The Gradient Statistic

George R. Terrell
Statistics Department, Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

Abstract: We propose an alternative to the usual large-sample test statistics, the gradient statistic. Its behavior is similar to the others asymptotically, but in many cases it takes a simpler form and is easier to compute. There is an Edgeworth-like series for improving probability approximations in the univariate case.
Key words: Likelihood methods; score statistic; Wald statistic; small-sample asymptotics.

I. Introduction

Large-sample tests of hypotheses about one or more parameters beyond the classical case of the normal linear model usually must resort to normal theory approximations. Wilks (1938) proposed the log-likelihood ratio statistic. Other influential proposals include the Wald (1943) statistic and the Rao (1947) score statistic. All have asymptotically chi-squared distribution. They vary in how convenient a form they take, how readily they may be computed, and how accurate is their asymptotic distribution theory (see e.g. Severini (2000)).

We will here propose a remarkably simple alternative test statistic, the gradient statistic, with the same asymptotic distribution. We will establish some properties, then develop a higher-order asymptotic theory in the single-parameter case. Our tool will be a two-point analog to an Edgeworth expansion.

II. Inference from Maximum Likelihood Estimates.

The test statistic we will investigate is closely related to each of the three more familiar test statistics for an hypothesis $\boldsymbol{\theta}$ about a vector parameter of dimension p, estimated from a vector of observations $\mathbf{x}$, with likelihood $L(\boldsymbol{\theta}, \mathbf{x})$. Let $l(\boldsymbol{\theta}, \mathbf{x}) = \log L(\boldsymbol{\theta}, \mathbf{x})$. The condition $\frac{\partial l(\hat{\boldsymbol{\theta}}, \mathbf{x})}{\partial \boldsymbol{\theta}} = 0$ will determine the maximum likelihood estimate. Remember that, under mild regularity conditions, the score vector $\frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}$ is asymptotically distributed $N(\mathbf{0}, \mathrm{I}_{\boldsymbol{\theta}})$. Then it is immediate that the Rao score statistic $R^2 = \frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \mathrm{I}_{\boldsymbol{\theta}}^{-1} \frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}$ is asymptotically chi-squared on p degrees of freedom.

Under much stronger regularity conditions, the maximum likelihood estimate $\ddot{\boldsymbol{\theta}}$ is asymptotically $N(\boldsymbol{\theta}, \mathrm{I}_{\boldsymbol{\theta}}^{-1})$, so that one version of the Wald test $W^2 = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathrm{T}} \mathrm{I}_{\boldsymbol{\theta}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ has the same asymptotic distribution. (Alternative versions of the test with the same asymptotic distribution use information calculated at the maximum likelihood or estimated from the data; these are often more convenient in practice.) Furthermore, it is a corollary to the usual proofs of the distribution of $\ddot{\boldsymbol{\theta}}$ that it and the score vector are asymptotically perfectly correlated.

To make the connection explicit, choose any square root A of the information matrix; that is, find a square solution to $\mathrm{A}^{\mathrm{T}}\mathrm{A} = \mathrm{I}_{\boldsymbol{\theta}}$. Then the standardized test vectors $(\mathrm{A}^{-1})^{\mathrm{T}} \frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}$ and $\mathrm{A}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ each have asymptotic distributions $\mathrm{N}(\mathbf{0}, \mathrm{I}_p)$; in fact, the two vectors converge to each other in probability. Then the inner product of these standardized test vectors also has asymptotically chi-squared p distribution:

$$\left[\left(\mathrm{A}^{-1}\right)^{\mathrm{T}}\frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}\right]^{\mathrm{T}} \mathrm{A}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \mathrm{A}^{-1}\mathrm{A}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$$

This last expression is remarkably simpler than the equivalent statistics from which it was derived, involving no knowledge nor estimates of the information, and no matrices. Notice also that, unlike its progenitors, it is symmetric in the hypothesized $\boldsymbol{\theta}$ and the observed $\ddot{\boldsymbol{\theta}}$ (remember that $\frac{\partial l(\hat{\boldsymbol{\theta}}, \mathbf{X})}{\partial \boldsymbol{\theta}} = \mathbf{0}$). This suggests that it is of independent interest.

