Elem. Math. 56 (2001) 157 – 162
0013-6018/01/040157-6 $ 1.50+0.20/0

Elemente der Mathematik

A test statistic whose derivation is simple and unusual

Osvaldo Marrero

Osvaldo Marrero studied mathematics at the University of Miami, and biometry and statistics at Yale University. After holding various positions in academia and in industry, he is now a Professor at Villanova University. His principal mathematical interests are combinatorics, computing, and statistics. Besides enjoying music, physical exercises, reading, and travel, he likes languages. He is fluent in English, French, and Spanish, and continues to improve his knowledge of Dutch and German.

Teaching students how to derive test statistics is an important part of an introductory course in mathematical statistics. Typically the course material includes both the development and the performance of likelihood-ratio tests. Thus, students learn the theoretical foundation for many of the standard tests. But students should also have the opportunity early on to see that sometimes a problem can lend itself well to the development of a specialized test statistic, tailor-made to a particular situation. One such development is given in this paper. Easy to visualize and understand, the derivation of the test statistic is based on simple geometrical and physical ideas. An example is given. Well received by students, this material is useful in mathematical-statistics courses.

1 Introduction

In the usual undergraduate course in mathematical statistics, students learn about likelihood ratios and other general methods to derive test statistics. Sometimes, however, it is possible to obtain a test statistic in another, specialized way that nicely suits a problem.

Bei vielen statistischen Problemen reicht es nicht aus, das vorliegende Material durch eine Häufigkeitsverteilung zu beschreiben. Vielmehr will man wissen, ob die aufgetretenen Abweichungen zufälliger Natur sind. Im entsprechenden Prüfverfahren geht man deshalb von der Annahme aus, dass die bei den Stichproben festgestellten Unterschiede zufälliger Natur sind (Nullhypothese) oder nicht (Alternativhypothese). Es besteht dann die Aufgabe, über Annahme oder Ablehnung der Nullhypothese zu entscheiden. Im vorliegenden Beitrag wird ein elementares Prüfverfahren vorgestellt, das kleine Stichprobenmengen zulässt, was bei den Standardverfahren zum Teil nicht möglich ist. *jk*

One such derivation is presented in this note; the test statistic is developed from simple geometrical and physical concepts that are familiar to students since their high-school days. The procedure is easy to visualize and, therefore, easy to understand. The underlying mathematics has the trigonometric flavor of Fourier or harmonic analysis. Intended for teachers and students of mathematical statistics, this material is a simplified, expository version of a test proposed by Marrero [2]; only monthly data are considered in the present paper. The application of the test is illustrated with an example. Students's reactions to this material have been very positive.

2 The problem

Biomedical researchers are interested in seasonal variation because of ecological considerations. For, if the incidence of a disease shows seasonal variation, then an environmental factor has to be considered in the etiology of that disease. Statistical tests for seasonality are valuable tools that can help to clarify the etiology of diseases that are poorly understood.

In biomedical seasonality studies both the sample size and the amplitude in the data are often small. When applied to such data, standard statistical procedures do not perform well. Therefore statisticians have developed specialized tests that perform better. One such test is discussed in this paper.

Assumed to come from a multinomial distribution with parameters n and $p_1, \ldots, p_{12}$, the data $N_1, \ldots, N_{12}$ are monthly frequencies over a year. The problem is to examine the data for seasonal variation. The null hypothesis is H_0: $p_1 = \cdots = p_{12} = 1/12$, and the alternative hypothesis H_A is that H_0 is false. Thus, the alternative hypothesis is not restricted to a particular kind of seasonal variation. Of course, in practice, the researcher specifies a priori one type of seasonal variation for H_A. As shown below, the test can be adapted to the pattern specified for H_A.

3 The derivation of the test statistic

To develop the test statistic, one takes advantage of the natural cyclic order of monthly data: the months always occur in the same ordering, and January returns after December. Thus, consider a unit circle centered at the origin of a rectangular coordinate system. The circle's circumference is divided into arcs of equal length, and, as explained below, the number of such arcs is determined by the pattern specified for H_A. Each month is identified with one of the arcs, allowing for the possibility that the same arc may correspond to more than one month. The data $N_1, \ldots, N_{12}$ are considered as point-masses, and each point-mass is placed on the center of the corresponding arc. It is as if the monthly data were wrapped around the circle. Thus, one obtains the *sample weighted circle*. It is helpful to visualize this circle as if it were suspended from its center. At the heart of the test there is a simple idea: if H_0 is true, then the center of mass of the weighted circle must be at the origin, and the circle rests at equilibrium. Therefore, the evidence in favor of H_A becomes stronger as the distance between the origin and the center of mass of the sample weighted circle increases.

