MOMENTS IN STATISTICS: APPROXIMATIONS TO DENSITIES AND GOODNESS-OF-FIT

Michael A. Stephens

TECHNICAL REPORT No. 469
MAY 11, 1993

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

MOMENTS IN STATISTICS: APPROXIMATIONS TO DENSITIES AND GOODNESS-OF-FIT

Michael A. Stephens

TECHNICAL REPORT No. 469
MAY 11, 1993

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Professor Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

[This paper appears also as a chapter (pp. 218-232) in *Proceedings of the Conference on Moments and Signal Processing*, Technical Report No. 459, Department of Statistics, Stanford University (September 21, 1992)]

MOMENTS IN STATISTICS: APPROXIMATIONS TO DENSITIES AND GOODNESS-OF-FIT

Michael A. Stephens

Summary

In this article we discuss ways in which moments are used (a) to approximate distributions, and (b) to test fit to a given distribution.

1 Approximating distributions using moments

Solomon and Stephens (1977) give a number of examples of statistics X for which the first few, or even all, the moments or cumulants may be found, but whose density $f(x)$ and distribution $F(x)$, assumed continuous, are intractable. A good example is the statistic S whose distribution is the weighted sum of independent chi-square variables, each with one degree of freedom, written

$$S = \sum_{i=1}^{k} \lambda_i (u_i)^2 \tag{1}$$

where u_i are i. i. d. $N(0,1)$, and λ_i are known weights. Many quantities in statistics have distributions (often asymptotic distributions) like S; for example, the Pearson X^2 statistic, used in testing fit to a distribution when the distribution tested contains unknown parameters which are estimated by maximising the usual likelihood, rather than the multinomial likelihood, has this distribution with some $\lambda_i \neq 1$. Other goodness-of-fit statistics, of Cramer–von Mises type, based on the empirical distribution function (EDF), also have such asymptotic distributions (see, for example, many examples in Stephens, 1986a).

One of the first examples of S to be tabulated, for $k = 2$, involved errors in target hitting during World War 2: tables for S were produced with some labour by Grad and Solomon (1955) using analytic methods. These have been extended by various authors to higher values of k, but the analysis after $k = 5$ or 6 rapidly

MOMENTS IN STATISTICS: APPROXIMATIONS TO DENSITIES AND GOODNESS-OF-FIT

Michael A. Stephens,
Simon Fraser University, Burnaby, B. C., Canada V5A 1S6

Summary

In this article we discuss ways in which moments are used (a) to approximate distributions, and (b) to test fit to a given distribution.

1 Approximating distributions using moments

Solomon and Stephens (1977) give a number of examples of statistics X for which the first few, or even all, the moments or cumulants may be found, but whose density $f(x)$ and distribution $F(x)$, assumed continuous, are intractable. A good example is the statistic S whose distribution is the weighted sum of independent chi-square variables, each with one degree of freedom, written

$$S = \sum_{i=1}^{k} \lambda_i (u_i)^2 \tag{1}$$

where u_i are i. i. d. $N(0,1)$, and λ_i are known weights. Many quantities in statistics have distributions (often asymptotic distributions) like S; for example, the Pearson X^2 statistic, used in testing fit to a distribution when the distribution tested contains unknown parameters which are estimated by maximising the usual likelihood, rather than the multinomial likelihood, has this distribution with some $\lambda_i \neq 1$. Other goodness-of-fit statistics, of Cramer–von Mises type, based on the empirical distribution function (EDF), also have such asymptotic distributions (see, for example, many examples in Stephens, 1986a).

One of the first examples of S to be tabulated, for $k = 2$, involved errors in target hitting during World War 2: tables for S were produced with some labour by Grad and Solomon (1955) using analytic methods. These have been extended by various authors to higher values of k, but the analysis after $k = 5$ or 6 rapidly

becomes very difficult. Thus in general it is difficult to find exact percentage points of S, but the cumulants κ_r, $r = 1, 2, \ldots$, are very easily obtained:

$$\kappa_r = \sum_{i=1}^{k} \lambda_i^r 2^{r-1}(r-1)! \tag{2}$$

2 Moments and cumulants

In this section we list definitions. The r-th moment about the origin of a random variable X, or equivalently of its distribution $f(x)$, will be called μ_r'; the r-th moment about the mean will be μ_r. The moment generating function $M_X(t)$ of X is defined by

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx; \tag{3}$$

when expanded as a Taylor series,

$$M_X(t) = 1 + \mu t + \frac{\mu_2' t^2}{2!} + \frac{\mu_3' t^3}{3!} + \cdots + \frac{\mu_r' t^r}{r!} + \cdots \tag{4}$$

where $\mu = \mu_1'$ is the mean of X.