Definition: Given an observation vector $\mathbf{x}$ hypothesized to arise from a density f dependent on a p-dimensional parameter vector $\boldsymbol{\theta}$, let the log-likelihood be denoted by $l(\boldsymbol{\theta}, \mathbf{x}) = \log f(\mathbf{x} \mid \boldsymbol{\theta})$. Let $\tilde{\boldsymbol{\theta}}$ be an estimate of $\boldsymbol{\theta}$. Then a **gradient statistic** for testing the hypothesis $\boldsymbol{\theta}$ will be $F^2 = \frac{\partial l(\boldsymbol{\theta}, \mathbf{x})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$.

Example 1: Let the counts in a set of k independent categories have observed counts n_i hypothesized to follow Poisson(m_i) distributions. Then of course $\hat{m}_i = n_i$. Up to an additive constant, the log-likelihood is $\sum_{i=1}^{k} n_i \log m_i - m_i$. Then $F^2 = \sum_{i=1}^{k} \frac{\left(n_i - m_i\right)^2}{m_i}$, the Pearson goodness of fit statistic.

III. Some Properties of F^2 .

Let us make explicit what we suggested in our informal derivation.

Theorem 1: Let our data $\mathbf{x}$ consist of an i.i.d. sample of n, and let regularity conditions A1 to A6 of Chapter 6.2 of Bickel and Doksum (2001) hold for our log-likelihood. Let $\tilde{\boldsymbol{\theta}}$ be an asymptotically efficient estimate of $\boldsymbol{\theta}$. Then $F^2 = \frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$ has asymptotically in n a chi-squared distribution with p degrees of freedom.

Proof: Bickel and Doksum establish in their (6.2.2) that the maximum likelihood estimate $\ddot{\boldsymbol{\theta}}$ is asymptotically efficient; and that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \mathrm{I}_{\boldsymbol{\theta}}^{-1} \frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}} + o_p\left(n^{-1/2}\right)$. Since $\frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}} = O_p\left(n^{1/2}\right)$, we have that

$$\frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \mathrm{I}_{\boldsymbol{\theta}}^{-1} \frac{\partial l(\boldsymbol{\theta}, \mathbf{X})}{\partial \boldsymbol{\theta}} + o_p(1) .$$

Then $F^2 = R^2 + o_p(1)$. Under condition A2, the score statistic has the indicated distribution, so the gradient statistic does too. Finally we note that $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + o_p\left(n^{-1/2}\right)$. Q.E.D.

Particular applications, such as linear regression with Laplace errors, suggest that condition A3, requiring that the likelihood be twice differentiable, may be too strong.

The gradient statistic has one peculiarity: it is not transparently non-negative, even though it must be so asymptotically. Since the likelihood-ratio statistic and the (expected) score statistic must obviously be non-negative; and the empirical score statistic and the Wald statistic are, too, under transparent conditions, the question is natural. (At the seminar at which the statistic was introduced, this was the first question from the floor.) Furthermore, in a subsequent section we shall look at its square root.

Definition: (Dharmadhikari and Joag-dev 1988) A function $f(\mathbf{x})$ is **star-unimodal** if there exists a point $\mathbf{x}$, its mode, such that on any ray extending from $\mathbf{x}$, $\mathbf{y} = \mathbf{x} + t\mathbf{z}$ where $t \bullet 0$, we have for $t_2 > t_1$ that always $f(x + t_1 z) \geq f(x + t_2 z)$.

That is, the function decreases uniformly as we proceed in any straight line from the mode. This condition not being a familiar one, let us connect it to another.

Definition: A function $f(\mathbf{x})$ is **strongly unimodal** at a maximizer $\mathbf{x}$ if log f is concave downward.

Proposition: If f is strongly unimodal, then it is star unimodal.

Proof: Notice that the condition for star unimodal is invariant under log transformation. For log f convex, there is a linear function $L_{\mathbf{x}}(\mathbf{y})$ for any $\mathbf{x}$ defined on each line $\mathbf{y} = \mathbf{x} + t\mathbf{z}$, $-\infty < t < \infty$, so that $L_{\mathbf{x}}(\mathbf{x}) = \log f(\mathbf{x})$ and always $L_{\mathbf{x}}(\mathbf{y}) \geq \log f(\mathbf{y})$. Now let $\mathbf{x}$ be the mode, and consider any direction $\mathbf{z}$. For $t_2 > t_1 \geq 0$, let $\mathbf{y}_1 = \mathbf{x} + t_1\mathbf{z}$. Then

$$L_{\mathbf{y}_1}(\mathbf{x}) \geq \log f(\mathbf{x}) \geq \log f(\mathbf{y}_1) = L_{\mathbf{y}_1}(\mathbf{y}_1) \geq L_{\mathbf{y}_1}(\mathbf{x} + t_2\mathbf{z}) \geq \log f(\mathbf{x} + t_2\mathbf{z}) \quad \text{Q.E.D}$$

We will be applying these concepts to a likelihood $L(\boldsymbol{\theta})$ whose maximum is at $\ddot{\boldsymbol{\theta}}$, where as usual $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.

