Soccer Players Doping Detection: Compare Three Methods with French Elite Players Data

[Summary Author]: Ding Sen A0083955R

Section 1. Introduction

Soccer player doping has been known to be difficult and controversial to detect. It is a well-known area where statistical methods are involved. To a layman, it may seem easy as with all kinds of medical or physiological detections. However, if we open up the white box, we realize all methods are not as straightforward and error resistant as they seem. Some more, soccer players' physiological markers are different than layman, and vary with individuals a lot. There is no established reference data. Statistical detections in professional soccer players thus offer good research areas for statistician to offer their expertise.

Statistical Issues

Typical procedures measure several biological markers, against a population baseline. This population baseline is not a trivial issue. Because elite soccer players have no comparable population baseline. So, they can only use their own data. Their own physiological data could vary according to age, exposure to altitude, and seasonality (winter or summer), thus it calls for a systematic workable solution to check abnormality against their own measurements.

Section 2: Methodology

The paper introduces and compares three z score-based methods. In soccer players, how to select appropriate method to test for doping?

The methods all use simple Tn against z score statistics. Or t(n-2) which asymptotically is closer to z score where n is large.

Method 0: Check nth data against (n-1) data

Method 0 is a baseline. Simple usual method. The well-known z score method, checking nth data against other (n-1) population baseline, Equation as follows:

$$T_n = \frac{X_n - \bar{X}_{n-1}}{\hat{\sigma}_{n-1}\sqrt{1+\frac{1}{n-1}}},$$

where $\bar{X}_{n-1}$ and $\hat{\sigma}^2_{n-1}$ are the empirical mean and variance of the sample $X_1, \ldots, X_{n-1}$, that is

$$\bar{X}_{n-1} = \frac{1}{n-1}\sum_{k=1}^{n-1} X_k \quad \text{and} \quad \hat{\sigma}^2_{n-1} = \frac{1}{n-2}\sum_{k=1}^{n-1}(X_k - \bar{X}_{n-1})^2.$$

Method 1: Check all possible n suspicious data against rest

Method 1 does not know which one among the n data points are the abnormal data, so they detect maximum z score to check the one abnormality among n data points against the rest. Equations, as follows:

$$T_n = \max_{i\in\{1,\dots,n\}} \left| \frac{X_i - \bar{X}_{n,-i}}{\hat{\sigma}_{n,-i}\sqrt{1+\frac{1}{n-1}}} \right|,$$

where

$$\bar{X}_{n,-i} = \frac{1}{n-1}\sum_{k=1,k\neq i}^{n} X_k \quad \text{and} \quad \hat{\sigma}^2_{n,-i} = \frac{1}{n-2}\sum_{k=1,k\neq i}^{n} (X_k - \bar{X}_{n,-i})^2.$$

Method 2: Consider more than 1 Suspicious Data

Method 2 relaxes the assumption that only 1 data points among n data points are abnormal.

It assumes a series of consecutive observations are abnormal. Equations as follows:

$$T_n = \max_{I\in\mathcal{I}} \left| \frac{\bar{X}_I - \bar{X}_{\bar{I}}}{\hat{\sigma}_{n,I}\sqrt{\frac{1}{|I|}+\frac{1}{n-|I|}}} \right|,$$

where $\mathcal{I}$ is the collection of all possible intervals I included in $\{1,\dots,n\}$ with length $1 \le |I| < n$,

$$\bar{X}_I = \frac{1}{|I|}\sum_{k\in I} X_k, \quad \bar{X}_{\bar{I}} = \frac{1}{n-|I|}\sum_{k\notin I} X_k,$$

- I includes more than 1 data points

Method 3: to consider Seasonality

Method 3 further assumes considers seasonality (winter and summer result in different biological markers results), thus partitioning population into two and considers them separately to draw the maximum abnormality. Equations as follows:

$$T_n = \max\left\{ \max_{i\in\{1,\ldots,n_1\}} \left| \frac{X_i - \bar{X}_{n_1,-i}}{\hat{\sigma}_{X,-i,Y}\sqrt{1+\frac{1}{n_1-1}}} \right|, \max_{j\in\{1,\ldots,n_2\}} \left| \frac{Y_j - \bar{Y}_{n_2,-j}}{\hat{\sigma}_{Y,-j,X}\sqrt{1+\frac{1}{n_2-1}}} \right| \right\},$$

where

$$\bar{X}_{n_1,-i} = \frac{1}{n_1-1}\sum_{k=1,k\neq i}^{n_1} X_k, \quad \bar{Y}_{n_2,-j} = \frac{1}{n_2-1}\sum_{k=1,k\neq j}^{n_2} Y_k,$$

$$\hat{\sigma}^2_{X,-i,Y} = \frac{1}{n-3}\left(\sum_{k=1,k\neq i}^{n_1} (X_k - \bar{X}_{n_1,-i})^2 + \sum_{k=1}^{n_2} (Y_k - \bar{Y}_{n_2})^2 \right),$$

and

$$\hat{\sigma}^2_{Y,-j,X} = \frac{1}{n-3}\left(\sum_{k=1}^{n_1} (X_k - \bar{X}_{n_1})^2 + \sum_{k=1,k\neq j}^{n_2} (Y_k - \bar{Y}_{n_2,-j})^2 \right).$$

<u>Summary of Methods</u>

Method	Issues
Method 0: Baseline	To Check nth among n
Method 1	To check any one among n
Method 2	To check a consecutive series among n
Method 3	To further consider seasonality

Table: A simple Comparison of z score based Methods and ideas

Section 3: Application with Soccer Player Data

Researchers gained access to a database of French elite leagues 1 and 2 soccer players. The data include **five biological markers** from **2577 male soccer players** of French elite leagues 1 and 2, include concentrations of ferritin, serum iron, haemoglobin, erythrocytes, and haematocrit levels, collected every 6 months, in July/Aug and in Jan/March from 2006 to 2012 for a total of **12 collections**. There are less valid data due to transfers between clubs, injuries and addition of new clubs. These three methods turn out to be easy to implement.

Reference:

- G. Sauliere et al. **Z-Score based methods and their application to biological monitoring: an example in professional soccer players.** Biostatistics (2019), 1, pp48-64.