Cumulants κ_r are defined through the cumulant generating function $C_X(t) = \log M_X(t)$, where "log" refers to natural logarithm. Then

$$C_X(t) = \kappa_1 t + \frac{\kappa_2 t^2}{2!} + \frac{\kappa_3 t^3}{3!} + \cdots + \frac{\kappa_r t^r}{r!} + \cdots \tag{5}$$

Thus in principle we must find $M_X(t)$ before finding $C_X(t)$.

The following relationships exist between low-order moments and cumulants: $\kappa_1 = \mu_1' = \mu$; $\kappa_2 = \mu_2 = \sigma^2$; $\kappa_3 = \mu_3$; $\kappa_4 = \mu_4 - 3\mu_2^2$. Further relationships may be found in Kendall and Stuart (1977, vol 1).

Suppose $Z = X_1 + X_2 + X_3 + \ldots + X_k$ where X_i are independent random variables. Then a property of moment generating functions is

$$M_Z(t) = M_{X_1}(t)\, M_{X_2}(t)\, M_{X_3}(t) \ldots M_{X_k}(t),$$

so that

$$C_Z(t) = C_{X_1}(t) + C_{X_2}(t) + \cdots + C_{X_k}(t), \tag{6}$$

and it quickly follows, using obvious notation, that

$$\kappa_r(Z) = \kappa_r(X_1) + \kappa_r(X_2) + \cdots + \kappa_r X_k. \tag{7}$$

This additive property makes it very easy to find cumulants of sums of independent random variables, and hence, for example, the cumulants of S.

Two important $M_X(t)$ are those of the $N(\mu, \sigma^2)$ distribution, $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$, and the χ^2_p distribution, $M_X(t) = 1/(1-2t)^{p/2}$. Finally, it is easily shown that $\mu_r(aX+b) = a^r \mu_r(X)$, for $r \geq 2$, where a and b are any real constants, and $\kappa_r(aX+b) = a^r \kappa_r(X), r \geq 2$.

As an example, consider S. If X has a χ^2_1 distribution, the MGF of X is $1/(1-2t)^{1/2}$; thus $C_X(t) = -\frac{1}{2}\log(1-2t)$, and expansion gives $C_X(t) = t + 2t^2 + \frac{8t^3}{3!} + \frac{48t^4}{4!} + \cdots$. Thus the r-th cumulant of X is $\kappa_r = 2^{r-1}(r-1)!$, that of $\lambda_i X$ is $\lambda_i^r \kappa_r$, and by the additive property (7), the r-th cumulant of S is given by the expression (2).

3 Mathematical approximations

The approximations in this section are called "mathematical" because they are based on mathematical analysis, with known properties of accuracy and convergence, in contrast to those to be considered later.

Suppose $n(t)$ is the standard normal density

$$n(t) = e^{-t^2/2}/\sqrt{2\pi} \tag{8}$$

and let $f(x)$ be the (continuous) density of X. Then it is (nearly always) possible to expand $f(x)$ as

$$f(x) = n(x)\left\{1 + \frac{1}{2}(\mu_2 - 1)H_2(x) + \frac{1}{6}\mu_3 H_3(x) + \frac{1}{24}(\mu_4 - 6\mu_2 + 3)H_4(x) + \ldots\right\} \tag{9}$$

called a Gram-Charlier series. The $H_r(x)$ are Hermite polynomials. Lists of Hermite polynomials, and also conditions for convergence, etc., are given in Kendall and Stuart (1977, vol. 1).

The basic technique involved in deriving (9) rests on the fact that Hermite polynomials are orthogonal with respect to the kernel $n(x)$; thus

$$\int_{-\infty}^{\infty} H_i(x)\, H_j(x)\, n(x)\, dx = \begin{cases} 0, i \neq j \\ j!, i = j. \end{cases} \tag{10}$$

Then if $f(x) = \sum_i c_i n(x) H_i(x)$, multiplication by $H_j(x)$ on both sides, and integration, gives

$$c_j = \int_{-\infty}^{\infty} f(x)\, H_j(x)\, dx/j!$$

. When worked out, $c_2 = (\mu_2 - 1)/2, c_3 = \mu_3/6$, etc.

