The Statistics of Sequential Testing

Patrick Le

Abstract

Keywords:

1. Introduction

2. Standard Group Sequential Testing

2.1. Setting

Data arrives over the course of K periods, each period containing N observations. At the end of each period k, $k = 1, ..., K$, a z-test is performed on the data which has accumulated thus far (so using kN observations) to determine if the mean is statistically different from zero. The z-statistic is compared to some critical value, and the null is rejected if the critical value is exceeded. The tests are performed sequentially until the null is rejected, or until all K tests have been performed.

If the critical value is the usual 1.96, then the fact that multiple tests are performed on the same data implies that Type I error will be inflated to be more than the desired 5%. The goal of sequential testing techniques is to control for Type I error. While Bonferonni corrections can be used, they are overly conservative (leading to Type I errors which are too small) and reduce the power of the test.

The one way to control for Type I error in sequential testing is to control the critical value in each test. Depending on the application, one might want the critical value to be the same for each test, or vary depending on the timing of the test. In this specific note, we show how to compute the shared critical value that can be used for each test.

This setting can be generalized to include applications like A/B testing.

2.2. The correlation structure of z-statistics

The fundamental insight of sequential testing over Bonferonni corrections is that there is significant correlation between the test at some period j and the test at some later period k. This is because both tests utilize some shared data. I now derive the correlation between the z-statistic at a period j, called z^j with the z-statistic at some other, later, period k, called z^k.

Suppose that the null is that the data, denoted by $X = (x_i)_{i=1,...,KN}$, has a mean of zero, and variance σ^2. Then using the first kN observations, the sample mean is $\bar{x}^k = \frac{\sum_{i=1}^{kN} x_i}{kN}$, which has standard deviation $\frac{\sigma}{\sqrt{kN}}$. The resulting z-statistic is given by

$$z^k = \frac{\bar{x}^k}{\frac{\sigma}{\sqrt{kN}}} = \frac{\sum_{i=1}^{kN} x_i}{kN \frac{\sigma}{\sqrt{kN}}}. \tag{1}$$

Similarly, the z-statistic for the test done at period j is given by

$$z^j = \frac{\bar{x}^j}{\frac{\sigma}{\sqrt{jN}}} = \frac{\sum_{i=1}^{jN} x_i}{jN \frac{\sigma}{\sqrt{jN}}}. \tag{2}$$

It is a standard result that the z-statistics are, under the null hypothesis, distributed asymptotically as a standard Normal (0,1) random variable.

Using the fact that these variables have mean zero, the covariance between z^j and z^k is given by:

$$Cov(z^j, z^k) = E(z^j, z^k) = E\left(\frac{\sum_{i=1}^{kN} x_i}{kN\frac{\sigma}{\sqrt{kN}}}\frac{\sum_{i=1}^{jN} x_i}{jN\frac{\sigma}{\sqrt{jN}}}\right). \tag{3}$$

Consider only the numerator for now, and rewrite the second sum as

$$\begin{aligned} E\left(\sum_{i=1}^{jN} x_i \sum_{i=1}^{kN} x_i\right) &= E\left(\sum_{i=1}^{jN} x_i \left(\sum_{i=1}^{jN} x_i + \sum_{i=jN+1}^{kN} x_i\right)\right) \\ &= E\left(\sum_{i=1}^{jN} x_i \sum_{i=1}^{jN} x_i\right) \\ &= E\left(\sum_{i=1}^{jN} x_i^2\right) \\ &= jN\sigma^2, \end{aligned}$$

where the second and third line use the fact that $E(x_i x_{i'}) = 0$ for any $i \neq i'$ (because the x_i's are i.i.d. mean zero), and the last line uses the fact that variance of x is σ^2.

Plugging the numerator into the expression in (3) gives

$$Cov(z^j, z^k) = E(z^j, z^k) = \frac{kN\sigma^2}{kN\frac{\sigma}{\sqrt{kN}}jN\frac{\sigma}{\sqrt{jN}}} = \sqrt{\frac{j}{k}}.$$

Therefore, $Cov(z^j, z^k) = \sqrt{\frac{j}{k}}$.

More generally, let Σ be a $K \times K$ variance covariance matrix for the z-statistics, then entries in Σ in the are given by

$$\Sigma(j,k) = \sqrt{\frac{\min(j,k)}{\max(j,k)}}, \tag{4}$$

for row j, column k.

2.3. Computation of constant critical Z value

Assume that an overall Type I error of α is desired, then the critical Z value, Z_c, needs to satisfy:

$$P(|z^1| \geq Z_c, \text{ or } |z^2| \geq Z_c, \text{ or ... or } |z^K| \geq Z_c) = \alpha.$$

In other words, the probability that any of the z-statistics from any of the K tests is greater or equal to Z_c is α.

