Phi\left(\alpha + \beta m^{-1} x\right)$ with Φ the standard normal cumulative distribution function and α, β real numbers, corresponds to the probabilistic stopping rule setting studied in e.g. Molenberghs et al. (2014).

In this paper, we will study the ordinary sample mean $\widehat{\mu}_N = \frac{1}{N} K_N$. It is reported in the literature that in the above described group sequential trial setting, bias may occur if $\widehat{\mu}_N$ is used to estimate μ (Hughes and Pocock, 1988; Emerson and Fleming, 1990; Liu and Hall, 1999). However, it was shown recently in Molenberghs et al. (2014) that if N only takes the values m and $2m$, and $\psi_m(x)$ takes the form $\Phi\left(\alpha + \beta m^{-1} x\right)$ or $\lim_{\beta \to \infty} \Phi\left(\alpha + \beta m^{-1} x\right) = 1_{\{\cdot \geq 0\}}(x)$, this bias vanishes as m tends to ∞. In this paper, we will establish explicit formulas for the expected length of the trial, the bias, and the mean squared error (MSE) in the general case, described by (a), (b), and (c). We deduce that, for fixed L, if $m_1 \to \infty$ (and hence $\forall i : m_i \to \infty$ and $n \to \infty$), the bias vanishes with rate $1/\sqrt{m_1}$, and the MSE vanishes with rate $1/m_1$. We will show that both rates are optimal. Furthermore, under a regularity condition, we will establish asymptotic normality in total variation distance for the sample mean if, for fixed L and $m_1, \ldots, m_L$, $n \to \infty$. In some cases, this validates the use of naive confidence intervals based on the sample mean if n is large.

The paper is structured as follows. In Section 2, we introduce the normal transform of a finite tuple of bounded Borel measurable maps of $\mathbb{R}$ into $\mathbb{R}$, for which we establish a recursive formula. We use the normal transform in Section 3 to obtain an explicit formula for the joint density of N and K_N. We establish a fundamental result in Section 4, which is used to calculate the expected length of the trial in Section 5, and the bias and the MSE in Section 6. It is shown that, for fixed L, if $m_1 \to \infty$ (and hence $\forall i : m_i \to \infty$ and $n \to \infty$), the bias vanishes with rate $1/\sqrt{m_1}$, and the MSE vanishes with rate $1/m_1$. Both rates are shown to be optimal. In Section 7, under a regularity condition, we establish asymptotic normality in total variation distance for $\widehat{\mu}_N$ if, for fixed L and $m_1, \ldots, m_L$, $n \to \infty$. We also derive a conclusion for naive confidence intervals

Please cite this article in press as: Berckmoes B., et al., On the sample mean after a group sequential trial. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.03.016.