If an *infinite* set of moments is available, as for S, the density can be approximated very accurately using a Gram-Charlier series of sufficient length, but there are many statistics in practical applications for which it is difficult even to get the first four moments — see Solomon and Stephens (1977) for examples. There are two other important drawbacks:

1. A k-term fit might, at any one value of x, be worse than a $(k-1)$-term fit.

2. Gram-Charlier series with finite numbers of moments can give a negative density $f(x)$, particularly in the tails.

3.1 Percentage points approximation

A Gram-Charlier-type expansion can also be found for $F(x)$, the distribution function of X; this can be inverted to give a percentage point for a given cumulative area α. Thus suppose $F(x_\alpha) = \alpha$; we want an approximation to x_α. A **Cornish-Fisher expansion** gives $x - \xi$ as a series in Hermite polynomials in x, or (more practically useful) in ξ, where ξ is the percentile corresponding to α for the normal distribution, that is, ξ is the solution of

$$\int_{-\infty}^{\xi} n(x)\,dx = \alpha. \tag{11}$$

Again, problems can arise with the convergence to the desired x_α. For more details on mathematical expansions of Gram-Charlier or Cornish-Fisher type, see Kendall and Stuart (1977, vol. 1).

4 Pearson curves and other systems

We now turn to a method of approximation which can be thought of as "laying one curve upon another" — the approximating curve has parameters which can be varied to make a good fit. The parameters are usually chosen by matching moments or cumulants. Percentage points of the approximating curve, which are tabulated or otherwise easily found, are then used as approximations to the desired points.

A family of approximating curves is the Pearson system, where the (continuous) density $f(x)$ is approximated by $f^*(x)$, given by

$$\frac{1}{f^*(x)}\frac{df^*(x)}{dx} = \frac{a+x}{b_0 + b_1 x + b_2 x^2}. \tag{12}$$

According to the values of the constants a, b_0, b_1, b_2, integration of the right-hand side will take many forms, giving great flexibility to the system of densities $f^*(x)$. With considerable algebra (see Elderton and Johnson, 1969, for details), the constants may be put in terms of the moments:

$$\text{Suppose } A = 10\mu_4\mu_2 - 18\mu_2^3 - 12\mu_3^2; \text{ then} \tag{13}$$

$$a = \frac{\mu_3(\mu_4 + 3\mu_2^2)}{A}, \tag{14}$$

$$b_0 = \frac{-\mu_2(4\mu_2\mu_4 - 3\mu_3^2)}{A}, \tag{15}$$

$$b_1 = -a; \tag{16}$$

$$b_2 = \frac{-(2\mu_2\mu_4 - 3\mu_3^2 - 12\mu_3^2)}{A}. \tag{17}$$

Thus knowledge of the first four moments or cumulants of X will fix the constants above: a further constant C enters on integrating, but is fixed by the fact that the total integral of $f^*(x)$ must be 1.

4.1 Percentage points

When the constants are known, the density $f^*(x)$ may be integrated and percentage points solved for numerically. Over the years, this was done, at first very laboriously, for a small range of possibilities, but a quite extensive tabulation was made, using electronic computers, in the late '60s. These tables are in *Biometrika Tables for Statisticians*, vol. II. The form of the tables is as follows. The percentage points for X, the *standardised* X-variable given by $X = (x - \mu)/\sigma$, are plotted in a two-way table indexed by the skewness and kurtosis parameters β_1 and β_2. These are defined by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}; \tag{18}$$

they have been defined to be scale-free, and $\sqrt{\beta_1}$ takes the sign of μ_3. β_1 measures skewness: a large (positive) $\sqrt{\beta_1}$ means the curve is skewed towards positive values (long tail is to the right) and *vice versa* for negative $\sqrt{\beta_1}$. A large β_2 (always positive) means the density has heavy tails. Of course, all symmetric distributions have $\beta_1 = 0$; a benchmark to measure kurtosis is the normal distribution for which $\beta_2 = 3$. Since $\kappa_4 = \mu_4 - 3\mu_2^2$, the parameter $\gamma_2 = \beta_2 - 3 = \kappa_4/\kappa_2^2$ can also be regarded as measuring kurtosis, with value $\gamma_2 = 0$ for the normal distribution.

Suppose, for a given S, we have $\sqrt{\beta_1} = 0.8$ and $\beta_2 = 4.6$. To use *Biometrika Tables*, one enters the appropriate $\sqrt{\beta_1}$ table, $\sqrt{\beta_1} = 0.8$, and travels down the left-hand column until the β_2 value, 4.6, is reached. Along the row are 17 tabulated percentage points for X, from $\alpha = 0.00$ to $\alpha = 1.00$. Interpolation must be used for $\sqrt{\beta_1}, \beta_2$ values not explicitly given.