More precisely, the sample weighted circle is constructed as follows. The test is adapted to the variation specified by the alternative hypothesis by means of an index t. Usually

one chooses $t := 1, 2$; more about t comes later. For each month, $i := 1, \ldots, 12$, one defines $\theta_i := t\pi i/6$; this is the midpoint, in radians, of the arc from $\theta_i - t\pi/12$ to $\theta_i + t\pi/12$ that corresponds to the ith month. Next, for each $i := 1, \ldots, 12$, the point-mass N_i is placed on θ_i, the arc's midpoint. The value of t determines the cyclic frequency $t/12$ and the corresponding period $12/t$ of this placing of point-masses. If the population point-masses were used instead, then one obtains the *population weighted circle*.

For monthly data, the seasonal variations usually seen in biomedical research are annual sinusoidal, semiannual sinusoidal, and annual unimodal.

If the pattern specified for H_A is annual sinusoidal or annual unimodal, then one chooses $t := 1$. This means that the circumference of the weighted circle is divided into twelve arcs such that $\pi/12$ to $\pi/4$ (15° to 45°) corresponds to January, $\pi/4$ to $5\pi/12$ (45° to 75°) corresponds to February, etc. Then, if H_A is true, the center of mass of the weighted circle will not be at the origin. Moreover, in this case, the circle will tilt in the direction of the heaviest concentration of mass, and thereby one can infer the time during the year when the population has maximum frequency. Of course, when H_A specifies annual sinusoidal variation, one would also infer that the population has minimum frequency six months away from the time of maximum frequency.

If the pattern specified for H_A is semiannual sinusoidal, then one chooses $t := 2$. In this case the circumference of the weighted circle is divided into six arcs such that $\pi/6$ to $\pi/2$ (30° to 90°) corresponds to January and July, $\pi/2$ to $5\pi/6$ (90° to 150°) corresponds to February and August, etc. Under H_A, the center of mass of the weighted circle will be away from the origin, and the circle will slant in the direction of the excess of mass. This allows one to infer the location of the two peaks and the two troughs during the year.

In rectangular coordinates, the center of mass of the sample weighted circle is at the point $(\bar{X}, \bar{Y})$, where

$$\bar{X} := \left(\sum_{i=1}^{12} N_i \cos\theta_i\right) \Big/ \left(\sum_{i=1}^{12} N_i\right) = \frac{1}{n}\sum_{i=1}^{12} N_i \cos\theta_i$$

and

$$\bar{Y} := \left(\sum_{i=1}^{12} N_i \sin\theta_i\right) \Big/ \left(\sum_{i=1}^{12} N_i\right) = \frac{1}{n}\sum_{i=1}^{12} N_i \sin\theta_i.$$

The random variables $\bar{X}$ and $\bar{Y}$ are, respectively, unbiased estimators of $\tau_{\bar{X}}$ and $\tau_{\bar{Y}}$, the true values that make up the population center of mass.

The squared distance between the center of mass of the sample weighted circle and its expected value is given by $\{\bar{X} - E(\bar{X})\}^2 + \{\bar{Y} - E(\bar{Y})\}^2$; the test statistic is a standardized version of this random variable. Explicitly, the test statistic T_t is defined by

$$T_t := \frac{\{\bar{X} - E(\bar{X})\}^2}{\mathrm{var}(\bar{X})} + \frac{\{\bar{Y} - E(\bar{Y})\}^2}{\mathrm{var}(\bar{Y})}.$$

