Journal of Statistical Software

October 2005, Volume 14, Issue 13. http://www.jstatsoft.org/

Monkeying with the Goodness-of-Fit Test

George Marsaglia
Florida State University

Abstract

The familiar $\sum(\text{OBS} - \text{EXP})^2/\text{EXP}$ goodness-of-fit measure is commonly used to test whether an observed sequence came from the realization of n independent identically distributed (iid) discrete random variables. It can be quite effective for testing for identical distribution, but is not suited for assessing independence, as it pays no attention to the order in which output values are received.

This note reviews a way to adjust or tamper, that is, monkey-with the classical test to make it test for independence as well as identical distribution—in short, to test for both the i's in iid, using monkey tests similar to those in the Diehard Battery of Tests of Randomness (Marsaglia 1995).

Keywords: goodness of fit, χ^2, monkey tests, overlapping m-tuples.

1. Introduction

We begin with an example that will illustrate the problem and suggest a solution. A general discussion will follow in Section 2.

Suppose you want to demonstrate use of the goodness-of-fit test on a sequence $x_1, x_2, \ldots, x_n$ purported to be iid, that is, independent, identically distributed discrete random variables taking values $0, 1, 2, \ldots, 25$ with equal probabilities, $1/26$. You use a computer to generate such a sequence with $n = 2600$, using, say, the random number generator provided by your programming language's function library.

If v_i is the number of times that the value i appears in the sequence of length 2600, then $E[v_i] = 100$ and $Q_1 = \sum(v_i - 100)^2/100$ is the standard goodness-of-fit statistic, viewed as a χ^2_{25} variate (25 degrees of freedom) if the process providing the 26 possible values is choosing them with equal probabilities. But the value of Q_1 is invariant under permutations of the 2600 elements in the test sequence. That Q_1 value provides no assessment for independence, only for identical distribution—the second, but not the first, i in iid.

To consider means for assessing independence, suppose we change the values 0,1,2,...,25 produced by our random variates to A,B,C,...,Z, in order to have a single symbol for each realization. Then the output of a sequence of realizations might look like that from the fabled monkey at a typewriter, randomly striking 26 keys. After n keystrokes, the whimsical output from the monkey might begin, then end, like this:

```
DCJFAVSPPFWMFHFCVYFMLNBQFNFSGCDQFLSZOIVLPTHEQDIIZSGNWJCJRY...
                           ...LUQMAJMLQJHKKOJXOPOYMLFWWKNEDXDOKMNQOLYBJPZZYPP
```

A satisfactory value for $Q_1 = \sum(v_i - n/26)^2/(n/26)$ will suggest that the frequencies of A's,B's,...,Z's are satisfactorily close to their expected values of $n/26$. But we might ask: how many times does CAT appear in that sequence? DOG? PDQ? TIT?—or, more generally, how many times does each of the possible 3-letter words appear, or, a favorite when using this as a classroom example, which forbidden 4-letter words appear, and how often?

If $n = 100 \times 26^3 = 1757600$, and if we have independent realizations, then we should expect around 100 instances of each of the possible 3-letter words, and we might be tempted to use the classical $\sum(\text{OBS} - \text{EXP})^2/\text{EXP}$ as a test statistic. But that will not work, because successive (overlapping) 3-letter words cannot reasonably be assumed to have come from a set of identically distributed, id, variates. It turns out, as we shall see in the next section, that the following is the proper way to test the 3-letter word counts, and thus extend to a full iid test, that is, test for independence as well as identical distribution: Let w_{ijk} be the number of times the word IJK appears in the sequence, and form $Q_3 = \sum_{i,j,k}(w_{ijk} - 100)^2/100$, the naive Pearson form. Also, with w_{ij} the number of times the word IJ appears in the sequence, set $Q_2 = \sum_{i,j}(w_{ij} - 2600)^2/2600$, the form for pairs of letters.