4.2 *Un peu d'histoire*

At this point, perhaps, it might be permitted to enliven the account with what the *Guide Michelin* calls *un peu d'histoire*. At the time *Biometrika Tables* Vol. II were being prepared, I was fortunate enough to know Professor E. S. Pearson, then retired but still very active, especially as Editor of *Biometrika*. He had collaborated with workers in the U. S. to get the tables (Johnson, Nixon,

Amos and Pearson, 1963) and had carefully compiled the full set by hand. He had introduced me to Pearson curves, which, to put it mildly, did not figure prominently in statistical training of the day, and had shown me how effective they could be. He gave me a copy of the tables to use. I undertook to write a Fortran program on the IBM 650, to interpolate and find points, given the first four moments. All 20 tables were then typed onto punched cards; in the end, I got it down to approximately 45 minutes per table. This is not such a dramatic piece of history as *Michelin* usually provides (assignations and assassinations often play a prominent role), but a diminishing generation of modern readers will still empathise with the fears of losing the boxes of cards, getting them wet in the snows of Montréal, etc., not to mention the awful discovery of a wrongly-typed number!

Since then, programs have been written to integrate the density equation for $f^*(x)$ numerically and to solve for x_α for given α, or to provide the tail area for given x; one of these, kindly given to me by Amos and Daniel (1971), has been added to my program; this greatly increases the range of β_1 and β_2 for which Pearson curve approximations can be found. However, points are still output from both the Amos and Daniel part of the program and by the *Biometrika Tables* part, ostensibly as a check where available, but truthfully as a sentimental tribute to E. S. P.

Later on, Charles Davis and I (Davis and Stephens, 1983) added to the program to enable a fit to be made using knowledge of an end point (for example, that the left-hand endpoint of S is zero) and *three* moments. This is especially valuable for the type of statistic for which each successive moment requires exponentially increasing hard work — for example, the distribution of areas, or perimeters, of polygons formed by randomly dropping lines on a plane — see Solomon and Stephens (1977). The Pearson-curve fitting program is available from the author.

Further developments have included algorithms to facilitate use of Pearson curves — see, for example, Bowman and Shenton (1979a, 1979b).

4.3 Accuracy of Pearson curve fits

(a) Pearson curve densities are unimodal, or possibly J- or U-shaped, but never multimodal. They are also never negative.

(b) Percentage points or tail areas found from Pearson curve fitting have been found, for unimodal long-tailed distributions, to be very accurate in the long tail, at least for tail areas bigger then 0.005, or the 0.5% point. Pearson and Tukey (1965) discuss this issue; Solomon and Stephens (1977) give comparisons. (In making comparisons, one must of course compare the Pearson curve fit with the correct x_α, or the correct area for given x, for a distribution which is *not itself* a member of the Pearson family.)

(c) Davis (1975) has made extensive comparisons with Gram-Charlier fits using only four moments. Pearson curve fits are better than Gram-Charlier fits everywhere except for distributions very close to the normal, as measured by the β_1, β_2 values.

4.4 Other systems

Johnson (1949) has proposed another family (divided into three parts) of curves defined by four moments: for example, the S_U curves are those given by the relation

$$\xi = \gamma + \delta \sinh^{-1} X \tag{19}$$

where $X = (x - \mu)/\sigma$, and γ, δ are to be chosen to make the distribution of ξ as close as possible to $N(0,1)$. A discussion, and tables to facilitate the calculation of γ and δ, are in *Biometrika Tables for Statisticians* Vol. II. Other authors have also proposed families of distributions, but they have not come into such common use for the purpose of approximating percentage points.

5 Use of higher moments

We now turn to the first of two interesting questions — can higher moments be used to improve the accuracy of Pearson curve fits in the long tail of the distribution? The long tail will be supposed to lie to the right, as for the distribution of S; then, since higher values of x will contribute more to the higher moments than smaller values, we might suppose that fits using higher moments will improve accuracy. Unfortunately it is not easy to establish the four constants in terms of higher moments — of course, only four of these would be needed to fix the constants. A recursion formula exists to generate higher moments, for $r = 2, 3, \ldots$:

$$r b_0 \mu'_{r-1} + \{(r+1)b_1 + a\}\mu'_r + \{(r+2)b_2 + 1\}\mu'_{r+1} = 0 \tag{20}$$

In this recursion, the constants a, b_0, b_1 and b_2 occur, and this means that one cannot reverse the recursion and generate , say, μ and σ^2 from μ_3, μ_4, μ_5 and μ_6.

