COMBINING EVIDENCE FROM SEVERAL SAMPLES FOR TESTING GOODNESS-OF-FIT TO A LOCATION-SCALE FAMILY

by

Donald A. Pierce
Oregon State University
and
Stanford University

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

TECHNICAL REPORT NO. 15

June 5, 1978

STUDY ON STATISTICS AND ENVIRONMENTAL
FACTORS IN HEALTH

PREPARED UNDER SUPPORT TO SIMS FROM

ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION (ERDA)
ROCKEFELLER FOUNDATION
SLOAN FOUNDATION
ENVIRONMENTAL PROTECTION AGENCY (EPA)
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Combining Evidence from Several Samples for Testing Goodness-of-Fit to a Location-Scale Family

Donald A. Pierce[1]
Oregon State University

SUMMARY

Consider the problem of testing goodness-of-fit to a specified location-scale family when evidence is to be combined from several independent samples, from populations with possibly different location and scale parameters. The procedure studied here is that of computing standardized residuals from each sample and then combining these into one set to be treated essentially as though they came from one sample. It is shown that the limiting distribution of any location-scale invariant goodness-of-fit statistic so applied is precisely the same as for the corresponding one-sample problem.

[1] This work was carried out while the author was visiting Stanford University.

1. INTRODUCTION

Consider the problem of testing goodness-of-fit to a specified location-scale family of distributions, such as the normal, when evidence is to be combined from several independent random samples, from populations which may differ in location and scale. That is, the null hypothesis is that the cumulative distribution functions of the k sampled populations are $F(x; \mu_j, \sigma_j) = H\{(x-\mu_j)/\sigma_j\}$, $j = 1,2,\ldots,k$, where $H(\cdot)$ is specified.

Several authors have discussed methods for this situation; see Wilk and Shapiro (1968), Pettitt (1977), and Quesenberry, et al (1976). In the first two of these the suggestion is basically to compute significance levels $P_1, P_2, \ldots, P_k$ from each sample by using some standard goodness-of-fit test, and then to combine these significance levels by Fisher's method. That is, compute $\chi^2 = -2\,\Sigma \log P_j$ which has under the null hypothesis a chi-square distribution with $2k$ degrees of freedom.

The objective here is to discuss a quite different approach, combining the samples <u>before</u> computing a test statistic, which seems more suitable for alternative hypotheses in which the populations have roughly the same shape, differing only in location and scale. This was also the rationale for the method of Quesenberry, et al (1976), but the approach here is much more direct than theirs.

Let x_{ji} denote the ith observation from the jth population; $j = 1,2,\ldots,k$; $i = 1,2,\ldots,n_j$; $\Sigma n_j = n$. Write $(\hat{\mu}_j, \hat{\sigma}_j)$ for the maximum likelihood estimators, computed separately for each of the k samples. Define residuals by $\hat{e}_{ji} = (x_{ji} - \hat{\mu}_j)/\hat{\sigma}_j$. It seems natural to infer a presumed common shape of the k populations by considering the distribution of the entire set of $n = \Sigma n_j$ residuals as one group. Graphical methods are useful for this purpose, and additionally one may wish to carry out a related formal significance test. The basic result to be pointed out here is that the limiting distribution of <u>any</u> location-scale invariant goodness-of-fit statistic computed from the entire set of n residuals is precisely the same as for the one-sample problem.

For testing normality a useful reference for results on the one-sample problem, with estimated parameters, is Pearson and Hartley (1972). Tables are given there for most of the common types of goodness-of-fit tests. For the case of the conventional chi-square test, it is well-known that the limiting distribution is not precisely chi-square; see Dahiya and Gurland (1972) for a table of critical values for testing normality. It should be noted that this test is rather poor in terms of power. See Shapiro, et al (1968) and Stephens (1974) for comparative studies. A modification of the conventional chi-square test which does have a limiting chi-square distribution was given by Rao and Robson (1974); see also Moore (1977). Thomas and Pierce (1978) gave results on the Neyman smooth test, with estimated parameters, which can be used for the cases of normal, exponential, and Weibull (extreme-value) hypotheses. For additional results on the Weibull case see Stephens (1977).

2. LIMITING DISTRIBUTIONS

Consider first the one-sample problem. We assume that a location-scale invariant goodness-of-fit statistic will be a function of the ordered values of the residuals $\hat{e}_i = (x_i - \hat{\mu})/\hat{\sigma}$. Equivalently, such a statistic will be a function of the empirical distribution function $\hat{H}_n$ of $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$; where $n\,\hat{H}_n(e)$ is the number of $\hat{e}_i \leq e$, $i = 1,2,\ldots,n$.

Under the null hypothesis the probability distribution of the stochastic process $\hat{y}_n$, where $\hat{y}_n(e) = \sqrt{n}[\hat{H}_n(e) - H(e)]$, depends on $H(\cdot)$ and n but not on (μ, σ) , since $\hat{y}_n(e)$ is a location-scale invariant statistic. Further $\hat{y}_n$ converges, as $n \to \infty$, to a stochastic process $\hat{y}$ described in Durbin (1973a,b). No details of that process are needed here, however, except that its distribution is the same for all (μ, σ) .

