The Canadian Journal of Statistics
Vol. 13, No. 3, 1985, Pages 253–256
La Revue Canadienne de Statistique

The asymptotic distribution of the correlation coefficient in testing fit to the exponential distribution*

R.A. LOCKHART

Simon Fraser University

Key words and phrases: Goodness-of-fit.
AMS 1980 *subject classifications*: Primary 62G10, 62E20.

ABSTRACT

The asymptotic distribution of certain tests of fit to the exponential distribution is obtained. The tests are based on regression of the order statistics on their expectations under a standard exponential distribution. Asymptotic normality at the rate $(\log n)^{1/2}$ is obtained for a family of statistics including the correlation coefficient.

RÉSUMÉ

L'auteur détermine la distribution asymptotique de certaines statistiques définissant des tests d'ajustement analytique de la loi exponentielle. Il s'agit en l'occurrence de la régression des statistiques d'ordre d'un échantillon par rapport à leur valeur espérée calculée relativement à une loi exponentielle standard. La normalité asymptotique d'ordre $(\log n)^{1/2}$ est établie pour une famille entière de statistiques, y compris le coefficient de corrélation.

Let $X_1 \le X_2 \le \cdots \le X_n$ be an ordered sample from an unknown distribution G. Suppose we wish to test the hypothesis H_0 that $G(x) = F((x - \alpha)/\beta)$, where F is some known standard distribution, α is an unknown location parameter, and β is an unknown scale parameter.

Set $W_i = (X_i - \alpha)/\beta$, and let $m_i = \mathscr{E}(W_i)$ and $\sigma_{ij} = Cov\ (W_i, W_j)$. Under H_0 we have

$$\begin{aligned} \mathscr{E}(X_i) &= \alpha + \beta m_i, \\ Cov\ (X_i, X_j) &= \beta^2 \sigma_{ij}. \end{aligned} \tag{1}$$

A number of tests of H_0 have been based on assessing the fit of the linear model (1).

The most obvious of these is $R^2(X, m)$, the square of the correlation coefficient between X and m. When F is standard normal the resulting test is quite powerful; indeed Leslie (1985) and Fotopolous, Leslie, and Stephens (1984) have shown that $R^2(X, m)$ is asymptotically equivalent to both the Shapiro-Wilk and Shapiro-Francia statistics.

In this note we investigate the asymptotic behaviour of $R^2(X, m)$ when F is the standard exponential distribution. This statistic has been studied by Smith and Bain (1976), who give finite-n Monte Carlo critical points. For a full discussion see Stephens and D'Agostino (to appear). In this case

*This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

$$m_i = \sum_{k=1}^{i} (n - k + 1)^{-1}$$

and

$$\sigma_{ij} = \sum_{k=1}^{i} (n - k + 1)^{-2} \qquad \text{for} \quad i \leq j$$

We establish below the asymptotic normality of the correlation coefficient and related statistics at the rate $(\log n)^{\frac{1}{2}}$. The result contrasts with the situation for the normal distribution and for Type II censoring, where a sum of weighted chi-squareds arises. [See deWet and Venter (1972) for the normal case and Lockhart and Stephens (1985) for the Type II censoring situation.]

THEOREM. *Suppose* $\hat{\alpha}$, $\hat{\beta}$ *are estimates of* α, β *with* $n^{\frac{1}{2}}(\hat{\alpha} - \alpha)$ *and* $n^{\frac{1}{2}}(\hat{\beta} - \beta)$ *bounded in probability. Suppose* $\tilde{\beta}$ *is an estimate of* β *with* $(\log n)^{\frac{1}{2}} (\tilde{\beta} - \beta) \rightarrow 0$ *in probability. If* $T_n = \tilde{\beta}^{-2} \sum_{i=1}^{n} (X_i - \hat{\alpha} - \hat{\beta} m_i)^2$, *then*

$$(4 \log n)^{-\frac{1}{2}}(T_n - \log n) \rightarrow \mathsf{N}(0, 1).$$

Note that $n\{1 - R^2(x, m)\}$ has the form T_n, where $\hat{\beta} = (\sum_{i=1}^{n} m_i x_i - n\bar{X})/(\sum_{i=1}^{n} m_i^2 - n)$, $\hat{\alpha} = \bar{X} - \hat{\beta}$, and $\tilde{\beta}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. The hypotheses of the theorem may be checked by direct calculation of the means and variances of $\bar{X}$, $\sum m_i X_i$, and $n^{-1} \sum_{i=1}^{n} X_i^2$. The identities $\sum_{i=1}^{n} m_i = n$ and $\sum_{i=1}^{n} m_i^2 = 2n - m_n$ are useful in the calculations.