Theorem 2: (Terrell and Ye). Let L be star unimodal, and differentiable at some $\boldsymbol{\theta}$. Then $F^2 = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^{\mathbf{T}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \geq 0$.

Proof: By way of contradiction, posit a $\boldsymbol{\theta}$ such that $\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^{\mathbf{T}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) < 0$. In particular $\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \neq \mathbf{0}$. Since this gradient vector is the direction in which l is locally most rapidly increasing, then $\ddot{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is a direction in which l is locally decreasing, because l is differentiable and the angle between the two directions is greater than a right angle. Thus there exists a sufficiently small $t_1 > 0$ such that $l(\boldsymbol{\theta}) > l\left[\boldsymbol{\theta} + t_1(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right] \leq l(\hat{\boldsymbol{\theta}})$. But $\boldsymbol{\theta} + t_1(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is in the line segment connecting $\boldsymbol{\theta}$ to $\ddot{\boldsymbol{\theta}}$. This contradicts star unimodality of l, and so of L. Q.E.D

In practice, it will be often be easiest to check the log-convexity of the likelihood; but there exist likelihoods, such as multivariate generalizations of the t-distributions, which are star unimodal but not strongly unimodal. Notice that the theorem does not guarantee the non-negativity when we use an estimate $\tilde{\boldsymbol{\theta}}$ that is not maximum likelihood.

One interesting fact about our statistic follows from a standard proof of the information inequality (Cramér-Rao bound). Let us consider the case where $\tilde{\boldsymbol{\theta}}$ is in fact an unbiased estimate. Then

$$\mathrm{Cov}\left(\tilde{\boldsymbol{\theta}}, \frac{\partial l(\boldsymbol{\theta}, \boldsymbol{X})}{\partial \boldsymbol{\theta}}\right) = \int \tilde{\boldsymbol{\theta}} \frac{\partial l(\boldsymbol{\theta}, \boldsymbol{X})}{\partial \boldsymbol{\theta}} L(\boldsymbol{X}) d\boldsymbol{X} = \int \tilde{\boldsymbol{\theta}} \frac{\partial L(\boldsymbol{X})}{\partial \boldsymbol{\theta}} d\boldsymbol{X}.$$

By regularity, $= \frac{\partial}{\partial \boldsymbol{\theta}} \int \tilde{\boldsymbol{\theta}} L(\boldsymbol{X}) d\boldsymbol{X} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = I$ by unbiasedness. Now consider the gradient statistic:

$$\mathrm{E}\left[\frac{\partial l(\boldsymbol{\theta}, \mathbf{x})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right] = \mathrm{E}\left\{\mathrm{tr}\left[\frac{\partial l(\boldsymbol{\theta}, \mathbf{x})}{\partial \boldsymbol{\theta}}^{\mathrm{T}} \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right]\right\} = \mathrm{E}\left\{\mathrm{tr}\left[\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \frac{\partial l(\boldsymbol{\theta}, \mathbf{x})}{\partial \boldsymbol{\theta}}^{\mathrm{T}}\right]\right\}$$

$$= \mathrm{tr}\,\mathrm{E}\left\{\left[\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \frac{\partial l(\boldsymbol{\theta}, \mathbf{x})}{\partial \boldsymbol{\theta}}^{\mathrm{T}}\right]\right\} = \mathrm{tr}\mathrm{Cov}\left[\tilde{\boldsymbol{\theta}}, \frac{\partial l(\boldsymbol{\theta}, \mathbf{x})}{\partial \boldsymbol{\theta}}\right] = \mathrm{tr}\,\mathrm{I}_p = p.$$

We now know

Theorem 3: For $\tilde{\boldsymbol{\theta}}$ unbiased, $\mathrm{E}(F^2) = p$.