The critical Z value is determined via the following simulation procedure.

- Draw a sample of size M of multivariate normal variables with mean zero and covariance matrix Σ, to obtain a matrix Z of z-statistics, with shape $M \times K$. This is essentially simulating the distribution of the z-statistics for each of the K sequential z-tests.
- For each draw, compute the maximum absolute value, to obtain a matrix Z_{abs} of maximum absolute z-statistics, with shape $M \times 1$
- Compute the $(1 - \alpha)$-percentile from the empirical distribution of Z_{abs}. This will be the critical value Z_c.

By construction, the above procedure yields a critical value Z_c that guarantees that the probability of the z-statistic from any of the z-tests exceeding the critical value Z_c is exactly α.

2.4. Remarks

There are a number of considerations when using this sequential testing procedure.

- It is assumed that the timing of the tests correlates with how much additional data has been collected. For instance, on day j there are jN observations, and on day k there are kN observations. This assumption is made to simplify exposition, and can be adjusted/accounted for in applications, such as A/B testing, where the number of new observations does not necessarily follow such a linear trend (due to variation in traffic, weekly/monthly trends, etc.)

3. Flexible Alpha-Spending Function Group Sequential Testing

The previous section outlines a relatively simple procedure to adjust the critical Z-value to account for repeated tests. The simplicity there derives from the imposition of the same critical Z-value. While such applications are simple to maintain and explain, they can be restrictive. To see this, let's plot the Type I error as a function of the number of tests.

[[INSERT chart here]]

This chart above is a type of *alpha-spending function* - it indicates how much Type I error is being committed each time a test is performed. It starts at zero, and ends at 0.05, the desired Type I error. The shape of this function is determined by the fact that we constrain the critical Z-value to be the same for each test. However, such a constraint doesn't need to apply. In other words, we can allow the critical Z-values to vary over time/tests, as long as we constrain the total Type I error to be the right amount.

The *flexible* alpha-spending function group sequential testing procedure allows for this flexibility by relying on a pre-specified alpha-spending function, and computes the appropriate critical Z-value based off of such a function. One can imagine applications where the experimenter is relatively confident of being able to detect a result early on, and thus would like to "front-load" the amount of Type I error that they want to spend on the early tests. Conversely, they might be skeptical of detecting an effect early on, and would like to "save" the Type I error for later rounds of testing. The flexible approach is able to accommodate all these risk preferences, while still respecting the overall Type I error constraint.

3.1. Setting

In the flexible alpha-spending function group sequential testing approach, we take the alpha spending function as given. An alpha-spending function is a monotonically increasing function $\boldsymbol{\alpha}$ from the interval $[0, 1]$ to the interval $[0, \alpha]$ where α is the desired overall Type I error. The domain of the function, the interval $[0, 1]$ is thought of as the fraction of information that's available at the time of testing. For instance, if the experiment runs for 4 weeks, then the information fraction at the end of the second week would be 0.5. In our application, since we perform z-tests at regular intervals, the information fractions at the time of testing are of the form $(\frac{1}{K}, \frac{2}{K}, \frac{3}{K}, ..., \frac{j}{K}, ..., \frac{k}{K}, ..., \frac{K}{K})$.

For examples of alpha-spending functions, refer to DeMets and Lan[1] (1994).

Given an alpha-spending function, the critical Z-values for each test are computed using the following simulation procedure.

- Draw a sample of size M of multivariate normal variables with mean zero and covariance matrix Σ, to obtain a matrix Z of z-statistics, with shape $M \times K$. This is essentially simulating the distribution of the z-statistics for each of the K sequential z-tests
- Using the simulated z-statistics for the first test, grid-search for critical value Z_1 such that $P(|z^1| \geq Z_1) = \boldsymbol{\alpha}(\frac{1}{K})$.
- Using the simulated z-statistics for the first 2 tests, and Z_1 above, grid-search for critical value Z_2 such that $P(|z^1| \geq Z_1, \text{ or } |z^2 \geq Z_2) = \boldsymbol{\alpha}(\frac{2}{K})$
- Iteratively, using the simulated z-statistics for the first k tests, and the critical values found for the first $k-1$ tests, grid-search for the critical value Z_k such that $P(|z^1| \geq Z_1, \text{ or... } |z^{k-1} \geq Z_{k-1}, \text{ or } |z^k| \geq Z_k) = \boldsymbol{\alpha}(\frac{k}{K})$.

[1] David L. Demets and K.K. Gorden Lan - Interim Analysis: The Alpha Spending Function Approach, Statistics in Medicine, Vol. 13, 1994.