Nevertheless, one can generate the fifth and sixth moments of the Pearson curve with the same first four moments of, say, S, and compare them with the *true* fifth and sixth moments of S. The first two moments are then slightly changed, and the procedure successively repeated, until the third, fourth, fifth and sixth moments of each curve match. This will mean that the mean and variance of the Pearson curve will not be exactly the same as those for S, although they will be close, and this will probably make a worse fit in the lower tail; but for higher x the fit could improve. I have made some comparisons using this procedure, but, as one might expect, there appears to be no systematic improvement. In discussion, when this paper was first presented, the suggestion

was made to use Least Squares to make "closest" fits, in order to compare the six moments. More work is needed to compare Pearson curve fits along these various lines, but it is not likely that the improvement will be sure, or will extend to points far into the tails. In the end it must be remembered that one curve is simply being laid on top of another, with only four parameters to vary, and there is no mathematical analysis that will *guarantee* accuracy.

Other methods for developing accuracy in the extreme tails include numerical inversion of the Characteristic Function (essentially the MGF with *it* replacing t, where $i = \sqrt{-1}$), or saddlepoint approximations. A method due to Imhof (1961) uses numerical inversion for distributions such as S, but the computer time needed increases rapidly as the distance into the tails increases (to give small tail areas). Field (1992) has recently examined saddle-point approximations for S. These would seem to give more promise of tail-end accuracy in the long run.

6 Use of sample moments

The second interesting question is: how accurate are Pearson curve fits when sample moments are used to make the fit? In the earliest days, this was the use to which Pearson curves were applied — to find a smooth density to describe a set of data, such as lengths of beans, or width of skulls. Kendall and Stuart (1977, Vol. 1) gives details of such a fit. In general, the Pearson curves will give very good fits to a unimodal set of data, or even to J-shaped or U-shaped sets, but it is important to assess the accuracy of extrapolation from the sample to the supposed population from which it came. More precisely, we ask how close the sample fit estimate of, say, the upper-tail 5% point is to the true population 5% point, and, further, whether or not the Pearson-curve point is better than the estimated point derived from choosing the appropriate order statistic — in a sample of 1000, the 951st value in ascending order, or in a sample of size 10000, the 9501st value. Some investigation of these questions has been undertaken in two quite different ways, by Johnstone (1988) and by myself (Stephens, 1991).

The accuracy of the Pearson curve point will depend on:

1. the sample size n,
2. the α-level (tail area) of the point required,
3. the true skewness and kurtosis of the density approximated,
4. higher moments.

Johnstone gives a small study, for samples from populations with the following range of parameters:

β_1	0.0	0.0	1.0	1.0	2.0
β_2	3.3	4.0	5.25	6.0	7.5

Johnstone gives plots of the estimated coefficient of variation, CV, of the Pearson curve x_α against $-\log\alpha$, where the base of logarithms is 10. Thus the CV of the estimated $x_{0.01}$ is plotted against 2, that of the estimated $x_{0.001}$ is plotted against 3, etc . The coefficient of variation is estimated using a Taylor series approximation. As one might expect, the CV goes up markedly as α gets smaller (so $-\log\alpha$ gets larger on the x-axis), and the steepness of the rise is greater for the more skew distributions .

In Stephens (1991), Monte Carlo samples were taken from populations for which exact percentage points could be found, and the exact points were compared with those obtained from (a) Pearson curve fits using the moments of each sample, and (b) the order statistic estimate from each sample. The order statistic estimate will be asymptotically unbiased, while one can say nothing exact about the point obtained by laying one curve on another; recall that sample moments, especially the third and fourth, are extremely variable, even for quite large samples. The results showed, as expected, that the Pearson curve points were more biased. However, somewhat surprisingly, they had smaller mean square error. Therefore, it might well be preferable to use the Pearson curve points, although, again, more investigations should be made especially if the points required are far into the tail.

7 Goodness of fit using moments

In this second part of the paper, we discuss how moments are used in Goodness-of-Fit, that is, to test whether a random sample comes from a given (continuous) distribution. The distribution will often have unknown parameters, which must be estimated from the given sample.