When should we consider the gradient statistic in preference to the classical test statistics? Its formal simplicity, which in practice often makes it the easiest to compute (see the next section), is always an attraction. In complex problems, having neither to create, estimate, nor invert an information matrix is a major plus. Of course the log-likelihood ratio statistic is even simpler in appearance; but in many instances, the derivative of the log-likelihood is quite a bit simpler than the log-likelihood itself. In the case of generalized linear models for ranks (Terrell 1998), these derivatives are within the reach of modern computing equipment, while the likelihood itself is currently computationally intractable. It will be suggested in later sections that the symmetry that the gradient statistic shares with the likelihood ratio causes the accuracy of the chi-squared approximation to deteriorate more slowly in the far tails of the null distribution than the score or Wald statistics.

Theorem 3 raises another important applications issue: it suggests that we may often improve the chi-squared approximation of the gradient statistic by using a less-biased estimate of $\boldsymbol{\theta}$. This may be accomplished in two ways. Note that our statistic lacks one desirable feature of the likelihood ratio and the original score statistic: it is not invariant under non-linear reparametrization of $\boldsymbol{\theta}$. But this means we may improve its behavior by choosing a parametrization in which the maximum likelihood estimate is unbiased. This happened in the example above in which it reproduced the Pearson statistic.

Another way to take advantage of Theorem 3 is to adjust the maximum likelihood estimate to a new asymptotically efficient estimate $\ddot{\boldsymbol{\theta}}$ that is unbiased or nearly so; and thereby improve the distributional approximation.

Example 2: In the case of a Poisson model for counts, the most usual parametrization is the log-linear model for the cell-means: $\eta_i = \log m_i$. The maximum likelihood estimate is of course $\hat{\eta}_i = \log n_i$. Then our goodness-of-fit static becomes $F^2 = \sum_{i=1}^{k} (n_i - m_i) \log \frac{n_i}{m_i}$. Notice that it is similar to (but slightly simpler than) the familiar likelihood ratio G^2. It is standard that the estimate of the log of a Poisson mean may have the order of its bias reduced by adding one-half to the observed count: $\tilde{\eta}_i = \log (n_i + \frac{1}{2})$. Our new gradient statistic becomes $F^2 = \sum_{i=1}^{k} (n_i - m_i) \log \frac{n_i + \frac{1}{2}}{m_i}$. Some experiments suggest that its distributional fit is indeed better than the unadjusted version; and may in fact be better than G^2.

IV. Application to Logistic Regression

A common model for which exact tests quickly become intractable is the linear logistic regression model, in which we wish to predict, from observed failures and successes, the probability of success when covariate information is available. The model is

$\mathrm{P}(success) = p = \dfrac{e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}}{1 + e^{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}}}$, where $\mathbf{x}$ is a vector of covariates and $\boldsymbol{\beta}$ is the hypothesized vector of coefficients. Let $\hat{p}$ be an indicator variable for success. We estimate $\ddot{\boldsymbol{\beta}}$ by maximum likelihood. It may quickly be checked by taking a second derivative that the likelihood is strongly unimodal.

Our gradient statistic is then $F^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \sum_{i=1}^{n} \mathbf{x}_i (\hat{p}_i - p_i)$; it tests the hypothesis $\boldsymbol{\beta}$, and has asymptotically a chi-squared distribution on degrees of freedom equal to the dimension of the space of covariates. It should be noted that this is simpler than any of the three classical statistics.

Example 3: Manly (1991) reports the mandible lengths in m.m. of 10 female and 10 male jackal skulls:

lng	105	106	107	107	107	108	110	110	110	111
sex	F	F	M	F	F	F	F	M	F	F
lngt	111	111	111	112	113	114	114	116	117	120
sex	M	F	F	M	M	M	M	M	M	M

We ask, can we predict the sex of a jackal from the mandible length of a skull? We will use the linear logistic model $\alpha + \beta(x_i - \bar{x})$ where x is the mandible length. The null hypothesis will assert that length is irrelevant, so $\beta = 0$. We will take care of the nuisance parameter a by conditioning on its maximum likelihood value under the null hypothesis when we compute the gradient; and then using its new (but close) maximum likelihood value when we estimate $\ddot{\beta}$. That estimate turns out to be $-.1508 - .6085(x - 111)$. For each of our competing test statistics, we will take a signed square root of the chi-squared on one degree of freedom, to get an approximate standard normal test.

The score statistic is –2.84; the Wald statistic is –2.19; the log-likelihood ratio statistic is –3.24; and our proposed gradient statistic is –3.82. There is quite a bit of variation here; a common experience when we fit non-normal models to small samples. However, they all turn out to reject the null hypothesis under common critical values; mandible length does provide a clue to sex.