Other natural statistics of the form T_n arise by using efficient estimates such as the bias-corrected maximum-likelihood estimates $\hat{\beta} = (n - 1)^{-1} \sum_{i=2}^{n} (X_i - X_1)$ and $\hat{\alpha} = X_1 - \hat{\beta}/n$. Use of the maximum-likelihood estimator for $\tilde{\beta}$ permits calculation of the cumulants of T_n. This allows fitting of Pearson curves to the distribution of T_n. [See Lockhart and Stephens (1985) for details.] This is important in view of the extraordinary slowness of the convergence in our Theorem. For instance, with $n = e^{16} \simeq 9 \times 10^6$ the normal approximation gives $P(T_n < 0) \simeq 0.023$.

Monte Carlo studies by Lockhart and Stephens (1985) show $R^2(X, m)$ has substantially lower power than EDF statistics such as the Anderson-Darling statistic A^2. Use of the maximum-likelihood estimates improves the power substantially but does not bring the power up to that of A^2. In fact Lockhart and McLaren (1985) have used the result presented here to show that statistics of the form T_n have 0 asymptotic efficiency relative to A^2 against a wide variety of contiguous alternatives.

Lockhart and Stephens (1985) have investigated more powerful statistics which avoid this problem by weighting the terms in T_n, say, as

$$U_n = \tilde{\beta}^{-2} \sum_{i=1}^{n} \sigma_{ii}^{-1} (X_i - \hat{\alpha} - \hat{\beta} m_i)^2.$$

Easy applications of weak-convergence results for the quantile process show U_n is generally asymptotically distributed as a sum of weighted chi-squareds.

The result obtained here has been extended by McLaren (1985) to other distributions with an exponential tail, such as the extreme-value and logistic.

Proof. Without loss of generality take $\alpha = 0$, $\beta = 1$. Let

$$T_n^* = \sum_{i=1}^{n} (X_i - \hat{\alpha} - \hat{\beta} m_i)^2$$
$$= \sum_{i=1}^{n} (X_i - m_i)^2 + n\hat{\alpha}^2 + (\hat{\beta} - 1)^2 \sum_{i=1}^{n} m_i^2$$
$$+ 2n\hat{\alpha}(\hat{\beta} - 1) - 2n\hat{\alpha}(\bar{X} - 1) - 2(\hat{\beta} - 1) \sum_{i=1}^{n} m_i(X_i - m_i).$$

The hypotheses on $\hat{\alpha}$, $\hat{\beta}$ together with calculations of $Var\ (\sum m_i(X_i - m_i))$ show

$$(4 \log n)^{-\frac{1}{2}} \left(T_n^* - \sum_{i=1}^{n} (X_i - m_i)^2 \right) \to 0 \qquad \text{in probability.}$$

We will use the martingale central limit theorem of Hall and Heyde (1981, p. 58) to show that

$$(4 \log n)^{-\frac{1}{2}} \left(\sum_{i=1}^{n} (X_i - m_i)^2 - \log n \right) \to \mathrm{N}(0, 1).$$

The result will follow, since then

$$(4 \log n)^{-\frac{1}{2}} (T_n - T_n^*) = \left((4 \log n)^{-\frac{1}{2}} (T_n^* - \log n) + \frac{(\log n)^{\frac{1}{2}}}{2} \right) \frac{1 - \tilde{\beta}^2}{\tilde{\beta}^2}$$
$$\to 0 \qquad \text{in probability.}$$

Let $V_1, V_2, \ldots$ be independently and identically distributed standard exponentials. Then $x_1, \ldots, x_n$ can be constructed (see Pyke 1965) as

$$\mathrm{X}_i = \sum_{j=n-i+1}^{n} j^{-1} V_j.$$

We have easily

$$\sum_{i=1}^{n} (X_i - m_i)^2 = \sum_{j=1}^{n} \sum_{i=1}^{n} \{\max\,(i, j)\}^{-1} (V_i - 1)(V_j - 1).$$