7.1 Tests based on skewness and kurtosis

Suppose the r-th sample moment m_r about the mean is defined by

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r. \tag{21}$$

The sample skewness and sample kurtosis are then defined by

$$b_1 = \frac{m_3^2}{m_2^3}, b_2 = \frac{m_4}{m_2^2}. \tag{22}$$

These statistics are not unbiased estimates of β_1 and β_2, but they are consistent, that is, the bias diminishes with increasing sample size. The sample skewness and kurtosis are time-honoured statistics for testing normality, having been used in a rather *ad hoc* manner for most of this century; b_1 is compared with zero,

and b_2 with β, the value of β_2 for the normal distribution. However, distribution theory of b_1 and b_2 is difficult, and it is only since computers have been available that extensive and reliable tables of significance points have existed for these statistics. Further, b_1 and b_2 can be combined to give one overall statistic (d'Agostino and Pearson, 1973, 1974; d'Agostino, 1986). For other distributions Bowman and Shenton (1986) have also given tables for these statistics. Studies have shown that skewness and kurtosis, especially combined, provide good omnibus tests for normality, although less is known for other distributions. For the important discrete distribution, the Poisson, all cumulants are equal to the mean, denoted by the parameter λ; a time-honoured test for the Poisson is based on the ratio of sample variance to sample mean, which of course should be about one. Again, this simple statistic appears to compete well with others in terms of power.

7.2 A formal technique based on moments

Perhaps because of the variability of sample moments, which makes calculation of significance points difficult for statistics based on these moments when calculated from samples of reasonable size, it took some time to formalize a technique based on moments. Gurland and Dahiya (1970) and Dahiya and Gurland (1972) have however devised a general procedure. The essential steps are as follows:

1. A vector ζ of length s, say, must be found, whose components ζ_i are functions of the theoretical moments, and such that each component ζ_i is linear in the parameters. (This might involve re-parametrising the distribution from its usual form).

2. The estimate h of ζ is obtained by replacing theoretical moments by sample moments.

3. The test statistic is then based on the difference $h - \zeta$.

Suppose that Σ is the covariance matrix of h, θ is the q-vector of unknown parameters, and W is the $s \times q$ matrix such that $\zeta = W\theta$. Then define

$$\hat{Q}_t = n(h - W\hat{\theta})'\hat{\Sigma}^{-1}(h - W\hat{\theta}),$$

where $\hat{\theta} = (W'\hat{\Sigma}^{-1}W)^{-1}W'\hat{\Sigma}^{-1}h$. The statistic $\hat{\theta}$ is the regression estimate of θ obtained by generalized least squares, and $\hat{\Sigma}$ is Σ with the estimate $\hat{\theta}$ used wherever θ appears.

Gurland and Dahiya (1970, 1972) showed that, asymptotically, the test statistic $\hat{Q}_t$ has the χ^2 distribution with $t = s - q$ degrees of freedom. Currie and Stephens (1986, 1990) have studied the procedure, and show several properties of $\hat{Q}_t$. Among these are the fact that the test statistic $\hat{Q}_t$ can be broken into t components, each with asymptotic distribution χ^2_1, and each testing different

features of the distribution. Each component is a function of moments or cumulants. For example, consider the test for normality, that is, for the distribution $N(\mu, \sigma^2)$. Gurland and Dahiya (1970) took $\zeta' = \{\mu, \log\mu_2, \mu_3, \log(\mu_4/3)\}$, so that $h' = \{\bar{x}, \log m_2, m_3, \log(m_4/3)\}$. The matrix W is $W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 2 \end{bmatrix}$, and $\theta = \begin{bmatrix} \mu \\ \log\sigma^2 \end{bmatrix}$. The test statistic $\hat{Q}_2$ becomes $\hat{c}_1 + \hat{c}_2$, where the two components are $\hat{c}_1 = nm_3^2/6m_2^3$ and $\hat{c}_2 = (3n/8)\{\log(m_4/3m_2^2)\}$. Thus the method leads to $nb_1/6$ and $(3n/8)\log(b_2/3)$ as test statistics, equivalent to the "old-fashioned" b_1 and b_2.

However, it should be noted that the components are not unique; they depend on how ζ is formed. Currie and Stephens (1986, 1990) discuss these questions in some detail.

8 Components of other goodness-of-fit statistics

Other goodness-of-fit statistics also have components which are functions of moments. The oldest of these was proposed by Neyman (1937), in connection with a test for uniformity.

A test for a fully specified continuous distribution (that is, all parameters known) can always be converted to a test for uniformity by means of the Probability Integral Transformation, and a test for the exponential distribution can also be so converted, even when the scale and origin parameters are not known, so that Neyman's test has wider applicability than it might at first appear. (For details of these transformations, see Stephens, 1986a, 1986b).