Under the null hypothesis, the means, variances, and covariance of $\bar{X}$ and $\bar{Y}$ are equal to simple expressions. These results depend upon the following orthogonality properties of the cosine and sine functions. If t is an integer such that $0 < t < 6$, then

$$\sum_{i=1}^{12} \cos(t\pi i/6) = \sum_{i=1}^{12} \sin(t\pi i/6) = \sum_{i=1}^{12} \cos(t\pi i/6)\sin(t\pi i/6) = 0$$

and

$$\sum_{i=1}^{12} \cos^2(t\pi i/6) = \sum_{i=1}^{12} \sin^2(t\pi i/6) = 6.$$

Thus, if H_0 is true and $t \in \{1, \ldots, 5\}$, one can show that $E_0(\bar{X}) = E_0(\bar{Y}) = 0$, $\mathrm{var}_0(\bar{X}) = \mathrm{var}_0(\bar{Y}) = 1/(2n)$, and $\mathrm{cov}_0(\bar{X}, \bar{Y}) = 0$. Moreover, when H_0 is true, it follows that the expression for the test statistic simplifies to

$$T_t = \frac{2}{n}\left\{ \left(\sum_{i=1}^{12} N_i \cos \frac{t\pi i}{6} \right)^2 + \left(\sum_{i=1}^{12} N_i \sin \frac{t\pi i}{6} \right)^2 \right\}.$$

If not already, the geometric meaning of the test statistic becomes transparent in terms of polar coordinates. Let (R, Θ) be the polar coordinates that correspond to $(\bar{X}, \bar{Y})$. Thus, the random variable R is defined by $R := (\bar{X}^2 + \bar{Y}^2)^{\frac{1}{2}}$, and, assuming $(\bar{X}, \bar{Y}) \neq (0, 0)$, the random variable Θ is defined by

$$\Theta := \begin{cases} \mathrm{Arctan}(\bar{Y}/\bar{X}), & \text{if } \bar{X} > 0, \\ \mathrm{Arctan}(\bar{Y}/\bar{X}) - \pi, & \text{if } \bar{X} < 0 \text{ and } \bar{Y} < 0, \\ \mathrm{Arctan}(\bar{Y}/\bar{X}) + \pi, & \text{if } \bar{X} < 0 \text{ and } \bar{Y} \geq 0, \\ -\pi/2, & \text{if } \bar{X} = 0 \text{ and } \bar{Y} < 0, \\ \pi/2, & \text{if } \bar{X} = 0 \text{ and } \bar{Y} > 0. \end{cases}$$

Then, if H_0 is true, $T_t = 2nR^2$; that is, T_t is equal to $2n$ times the squared distance between the center of mass of the sample weighted circle and the origin. When H_0 is rejected, the value of Θ serves to infer the times during the year when the population has extreme frequencies; this is illustrated in the example below.

To obtain the null distribution of T_t, one assumes that the probability law of the weighted averages $\bar{X}$ and $\bar{Y}$ can be approximated well by a nonsingular bivariate normal distribution. In this case, in particular, $\mathrm{cov}_0(\bar{X}, \bar{Y}) = 0$ is equivalent to the independence of the random variables $\bar{X}$ and $\bar{Y}$. Then the null distribution of T_t is the chi-square with two degrees of freedom. By referring the value of T_t to the chi-square distribution with two degrees of freedom, one can compute the p-value associated with an application of the test. The *p-value* is the probability of the test statistic taking on a value equal to or more extreme than the observed value when H_0 is true; "more extreme" means a result in a direction that favors the alternative hypothesis. Thus, the p-value is a numerical summary of the evidence: the smaller the p-value, the stronger is the sample evidence

against H_0 and in favor of H_A. Customarily set at 5%, the *significance level* defines a priori the upper bound in the following commonly accepted rule: reject H_0 if and only if the p-value is less than or equal to the significance level. A *type I error* occurs when one erroneously rejects H_0. Thus, the significance level that one chooses may be viewed as some sort of insurance against making a type I error during the application of a statistical test.

The normality assumption on $\bar{X}$ and $\bar{Y}$ is a natural one. Roughly, one would think so because $\bar{X}$ and $\bar{Y}$ are weighted averages, and it is widely known that standardized averages of independent and identically distributed random variables follow approximately the standard normal distribution. A study by Marrero [2] produced two important conclusions about the performance of the test statistic T_t. First, the validity of the said assumption on $\bar{X}$ and $\bar{Y}$ was confirmed; the study showed that under H_0 and at the usual nominal significance levels of 1% and 5%, the type I error rate is correct for sample sizes as low as fifteen. Second, the study showed that the test statistic T_t performs very well, outperforming competing tests.