V. Univariate Asymptotic Theory.

The author became interested in the test-statistic problem through small-sample asymptotics, the search for corrections to improve the fit to the limiting distribution. If we can do the same thing for the gradient statistic, we may make it more useful; and in addition we will obtain estimates of the accuracy of the approximation.

We will address the rate of convergence of the gradient statistic to its asymptotic form in the single parameter case. Consider the simplest class of models, the natural exponential family. Then $l(\theta, x) = H(x) + \theta x - K(\theta)$. The moment generating function of this random variable is $m(s) = \mathrm{E}\left(e^{xs}\right) = e^{K(\theta + s) - K(\theta)}$; therefore, its cumulant generating function is $k(s) = \log m(s) = K(\theta + s) - K(\theta)$. Thus $k'(0) = \mathrm{E}(X) = \mu$ under the null hypothesis θ. The condition for a maximum likelihood estimate, $\dfrac{\partial l(\hat{\theta})}{\partial \theta} = 0$, is then $K'(\hat{\theta}) = x$. Let $t = \ddot{\theta} - \theta$. Then $k'(t) = K'(\ddot{\theta}) = x$; this is then E($X$) under the empirical alternative hypothesis $\ddot{\theta}$. Furthermore, Var(X) is $k''(0)$ under the null hypothesis and $k''(t)$ under the alternative hypothesis.

In this univariate case, we study normal approximation to the distribution of the sufficient statistic x; squared and in the multivariate case that becomes a chi-squared statistic.

Since for a Normal(μ, σ^2) random variable, $k(s) = \mu s + \sigma^2 s^2/2$, our approach will be to find various simple quadratic approximations to $k(s)$. Initially, we will require that the approximation integrate to 1; i. e., that $k(0) = 0$. We then determine the approximation by imposing two further conditions on k to obtain μ and σ. Since we have information about k at the two points 0 and t, that may be accomplished in several alternative ways.

(1) Require $k'(0) = \mu$ and $k''(0) = \sigma^2$, the null values. Then the standard normal test statistic is $z = \frac{x - \mu}{\sigma}$. Expressing that in terms of the log-likelihood, $z = \frac{\partial l(\theta)}{\partial \theta}\left(-\frac{\partial^2 l(\theta)}{\partial \theta^2}\right)^{-1/2}$. Squaring that, you see the Rao (1947) score statistic (in the natural exponential family, the distinction between observed and expected information disappears).

(2) Require $k'(t) = x$ and that $k''(t)$ interpolate the correct value. Then our test statistic is $z = t\sqrt{k''(t)}$. In terms of likelihood, this is $z = (\hat{\theta} - \theta)\left(-\frac{\partial^2 l(\hat{\theta})}{\partial \theta^2}\right)^{1/2}$; which corresponds to the Wald (1943) statistic.

(3) Require $k'(t) = x$ and that $k(t)$ interpolate the correct value. Then $z = \operatorname{sgn}(t)\sqrt{2\left[xt - k(t)\right]}$. The likelihood form is $z = \operatorname{sgn}(\hat{\theta} - \theta)\sqrt{2\left[l(\hat{\theta}) - l(\theta)\right]}$; the univariate version of Wilks' (1938) log-likelihood ratio statistic, called the deviance.

(4) Require $k'(0) = \mu$ and $k'(t) = x$. Then we have $z = \operatorname{sgn}(t)\sqrt{t(x - \mu)}$; and $z = \operatorname{sgn}(\hat{\theta} - \theta)\sqrt{\frac{\partial l(\theta)}{\partial \theta}(\hat{\theta} - \theta)}$ is the univariate gradient statistic, which we shall call the **gradience**.

Other possibilities seem less interesting.

Example 4: A Gamma(α) random variable is exponential with natural parameter the inverse of its scale. Let $\alpha = 16$, and look at tail probabilities (the smaller of $X \bullet x$ and $X \bullet x$): In this family, the Rao and Wald statistics happen to coincide, though this is not generally true.

x	exact P	Rao	Wilks	gradience
8	.00823	.02275	.006458	.00234
12	.1556	.1587	.1361	.1241
20	.1565	.1587	.1770	.1855
24	.0344	.0228	.0410	.0512
28	.00543	.00135	.00679	.01167

It seems to often be the case that the likelihood-ratio behaves best over a wide range, the score and Wald statistics behave best near μ, and the gradience is in between.