Neyman's test is as follows: suppose the test is that Z has a uniform distribution between 0 and 1, written $U(0,1)$. On the alternative, let the logarithm of the density of Z be expanded as a series of Legendre polynomials:

$$\log(f(z)) = A(c)\{1 + c_1 L_1(z) + c_2 L_2(z) + c_3 L_3(z) + \cdots\}, \qquad (23)$$

where the c_i are coefficients, components of the vector c, $L_i(z)$ is the i-th Legendre polynomial, and $A(c)$ is a normalising constant.

A test for uniformity is then a test that all $c_i = 0$. The estimates of c_i are

$$\hat{c}_i = \sum_{j=1}^{n} L_i(z_j) \qquad (24)$$

where $z_1, z_2, \ldots, z_n$ is the given sample.

The first few Legendre polynomials are best expressed in terms of $y = z - 0.5$. Then

$$L_1(z) = 2\sqrt{3}y, \tag{25}$$

$$L_2(z) = \sqrt{5}(6y^2 - 0.5), \tag{26}$$

$$L_3(z) = \sqrt{7}(20y^3 - 3y), \tag{27}$$

so that the estimate $\hat{c}_1$ becomes a function of the first moment about the known mean 0.5, the second estimate $\hat{c}_2$ becomes a function of the second moment, $\hat{c}_3$ a function of both the third and the first moments, etc.

Neyman shows that the suitably normalised $\hat{c}_i$ have asymptotic $N(0,1)$ distributions, and his overall test statistic is the sum of the squares of these normalised estimates. Thus the overall statistic has an asymptotic χ^2 distribution, just as for the Dahiya-Gurland statistic, and the individual terms, based on moments, are the components of the overall test statistic.

9 EDF statistics

Another important family of goodness-of-fit statistics is that derived from the Empirical Distribution Function (EDF) of the z-sample. This family includes the well-known Kolmogorov-Smirnov statistic and the Cramer-von Mises family of statistics (for details and tests for many distributions based on these, see Stephens, 1986a).

One of the most important of the Cramer-von Mises class is A^2, introduced by Anderson and Darling (1954). The definition of A^2 is based on an integral involving the difference between the EDF and the tested distribution $F(x)$ (with parameters estimated if necessary). The working formula is

$$A^2 = -n - \frac{1}{n}\sum_i (2i-1)\left[\log z_{(i)} + \log(1 - z_{(n+1-i)})\right], \tag{28}$$

where $z_i = F(x_i)$, and $z_{(i)}$ are the order statistics.

As an omnibus test statistic, A^2 has been shown to perform well in many test situations.

Anderson and Darling showed that the asymptotic distribution of A^2 is, like S of Section 1, a sum of weighted χ^2 variables. The individual terms in the sum can again be regarded as components of the entire statistic, and Stephens (1974) has investigated these components in some detail. A remarkable result is that they too are based on Legendre polynomials, so that they are effectively the same as the Neyman components, based on moments of the z-sample. There has been some investigation of components of these and other statistics, as individual test statistics for the distribution under test; references are given by Stephens(1986a). As for the Gurland-Dahiya components, they can

be expected to be sensitive to different departures from the tested distribution. The complete test statistics of Neyman and of Anderson-Darling combine the same components, but with different weightings.

References

Amos, D.E. and Daniel, S.L. (1971) Tables of percentage points of standardized Pearson distributions. Research Report SC-RR-71 0348, Sandia Laboratories, Albuquerque, New Mexico.

Anderson, T.W. and Darling, D.A. (1954) A test of goodness of fit. *J. Amer. Statist. Assoc.* **49** 765–769.

Bowman, K. O. and Shenton, L.R. (1979a) Approximate percentage points for Pearson distributions. *Biometrika* **66** 145–151.

Bowman, K. O. and Shenton, L.R. (1979b) Further approximate Pearson percentage points and Cornish-Fisher. *Comm. Stat. B — Simula. Computa.* **8** 231–234.

Bowman, K. O. and Shenton, L. R. (1986) Moment $(\sqrt{b_1}, b_2)$ techniques. In d'Agostino, R. B. and Stephens, M. A., eds., 1986.

Currie, I. D. and Stephens, M. A. (1986a) Gurland-Dahiya statistics for the exponential distribution. Technical report, Dept. of Mathematics and Statistics, Simon Fraser University.

Currie, I. D. and Stephens, M. A. (1986b) Gurland-Dahiya statistics for the inverse Gaussian and gamma distributions. Technical report, Dept. of Mathematics and Statistics, Simon Fraser University.