Returning to the k-sample problem, write $\hat{H}^{(j)}_{nj}$, $\hat{y}^{(j)}_{nj}$ for the above processes as defined for each sample, and $\hat{H}^*_n$, $\hat{y}^*_n$ for these processes as defined for the pooled set of n residuals. It is easily seen that

$$\hat{y}^*_n = \sum_{j=1}^{k} (n_j/n)^{\frac{1}{2}}\, \hat{y}^{(j)}_{nj} ,$$

and, as the $n_j \to \infty$ in such a way that $n_j/n \to \alpha_i > 0$ for all j , the process $\hat{y}^*_n$ converges under the null hypothesis to the same process $\hat{y}$ as do each of the $\hat{y}^{(j)}_{nj}$.

The limiting distribution under the null hypothesis of any goodness-of-fit statistic which is a suitably continuous function of $\hat{H}_n^*$ is then determined by the distribution of $\hat{y}$, and is the same as for the corresponding one-sample problem. See Durbin (1973b, Sec. 3.2, 4.4, 4.5) for discussion of the required continuity, which holds for any of the conventional goodness-of-fit statistics.

3. AN EXAMPLE

An example of the increased sensitivity which may be gained by this approach, as opposed to combining separate tests by Fisher's method, some simulated sampling results are given here for the case of testing normality. The alternative taken is the type for which the test is designed, that is all of the samples are from populations with precisely the same shape, specifically a Weibull distribution with $H(e) = 1 - \exp(-e^2)$.

The Anderson-Darling test was used both for the separate samples and for the pooled residuals. Pettitt (1977) gave formulas for significance levels for finite sample size, which were used here. Using these results significance levels $P_1, P_2, \ldots, P_k$ were computed for each sample; these were combined into $\chi^2 = -2\,\Sigma \log P_j$ and an overall significance level $P^{(1)}$ was computed from the chi-square distribution with $2k$ degrees of freedom. On the other hand, a significance level $P^{(2)}$ was computed by applying the Anderson-Darling test to the entire set of pooled residuals, using the asymptotic distribution.

Only a few trials of this experiment are needed to get a clear indication of the relative sensitivities, provided that the comparison is made in a pairwise sense for each trial, that is for each set of k samples. Table 1 gives pairs of significance levels for 15 trials, rounded to two decimal places, for the case of 5 samples each of size 30. In addition, 100 trials were carried out under the null hypothesis and the distributions of both $P^{(1)}$ and $P^{(2)}$ were uniform to within sampling error.

Table 1. Paired significance levels for 5 samples of size 30 from Weibull (2), testing for normality

Fisher's Method: $P^{(1)}$	Pooling Residuals: $P^{(2)}$
.36	.08
.01	.00
.02	.00
.13	.00
.19	.01
.16	.00
.01	.00
.01	.00
.25	.05
.07	.05
.42	.31
.19	.04
.00	.00
.09	.01
.42	.01

REFERENCES

Dahiya, R. C. and Gurland, J. (1972). Pearson chi-square test of fit with random intervals, I. Null case. *Biometrika* *59*, 147-53.

Durbin, J. (1973a). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics* *1*, 279-90.

Durbin, J. (1973b). *Distribution Theory for Tests Based on the Sample Distribution Function*. Philadelphia: Society for Industrial and Applied Mathematics.

Moore, D. S. (1977). Generalized inverses, Wald's method, and the construction of chi-squared tests of fit. *Journal of the American Statistical Association* *72*, 131-37.

Pearson, E. S. and Hartley, H. O. (1972). *Biometrika Tables for Statisticians, Vol. 2*. Cambridge: Cambridge University Press.

Pettitt, A. N. (1977). Testing normality of several independent samples using the Anderson-Darling statistic. *Applied Statistics* *26*, 156-61.

Quesenberry, C. P., Whitaker, T. B., and Dickens, J. W. (1976). On testing normality using several samples: An analysis of peanut aflatoxin data. *Biometrics* *32*, 753-59.

Rao, K. C. and Robson, D. S. (1974). A chi-square statistic for goodness-of-fit tests within the exponential family. *Communications in Statistics* *3*, 1139-53.

Shapiro, S. S., Wilk, M. B., and Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association* *63*, 1343-72.

Stephens, M. A. (1974). EDF statistics for goodness-of-fit and some comparisons. *Journal of the American Statistical Association* *69*, 730-37.

Stephens, M. A. (1977). Goodness-of-fit for the extreme value distribution. *Biometrika* *64*, 583-88.

Thomas, D. R. and Pierce, D. A. (1978). Neyman's smooth goodness-of-fit test when the hypothesis is composite. To appear in the *Journal of the American Statistical Association*.

Wilk, M. B. and Shapiro, S. S. (1968). The joint assessment of normality of several independent samples. *Technometrics* *10*, 825-39.