Daniels (1954) proposed using the saddlepoint principle to improve the approximation of the Wald statistic. The effect in the current context is to interpolate the correct value $k(t)$ instead of $k(0) = 0$. Then the approximation no longer integrates to 1; but the density at x and the tail probability from x are usually more accurate. The effect is to multiply the probability approximation by $e^{k(0) - xt + k''(0)t^2/2}$. As Daniels noted, this actually improves the order of an estimate of the density; but not of the tail probability.

Of course the log-likelihood statistic already interpolates $k(t)$. When the gradience is adjusted to interpolate this value, the effect is to multiply each density and probability by $e^{k(0)-t(\mu+\frac{1}{2})}$.

Example 5: In the Gamma example (4)

x	exact P	tilted Wald	tilted grad
8	.00823	.00765	.00581
12	.1556	.1432	.1323
20	.1565	.1703	.1801
24	.0344	.0371	.0428
28	.00543	.00578	.00726

The effect of tilting to the empirical value is to hurt the Wald approximation near 0, but to improve it greatly away from zero. The approximation of the gradience is improved, but only modestly. We might understand this by noting that the Wald statistic concentrates its information at the empirical value, and tilting only accentuates that. The gradience is a compromise between the null and observed value, so tilting toward one or the other has less effect.

VI. Higher-Order Corrections.

The classical way to construct higher-order approximations to the score statistic is the **Edgeworth expansion**. (see e.g. Abramowitz and Stegun (1972)) Take an extended Taylor expansion about zero for the cumulant generating function:

$$k(s) = s\mu + \sigma^2 s^2/2 + k'''(0)s^3/6 + k^{iv}(0)s^4/24 + k^{v}(0)s^5/120 + \cdots .$$

Now standardize by $z = \frac{x-\mu}{\sigma}$ to get $k_z(s) = s^2/2 + k_3 s^3/6 + k_4 s^4/24 + k_5 s^5/120 + \cdots$ where those standardized cumulants are $k_m = k^{(m)}(0)/\sigma^m$. Now exponentiate, expanding the correction terms, to get

$$m_z(s) = e^{s^2/2}\left[1 + k_3 s^3/6 + \left(k_3^2 s^6/72 + k_4 s^4/24\right) + \left(k_3^3 s^9/1296 + k_3 k_4 s^7/144 + k_5 s^5/120\right) + \cdots\right].$$

The terms have been grouped by orders of $n^{-1/2}$ where n is the sample size. Invert the Laplace transform to get

$$f(z) = \phi(z)\left[1 + k_3 H_3/6 + \left(k_3^2 H_6/72 + k_4 H_4/24\right) + \left(k_3^3 H_9/1296 + k_3 k_4 H_7/144 + k_5 H_5/120\right) + \cdots\right]$$

where the Hs are Hermite polynomials in z. Using the standard fact that $[H_k\phi]' = -H_{k+1}\phi$, we integrate the expression above to get a tail area expansion

$$P(X > x) = \Phi(z) + \phi(z)\left[k_3 H_2/6 + \left(k_3^2 H_5/72 + k_4 H_3/24\right) + \left(k_3^3 H_8/1296 + k_3 k_4 H_6/144 + k_5 H_4/120\right) + \cdots\right]$$

where Φ is the standard normal tail probability.

Daniels (1987) pointed out that one could construct a *tilted* Edgeworth expansion to achieve the same end for the one-parameter Wald statistic. Starting with a Taylor's expansion of the cumulant generating function about the value t determined by $k'(t) = x$; and standardizing by $z = t\sqrt{k''(t)}$, proceed as above to get

$$f(w) = e^{k(0) - xt + k''(0)t^2/2}\phi(w)\left[1 + k_3^* H_3(w - z)/6 + \left(k_3^{*2} H_6(w - z)/72 + k_4^* H_4(w - z)/24\right) + \cdots\right]$$

where $k_m^* = k^{(m)}(t)/k''(t)^{m/2}$ are tilted cumulants. A central insight of Daniels (1954) was that when we use this to evaluate $f(z)$, the first correction term drops out, leading to an approximation of order $O(n^{-1})$ to the density of x. But the tail area can be integrated directly to get

$$\mathrm{P}(X > x) = e^{\kappa(0) - xt + z^2/2}\left\{\Phi(z) - \frac{\kappa_3(t)}{6}L_3(z)\right.$$
$$\left. + \left[\frac{\kappa_4(t)}{24}L_4(z) + \frac{\kappa_3(t)^2}{72}L_6(z)\right] - \left[\frac{\kappa_5(t)}{120}L_5(z) + \frac{\kappa_4(t)\kappa_3(t)}{144}L_7(z) + \frac{\kappa_3(t)^3}{1296}L_9(z)\right] + \cdots\right\}$$

where the L functions, which play here the same role as the Hermite functions are defined in Terrell (2001).