Then, if the monkey is randomly and independently striking the typewriter keys, $Q3 - Q2$ should follow a χ^2 distribution with $26^3 - 26^2 = 16900$ degrees of freedom, (d.o.f.).

For such a large d.o.f. we can take $Q_3 - Q_2$ to be normal with mean 16900 and variance 33800, so the question reduces to: how many sigmas is $Q3 - Q2$ from its mean, i.e., what is the value of the standard normal variate $(Q_3 - Q_2 - 16900)/183.85$?.

For a truly iid sequence of $26^3 \times 100$ uniform variates from a set of 26, we should certainly expect $(Q_3 - Q_2 - 16900)/183.85$ to be within ±3. For the `rand()` function of the `C` compiler I used, (a congruential RNG), $Q_3 - Q_2$ was -1.27 sigmas from its mean, quite satisfactory.

However, we get a different story if we change the `rand()` function to a LFSR generator based on the primitive polynomial $x^{31} + x^3 + 1$, the very generator Whittlesey (1969) touted after I had established the lattice structure of congruential RNGs (Marsaglia 1968). The standard goodness-of-fit measure, which we might call Q_1, passed well; the letters are apparently produced with the proper frequencies. But are they produced independently? No, not even close: $Q_3 - Q_2$ was over 62000 sigmas from its mean, and $Q_2 - Q_1$ was worse, over 136000 sigmas from its mean.

Monkey tests such as these and others from Marsaglia (1995) have been be used for many RNGs. A more extensive summary, and further references relating to monkey tests are in that CDROM, mostly in the file `monkey.pdf`, a copy of which is available from `http://www.jstatsoft.org/v14/i13/monkey.pdf`. See also Knuth (1999, pp. 62, 78, 565) for discussion of the $Q_r - Q_{r-1}$ approach. The main purpose of this note is to merely point out to interested readers that extending the standard $\sum(\text{OBS} - \text{EXP})^2/\text{EXP}$ to more than just individual frequencies, (e.g., to $Q_2 - Q_1$ or $Q_3 - Q_2$), can provide a better assessment of a purported iid sequence. Using overlapping pairs, triples, quadruples and the resulting χ^2 distributions for $Q_2 - Q_1$, $Q_3 - Q_2$ or $Q_4 - Q_3$, can provide means to test for both of the i's in iid.

2. A sketch of background theory

If $x_1, x_2, \ldots, x_n$ is a sequence of iid discrete variates taking values $v_1, v_2, ..., v_k$ with probabilities $p_1, p_2, \ldots, p_k$, and if $V_1, V_2, \ldots, V_k$ are counts for the number of times v_i appears in the sequence of x's, then $V_1, \ldots, V_k$ are, as $n \to \infty$, asymptotically jointly normal with mean vector $(np_1, \ldots, np_k)$ and a certain covariance matrix C. That covariance matrix will have rank $k-1$ because $V_1 + \cdots + V_k = n$. For such a jointly normal distribution, if C^- is any weak inverse of C, $(CC^-C = C)$, then the mean-adjusted quadratic form in the V's, with matrix C^-, will be reduced to the sum of squares of $k-1$ standard normal variates, and thus have a χ^2_{k-1} distribution. This quadratic form can be expressed as $\sum(V_i - np_i)^2/(np_i)$ in the above case, the familiar $\sum(\mathrm{OBS} - \mathrm{EXP})^2/\mathrm{EXP}$.

Now let z_i be the value taken by x_i, and consider the concatenation $z_1 z_2 z_3 \cdots z_n$, as though it were the output of our monkey randomly hitting k different keys with probabilities $p_1, p_2, \ldots, p_k$.

From the assumptions, the probability that any particular 3-letter 'word' in this string will take a specific value, say $v_7 v_4 v_3$, is the product $p_7 p_4 p_3$, and the number of appearances of that particular word will have an (asymptotically) normal distribution, part of the jointly normal distribution for the counts of all 3-letter words. Thus, if we could find the covariance matrix C for the joint 3-letter word counts, and find a weak inverse C^- for C, then we could take a mean-adjusted quadratic form in the joint 3-letter word counts, with matrix C^-, and have a χ^2_r distribution, with r the rank of C.