Currie, I. D. and Stephens, M. A. (1990) Tests of fit using sample moments. *Scand. J. Statist.* **17** 147–156.

d'Agostino, R. B. (1986) Tests for the normal distribution. Chapter 9 in *Goodness-of-Fit Techniques*, (d'Agostino, R. B. and Stephens, M. A., eds.) Marcel Dekker: New York.

d'Agostino, R. B. and Pearson E. S. (1973) Tests for departures from normality. Empirical results for the distribution of b_2 and $\sqrt{b_1}$. *Biometrika* **60** 613–622.

d'Agostino, R. B. and Pearson E. S. (1974) Correction and amendment. Tests for departures from normality. Empirical results for the distribution of b_2 and $\sqrt{b_1}$. *Biometrika* **61** 647.

Dahiya, R. C. and Gurland, J. (1972) Goodness-of-fit tests for the gamma and exponential distributions. *Technometrics* **14** 791–801.

Davis, C. S. (1975) Unpublished MSc. thesis, Dept. of Mathematics, McMaster Univ., Hamilton, Ontario.

Davis, C. S. and Stephens, M. A. (1983) Approximate percentage points using Pearson curves. Algorithm AS 192, *Appl. Stat.* **32** 322-327.

Elderton, W. P. and Johnson, N. L. (1969) *Systems of frequency curves.* Cambridge University Press: London.

Field, C. (1992) Tail areas of sums of weighted chi-squares. To appear.

Grad, A. and Solomon, H. (1955) Distribution of quadratic forms and some approximations. *Ann. Math. Statist.* **26** 464–477.

Gurland, J. and Dahiya, R. C. (1970) A test of fit for continuous distributions based on generalised minimum chi-squared. *Statistical Papers in honor of G. W. Snedecor* 115–127 (ed. T. A. Bancroft), Iowa State Univ. Press.

Imhof, J. P. (1961) Computing the distribution of quadratic forms in normal variables. *Biometrika* **48** 417–426.

Johnson, N. L. (1949) Systems of frequency curves generated by methods of translation. *Biometrika* **36** 149–176.

Johnson, N. L., Nixon, E., Amos, D. E. and Pearson, E. S. (1963) Table of percentage points of Pearson curves, for given $\sqrt{\beta_1}$ and β_2, expressed in standard measure. *Biometrika* **50** 459–498.

Johnstone, I. (1986) A program for estimating uncertainties in quantile estimates derived from empirical Pearson fits. Technical report no. 381, Dept. of Statistics, Stanford Univ.

Kendall and Stuart (1977) *The Advanced Theory of Statistics*, Vol. 1, Charles Griffin: London.

Neyman, J. (1937) "Smooth" tests for goodness-of-fit. *Skand. Aktuarietidskr.* **20** 149–199.

Pearson, E. S. and Tukey, J. W. (1965) Approximate means and standard deviations based on distances between percentage points of frequency curves. *Biometrika* **52** 533–546.

Solomon, H. and Stephens, M. A. (1977) Distribution of a sum of weighted chi-square variables. *J. Amer. Statist. Assoc.* **72** 881–885.

Stephens, M. A. (1974) Components of goodness-of-fit statistics. *Ann. Inst. Henri Poincaré* B. **10** 37–54.

Stephens, M. A. (1986a) Tests based on EDF statistics. Chapter 4 in *Goodness-of-Fit Techniques*, (d'Agostino, R. B. and Stephens, M. A., eds.) Marcel Dekker: New York.

Stephens, M. A. (1986b) Tests for the exponential distribution. Chapter 11 in *Goodness-of-Fit Techniques*, (d'Agostino, R. B. and Stephens, M. A., eds.) Marcel Dekker: New York.

Stephens, M. A. (1991) Computer problems in goodness-of-fit. *The Frontiers of Statistical Computation, Simulation and Modeling* (Dudewicz, J., Nelson, P., Ozturk, A., eds.) American Sciences Press: Syracuse.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Report No. 469	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Moments in Statistics: Approximations to Densities and Goodness-of-Fit		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Michael A. Stephens		8. CONTRACT OR GRANT NUMBER(s) N0025-92-J-1264
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305-4065		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 111		12. REPORT DATE May 11, 1993 13. NUMBER OF PAGES 18
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Solomon and Stephens examples of statistics X; statistic S; Pearson statistic; goodness-of-fit, EDF (empirical distribution function), errors in target-hitting, World War 2.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this article we discuss ways in which moments are used (a) to approximate distributions. and (b) to test fit to a given distribution.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)