VII. A Two-Point Edgeworth Expansion.

We will now develop an Edgeworth-like expansion for the gradience. Its key feature is that it is neither based on an expansion of k about zero nor about t; it is, rather, symmetric in the null and observed values. We will therefore develop a Taylor-like expansion that is *almost* symmetric in 0 and t. We will begin with an expansion about 0:

$$k(s) = sk'(0) + \frac{s^2}{2}k''(0) + \frac{s^3}{6}k'''(0) + \frac{s^4}{24}k^{iv}(0) + \frac{s^5}{120}k^{v}(0) + \cdots .$$

To extend it to the two points, we will use the Euler-Maclaurin formula

$$g(t) - g(0) = \frac{t}{2}\left[g'(t) + g'(0)\right] - \frac{t^2}{12}\left[g''(t) - g''(0)\right] + \frac{t^4}{720}\left[g^{iv}(t) - g^{iv}(0)\right] - \cdots$$

(Whittaker and Watson 1927). Solve for $g'(0)$ to get

$$g'(0) = \frac{g(t) - g(0)}{t} - \frac{1}{2}\left[g'(t) - g'(0)\right] + \frac{t}{12}\left[g''(t) - g''(0)\right] - \frac{t^3}{720}\left[g^{iv}(t) - g^{iv}(0)\right] + \cdots .$$

Now substitute for the derivatives beginning with k'':

$$k(s) = sk'(0) + \frac{s^2}{2}\left[\frac{k'(t) - k'(0)}{t} - \frac{t}{2}\frac{k''(t) - k''(0)}{t} + \frac{t^2}{12}\frac{k'''(t) - k'''(0)}{t} - \cdots\right]$$
$$+ \frac{s^3}{6}\left[\frac{k''(t) - k''(0)}{t} - \frac{t}{2}\frac{k'''(t) - k'''(0)}{t} + \frac{t^2}{12}\frac{k^{iv}(t) - k^{iv}(0)}{t} - \cdots\right]$$
$$+ \frac{s^4}{24}\left[\frac{k'''(t) - k'''(0)}{t} - \frac{t}{2}\frac{k^{iv}(t) - k^{iv}(0)}{t} + \cdots\right] + \frac{s^5}{120}\left[\frac{k^{iv}(t) - k^{iv}(0)}{t} + \cdots\right] + \cdots .$$

This suggests a sort of mean mth cumulant, $k_m = \dfrac{k^{(m-1)}(t) - k^{(m-1)}(0)}{t}$, which by the mean value theorem is a tilted mth cumulant, tilted to a point intermediate between 0 and t. Grouping, we get

$$k(s) = sk'(0) + \frac{s^2}{2}k_2 + \frac{k_3}{12}\left(2s^3 - 3s^2t\right) + \frac{k_4}{24}\left(s^4 - 2s^3t + s^2t^2\right) + \frac{k_5}{720}\left(6s^5 - 15s^4t + 10s^3t^2\right) + \cdots$$

Notice that when we evaluate this expansion at t, we get

$$k(t) = \frac{t}{2}\left[k'(t) + k'(0)\right] - \frac{t^2}{12}\left[k''(t) - k''(0)\right] + \frac{t^4}{720}\left[k^{iv}(t) - k^{iv}(0)\right] - \cdots .$$

the asymptotic Euler-Maclaurin expansion. In general, then, a partial expansion will not interpolate $k(t)$ when it is constructed to interpolate $k(0) = 0$. If we now carry out a parallel derivation using a Taylor expansion about t, by symmetry we will achieve the same expansion for $k(s)$, but now interpolating $k(t)$ by introducing a constant term (e.g. for our expansion through the fifth power)

$$c_4(t) = k(t) - \frac{t}{2}\left[k'(t) + k'(0)\right] + \frac{t^2}{12}\left[k''(t) - k''(0)\right] - \frac{t^4}{720}\left[k^{iv}(t) - k^{iv}(0)\right] .$$

The constant is thus the error in an Euler-Maclaurin expansion. This is the sense in which our expansion is almost symmetric in t.