4 Example

Anencephaly is the congenital absence of all or most of the brain. Discussed by Marrero [2], the data are the number of first-born anencephalics in Birmingham, England, 1940–1947, pooled into twelve monthly frequencies (Edwards 1961, Table 1, p. 85). The ordered list of observations is $(10, 19, 18, 15, 11, 13, 7, 10, 13, 23, 15, 22)$.

The alternative hypothesis is annual sinusoidal variation; therefore, one chooses $t := 1$. The resulting test-statistic value $T_1 = 6.64$, whose corresponding p-value $p = 0.0362$. Therefore, one rejects the null hypothesis of uniformity in favor of the alternative hypothesis.

In polar coordinates, the center of mass of the sample weighted circle is located at $(R, \Theta) = (0.137, 12.8^\circ)$. Since $12.8^\circ \in (345^\circ, 375^\circ = 15^\circ)$, one concludes that the population appears to have annual sinusoidal variation, with maximum frequency near the end of December and minimum frequency six months away.

These conclusions agree with Figure 1, where the data are shown together with the fitted annual sinusoidal model $n_i = \alpha_0 + \alpha_1 \cos(\pi i/6) + \beta_1 \sin(\pi i/6) + e_i$, whose least-squares parameter estimates (and respective estimated standard errors) are $\widehat{\alpha}_0 = 14.7$ (1.3), $\widehat{\alpha}_1 = 3.9$ (1.8), and $\widehat{\beta}_1 = 0.9$ (1.8).

Based on this analysis, one concludes that the development of anencephaly is not definitively determined at conception, and that the subsequent development of this disorder may be influenced by environmental factors.

Acknowledgement

The author thanks a referee for helpful, constructive comments.

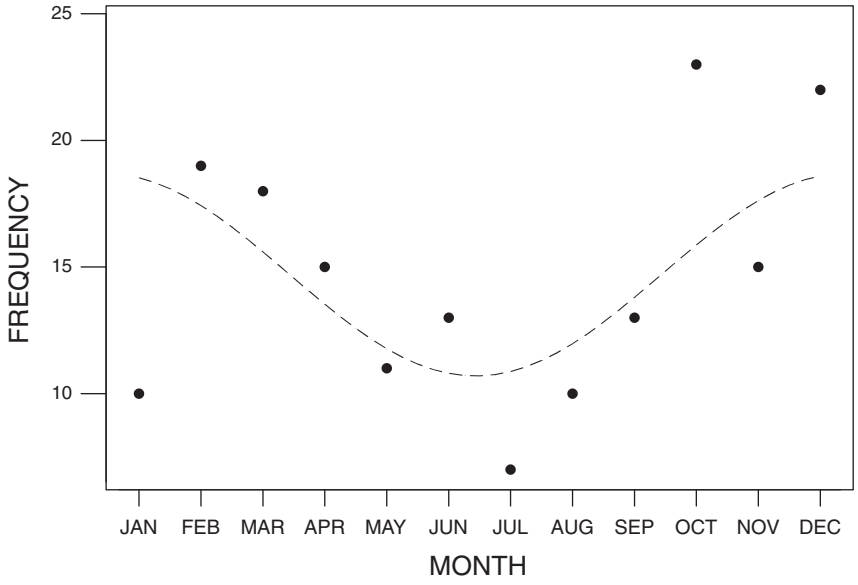

Fig. 1 Monthly frequency of first-born anencephalics in Birmingham, England, 1940–1947: Data and fitted annual sinusoidal model

References

[1] Edwards, J.H.: The Recognition and Estimation of Cyclic Trends, *Annals of Human Genetics* 25 (1961), 83–86. (See also the addendum: Smith, C.A.B.: Note on the Error Variance, *Ibid.*, 86–87.)

[2] Marrero, O.: L'analyse de la variation saisonnière quand l'amplitude et la taille sont faibles, *Revue canadienne de statistique* (*Canadian Journal of Statistics*) 27 (1999), 875–882.

Osvaldo Marrero
Department of Mathematical Sciences
Villanova University
Villanova, Pennsylvania 19085-1699, USA
e-mail: Osvaldo.Marrero@villanova.edu