Let

$$\mathcal{F}_k = \sigma\{V_1, \ldots, V_k),$$
$$\sigma_n^2 = Var \left(\sum_{1}^{n} (X_i - m_i)^2 \right)$$
$$= 4 \left(\sum_{1}^{n} j^{-1} + \sum_{1}^{n} j^{-2} \right),$$

and

$$S_{k,n} = \sigma_n^{-1} \left(\sum_{i=1}^{k} \sum_{j=1}^{k} \{\max(i, j)\}^{-1} (V_i - 1)(V_j - 1) - \sum_{j=1}^{k} j^{-1} \right).$$

It is easily checked that $(S_{k,n}, \mathcal{F}_k, k = 1, 2, \ldots, n)$ is a mean-0 martingale. Since

$$\sum_{j=1}^{n} j^{-1} = \log n + O(1),$$

it suffices to show $S_{n,n} \to \mathrm{N}(0, 1)$. This will follow from the martingale central limit theorem if we show

$$\sum_{1}^{n} (S_{k,n} - S_{k-1,n})^2 \rightarrow 1 \qquad \text{in probability,} \tag{2}$$

$$\mathscr{E}[\max\{(S_{k,n} - S_{k-1,n})^2;\ 1 \le k \le n\}] \text{ is bounded in } n, \tag{3}$$

and

$$\max\{(S_{k,n} - S_{k-1,n})^2;\ 1 \le k \le n\} \rightarrow 0 \qquad \text{in probability.} \tag{4}$$

The conditions (2) and (3) follow from easy moment calculations. Note that

$$\max\{(S_{k,n} - S_{k-1,n})^2;\ 1 \le k \le n\} \le 2\sigma_n^{-2}\Bigg[\max\{k^{-2}[(V_k - 1)^2 - 1];$$

$$1 \le k \le n\} + 4\max\left\{k^{-2}(V_k - 1)^2\left(\sum_{1}^{k-1} (V_j - 1)^2\right);\ 1 \le k \le n\right\}\Bigg]$$

Applying Markov's inequality and the Borel-Cantelli lemma to $k^{-\frac{3}{2}}(V_k - 1)^6$ shows $k^{-\frac{1}{2}}(V_k - 1)^2 \rightarrow 0$ a.s. The law of the iterated logarithm shows $k^{-\frac{3}{2}}\sum_{1}^{k-1}(V_j - 1)^2 \rightarrow$ a.s. The condition (4) follows easily, since $\sigma_n \rightarrow \infty$.

ACKNOWLEDGEMENTS

The author would like to thank Professor M. A. Stephens for suggesting the problem studied here, and a referee for several helpful comments.

REFERENCES

DeWet, T., and Venter, J.H. (1972). Asymptotic distributions of certain test criteria of normality. *South African Statist. J.*, 6, 135–149.

Fotopolous, S.; Leslie, J.R., and Stephens, M.A. (1984). Errors in approximations for expected normal order statistics with an application to goodness-of-fit. Technical Report #342, Stanford University.

Hall, P., and Heyde, C.C. (1981). *Martingale Limit Theory and Its Application*. Academic Press, New York.

Leslie, J.R. (1985). Asymptotic properties and new approximations for both the covariance matrix of normal order statistics and its inverse. Technical Report, Department of Statistics, Birbeck College.

Lockhart, R.A., and McLaren, C.G. (1985). Asymptotic relative efficiency of correlation tests-of-fit. Technical Report, Department of Mathematics and Statistics, Simon Fraser University.

Lockhart, R.A., and Stephens, M.A. (1985). Tests of fit to the exponential distribution based on regression. Technical Report, Department of Mathematics and Statistics, Simon Fraser University.

McLaren, C.G. (1985). *Some Contributions to Goodness-of-Fit*. Ph.D. Thesis, Department of Mathematics and Statistics, Simon Fraser University.

Pyke, R. (1965). Spacings. *J. Roy. Statist. Soc. Ser. B*, 27, 395–449.

Smith, R.N., and Bain, L.J. (1976). Correlation-type goodness-of-fit statistics with censored samples. *Comm. Statist. A—Theory Methods*, 5, 119–132.

Stephens, M.A., and D'Agostino, R.B. (1985). *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

Received 23 March 1985
Revised 22 May 1985
Accepted 11 June 1985

Department of Mathematics and Statistics
Simon Fraser University
Burnaby, British Columbia V5A 1G6