When I first tried this, I was stuck with an awkward covariance matrix for the joint 3-letter word counts, because the first 3-letter word has no left neighbor, the last has no right neighbor, the second has only one left neighbor, and so on, while the more central 3-letter words have two left- and two-right neighbors that contribute to finding the joint covariances.

I was able to overcome this difficulty by assuming the output was circular, so that each 3-letter word had two left- and two right-neighbors. The resulting covariance matrix had a complicated form for which I was able to find a weak inverse, and then discover a remarkable fact: *The mean-adjusted quadratic form in the weak inverse for that covariance matrix can be represented as the difference, $Q_3 - Q_2$, between two familiar $\sum(OBS - EXP)^2/EXP$ forms, Q_3 for 3-letter and Q_2 for 2-letter words.*

The general result is readily inferred from that for 3-letter words, which we use to simplify notation: If $z_1 z_2 z_3 \cdots z_n$ is the (circular) concatenated output of a sequence $x_1, x_2, \ldots, x_n$ of iid discrete variates taking values $v_1, v_2, \ldots, v_k$ with probabilities $p_1, p_2, \ldots, p_k$, and if Q_3 is the naive Pearson form for 3-letter words:

$$Q_3 = \sum_{i.e.} \frac{(V_{ijk} - np_i p_j p_k)^2}{np_i p_j p_k},$$

where V_{ijk} is the count for the number of appearances of the 3-letter word $z_i z_j z_k$, and if Q_2 is the naive Pearson form for 2-letter words,

$$Q_2 = \sum_{i,j} \frac{(V_{ij} - np_i p_j)^2}{np_i p_j},$$

then $Q_3 - Q_2$ is (asymptotically) χ^2 distributed with $k^3 - k^2$ degrees of freedom, and more generally, for example, $Q_5 - Q_4$ will be asymptotically χ^2 distributed with $k^5 - k^4$ degrees of freedom, $Q_4 - Q_3$ will be $\chi^2_{k^4 - k^3}$ distributed, and so on.

In practice, $Q_2-Q_1, Q_3-Q_2, Q_4-Q_3$, etc. are themselves discrete variates whose distributions get closer to the χ^2 limiting forms as n increases. In particular, expected counts such as $np_ip_jp_k$ may require a large value of n in order to reach the lower limit of ten or so that we often require for expected cell counts. Otherwise, the distributions of forms such as Q_3-Q_2, Q_4-Q_3, etc., may not be close enough to the limiting χ^2 distribution, and extensive simulation may be required for more accurate hypothesis testing.

References

Knuth D (1999). *The Art of Computer Programming*, volume II. Addison-Wesley, Reading, Massachusets, third edition.

Marsaglia G (1968). “Random Numbers Fall Mainly in the Planes.” *Proceedings of the National Academy of Sciences*, **61**, 25–28.

Marsaglia G (1995). “The Marsaglia Random Number CDROM, including the Diehard Battery of Tests of Randomness.” Developed at Florida State University under a grant from The National Science Foundation. The original 1000 free CD’s are long gone, but access to a master copy is available at http://stat.fsu.edu/pub/diehard/. An improved version of the Diehard battery is at http://www.cs.hku.hk/~diehard/.

Whittlesey J (1969). “On the Multidimensional Uniformity of Pseudorandom Number Generators.” *Communications of the ACM*, **12**, 247.

Affiliation:

George Marsaglia
Professor Emeritus, Statistics
Florida State University
Mail address: 1616 Golf Terrace Drive
Tallahassee FL 32301, United States of America
E-mail: geo@stat.fsu.edu

Journal of Statistical Software
October 2005, Volume 14, Issue 13.
http://www.jstatsoft.org/

Submitted: 2005-05-01
Accepted: 2005-09-20