Lies or Misuse?

Comment on "Lies, Damned Lies, and Statistics (in Geology)"

PAGES 65–66

To demonstrate a concern in geological interpretation after statistical hypothesis testing, writing that "geological hypotheses are never 'true'—they will always be rejected if lots of data are available," P. Vermeesch (*Eos, 90*(47), 443, doi:10.1029/2009EO470004, 2009) considers a null hypothesis H_0 of earthquake occurrences not depending on the day of the week. He found that his testing result rejects H_0, and he argues that the hypothesis testing does not reveal any geological significance. We argue that his conclusion basically demonstrates a Type I statistical error, where the null hypothesis is rejected despite being true.

Because the use of hypothesis testing crucially relies on three criteria—the correct null hypothesis, a plausible probability distribution, and an appropriate testing statistic—one will easily obtain an incorrect interpretation of statistical significance if one of these criteria is not met. Vermeesch's argument does not exhaustively address whether the last two criteria are met and is insufficient to claim that statistically the hypothesis should be rejected.

When the null hypothesis is treated as having a normal, instead of a uniform, probability distribution of occurrence, the chi-square test shows the same as that obtained by Vermeesch, where H_0 is rejected. Yet our analyses on both the raw data and subsets with a tenth of the original size randomly sampled from the raw data show a consistent H_0 rejection. This result disagrees with that of Vermeesch, who argues that sample size influences statistical results, which if true should hold for any distribution. This result also suggests that using uniform distribution for the null hypothesis may not reveal truth.

Furthermore, besides the chi-square test, a two-sample Kolmogorov-Smirnov (KS) test (see F. J. Massey, *J. Am. Stat. Assoc., 46*(253), 68–78, 1951) is also considered to see whether the probability distribution of one data set can be said with confidence to match another, information that helps to show that they describe the same phenomena. The KS test shows that the raw data (or the subsets of Vermeesch's earthquake data) and a normal probability distribution are almost the same, indicating that one cannot reject H_0. Even when a uniform probability distribution is used, the KS test still cannot reject H_0. The hypothesis of earthquake occurrences not depending on the day of the week is thus statistically significantly accepted by the KS test.

We thus suggest that users must pay considerable attention to determine the geological significance of the "uniformity" or "null hypothesis." One must be cautious with the application of hypothesis testing before confidently drawing conclusions after analysis.

—CHIH-YUAN TSENG, Department of Oncology, University of Alberta, Edmonton, Alberta, Canada; and CHIEN-CHIH CHEN, Department of Earth Sciences and Graduate Institute of Geophysics, National Central University, Jhongli, Taiwan; E-mail: chencc@earth.ncu.edu.tw

Statistical Significance Does Not Equal Geological Significance

Reply to Comments on "Lies, Damned Lies, and Statistics (in Geology)"

PAGE 66

In my *Eos* Forum of 24 November 2009 (*90*(47), 443), I used the chi-square test to reject the null hypothesis that earthquakes occur independent of the weekday to make the point that statistical significance should not be confused with geological significance. Of the five comments on my article, only the one by *Sornette and Pisarenko* [2011] disputes this conclusion, while the remaining comments take issue with certain aspects of the geophysical case study. In this reply I will address all of these points, after providing some necessary further background about statistical tests.

Two types of error can result from a hypothesis test. A Type I error occurs when a true null hypothesis is erroneously rejected by chance. A Type II error occurs when a false null hypothesis is erroneously accepted by chance. By definition, the p value is the probability, under the null hypothesis, of obtaining a test statistic at least as extreme as the one observed. In other words, the smaller the p value, the lower the probability that a Type I error has been made. In light of the exceedingly small p value of the earthquake data set, *Tseng and Chen*'s [2011] assertion that a Type I error has been committed is clearly wrong. How about Type II errors?

If β is the probability of a Type II error, then the "power" of a statistical test is given by $1 - \beta$. It is well known that the power of a test increases with sample size n [*Cohen,* 1992]. Given the extremely large sample size of the earthquake data set (n = 118,415), the chance of a Type II error is also vanishingly small. The concept of statistical power lies at the heart of the problem at hand, as acknowledged in the comment by *Taylor and Anderson* [2011]. For example, the Kolmogorov-Smirnov test used by Tseng and Chen fails to reject the null hypothesis because it has very low power in this context.

The outcome of a statistical test depends on three parameters: the significance criterion (α, typically taken as 0.05), the sample size (n), and the so-called "effect size" (w). The key point in the present discussion is that whereas the absolute value of the effect size (for example, the height difference between two people) is scientifically interesting, the question of whether it is "significantly" different from zero is not only irrelevant but also can actually cause harm [*Ziliak and McCloskey,* 2008]. For a multinomial distribution (a.k.a. a histogram), the relationship between α, n, and w is well known and can be calculated analytically (for details, see the online supplement to this *Eos* issue (http://www.agu.org/eos_elec/)). Some key values for the earthquake problem are given in Table 1 and strongly contradict the claim by Sornette and Pisarenko that the p value should not depend on sample size.

The comments by *Kravtsov and Saunders* [2011], *Taylor and Anderson* [2011], and *Weigel* [2011] attribute the nonuniformness of the weekly earthquake distribution to the occurrence of aftershocks. I would like to remark that serial dependence and clustering of observations are not problematic, per se, unless they act on time scales that are shorter than the time span of the histogram, which is indeed the case for the weekly earthquake distribution. Sornette and Pisarenko remove the aftershock clusters by applying a number of progressively more restrictive "filters" to the data. Not surprisingly, they eventually manage to produce a data set that passes the chi-square test (p = 0.46). Not only is this a classical case of circular reasoning, but also the fact that the filtered data set contains only 5636 earthquakes (n/21) is an excellent illustration of the point my paper was trying to make. Using

Table 1. Power Calculation of a Multinomial Distribution (Seven Bins, n = 118,415)[a]

Sample Size	Expected Chi-Squared Value	Expected p Value
n	94	4.5×10^{-18}
n/2	50	4.7×10^{-9}
n/4	28	9.4×10^{-5}
n/8	17	9.3×10^{-3}
n/13	13	0.05
n/20	10	0.11

[a]The expected chi-squared values are based on a noncentral chi-square distribution using a very small effect size of 0.02726 [*Cohen,* 1992]. See the online supplement to this *Eos* issue for details.