Using the standardization $z = \operatorname{sgn}(t)\sqrt{t(x-\mu)}$ for the gradient statistic, we get a standardized cumulant generating function

$$k_z(s) = \frac{s^2}{2} + \frac{k_3^*}{12}\left(2s^3 - 3s^2 z\right) + \frac{k_4^*}{24}\left(s^4 - 2s^3 z + s^2 z^2\right) + \frac{k_5^*}{720}\left(6s^5 - 15s^4 z + 10s^3 z^2\right) + \cdots$$

where those standardized mean cumulants are of course $k_m^* = \frac{k_m}{k_2^m}$. Our moment generating function may then be expanded

$$m_z(s) = e^{s^2/2}\left\{1 + \frac{k_3^*}{12}\left(2s^3 - 3s^2 z\right) + \left[\frac{k_4^*}{24}\left(s^4 - 2s^3 z + s^2 z^2\right) + \frac{k_3^{*2}}{288}\left(4s^6 - 12s^5 z + 9s^4 z^2\right)\right] + \right.$$

$$\left.\left[\frac{k_5^*}{720}\left(6s^5 - 15s^4 z + 10s^3 z^2\right) + \frac{k_3^* k_4^*}{288}\left(2s^7 - 7s^6 z + 8s^5 z^2 - 3s^4 z^3\right) + \frac{k_3^{*3}}{10368}\left(8s^9 - 36s^8 z + 54s^7 z^2 - 27s^6 z^3\right)\right] + \cdots\right\}$$

Inverting the Laplace transform and integrating, we get the tail-probability estimate

$$P(X \geq x) = \Phi(z) + \phi(z)\left\{-\frac{k_3^*}{12}\left(z^2 + 2\right) + \left[-\frac{k_4^*}{24} z + \frac{k_3^{*2}}{288}\left(z^5 + 5z^3 + 24z\right)\right] + \right.$$

$$\left.\left[\frac{k_5^*}{720}\left(z^4 - z^2 + 18\right) + \frac{k_3^* k_4^*}{288}\left(z^4 + 9z^2 - 30\right) - \frac{k_3^{*3}}{10368}\left(z^8 + 8z^6 + 75z^4 + 390z^2 - 840\right)\right] + \cdots\right\} .$$

Example 6: We continue the Gamma example with first and second order corrections to the gradience:

x	exact P	1st order	2nd order
8	.00823	.00557	.00778
12	.1556	.1529	.15585
20	.1565	.1542	.15637
24	.0344	.0304	.03443
28	.00543	.00224	.00598

These become very accurate over most of the range of the distribution, then deteriorate rapidly in the far tails.

We may tilt our cumulant generating function to interpolate $k(t)$ as above. For the first and second order corrections, the tilting factor is the same: $e^{c_2(0)}$ multiplies each tail

probability, where $c_2(t) = k(t) - \frac{t}{2}\left[k'(t) + k'(0)\right] + \frac{t^2}{12}\left[k''(t) - k''(0)\right]$. In the example above, the reader may check that the tilt is small compared to the error in the approximation. This is more evidence for our assertion that our method is almost symmetric in the null and empirical values of the parameter.

VIII. Bibliography.

Abramowitz, M. and Stegun, I. (1972). **A Handbook of Mathematical Functions.** Dover: New York.

Bickel, P. and Doksum, K, (2001). **Mathematical Statistics, Volume I, Second Edition.** Prentice Hall: New Jersey.

Daniels, H. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* 25 pp. 631-650.

Daniels, H. (1987). Tail probability approximations. *International Statistical Review* 55 pp. 37-48.

Dharmadhikari, S., and Joag-dev, K. (1988). **Unimodality, Convexity, and Applications.** Academic Press. p. 38.

Manly, B. (1991). **Randomization and Monte Carlo Methods in Biology.** Chapman and Hall: New York.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society* 44 pp. 40-57.

Severini, T. A. (2000). **Likelihood Methods in Statistics.** Oxford University Press.

Terrell, G. R. (1998). Gibbs sampling for estimation of linear models for ranks. **Proceedings of the Symposium on the Interface: Computing Science and Statistics**. 1998, Houston, Texas, pp. 267-270.

Terrell. G. R. (2001). Tail probabilities by density extrapolation. *Proceedings of the Statistical Computing Section, Joint Statistical Meetings*. Atlanta. to appear.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* 54 pp. 426-482.

Whittaker, E. T., and Watson, G. N. (1927). **A Course of Modern Analysis**. Cambridge Mathematical Library. pp. 127-128.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* 9 pp. 60-62.

Ye, K.-Y. (2001) Personal communication.