Exact p-values in Goodness of fit

Federico J. O’Reilly

IIMAS-UNAM, México

XXVI European Meeting of Statisticians, Torun, Poland. July 2006

Key Words: randomization; goodness-of-fit; conditional Monte Carlo.

ABSTRACT

Recent, separate articles are analyzed in their effort to provide exact evaluation of p-values in the context of testing fit to a distribution in the presence of parameters beyond; the well known location-scale cases. An interesting application of a result by J. Durbin, (1961) is made. Resulting methods for different non location scale families as inverse-Gaussian, Gamma and others, are discussed.
1. INTRODUCTION

In simple goodness-of-fit, tests were initially based on their asymptotics. Finite\(-n\) distributions (known in some cases but in general found by simulations), found close to their asymptotic distributions. Not general rule when problem is composite; approximation varies with null family; on top, there are families for which asymptotics depend on values of parameters.

In case of shape parameters and most discrete families. It has been a practice to enter tables, when the sample size \(n\) is large, estimating parameters (see Lockhart and Stephens, 1985 for gamma, O’Reilly and Rueda, 1992 for inverse-Gaussian and Spinelli and Stephens, 1997 for Poisson).
Early efforts to obtain same limit (continuous case) when parameters unknown, as if known, succeeded with various techniques. Durbin’s half sample (see Durbin, 1973, page 59, and closely connected Rao’s expression) are among them; in a different way, the random substitution in Durbin (1961). The first two rely on a randomly selected half-sample and the third on a randomly selected value for the sufficient statistic, fixing the parameter value of its distribution.

These methods, besides poor power, had inconvenience of not yielding invariant results when used by two statisticians. Based on an exogenous randomization; violate Likelihood Principle. Coloquially refer to efficient methods that use the “simple” goodness-of-fit distributions of tests in “composite” problems, as a “statistician’s dream”.
In Section 2 Durbin’s theorem is recalled.

In Section 3, modern standpoint using computer intensive procedures is reviewed. Parametric bootstrap to obtain (approximate) p-values, and on-line algorithms to compute exact p-values. Simulation procedures, have now a big advantage and apply to discrete problems also. In this same section results in Lindqvist and Taraldsen (2005), Engen and Lillegård (1997), and in O’Reilly and Gracia-Medrano (2006) may be connected Durbin’s result.

Finally in Section 4 comment on the changing paradigm in goodness-of-fit, is made. What now may seem unjustified, was perfectly justified with the available resources, when it was proposed.

2. COMPOSITE TO SIMPLE

Recall that in Durbin (1961), a procedure is given to change the original $x_1, \ldots, x_n$ by what is termed there, a random substitution.
The result shown with all details for normal with $\mu$ and $\sigma^2$. With one realization from the joint distribution for $\bar{X}$ and $S$ as if they corresponded to a $N(0,1)$ sample, a realization, denoted $\bar{x}'$ and $s'$, each $x_i$ is transformed into an $x'_i$ with the property that these new $x'_i$, $i = 1, \ldots, n$ behave as an independent sample from a $N(0,1)$ distribution. Thus, transforming the original problem into one of testing a standard normal; a simple hypothesis. The definition of the $x'_i$:

$$x'_i = (s'/s)(x_i - \bar{x}) + \bar{x}'.$$  

But as mentioned in the article:

"The price that has to be paid for the elimination of nuisance parameters by this method is that an element of randomization is introduced in the analysis of the data. The device can be objected on the ground that it permits different investigators to draw different conclusions from the same set of data".
The author clearly mentions, the generality of his procedure, which we reproduce next.

**Theorem (Durbin, 1961)**

Assume $X_1, \ldots, X_n$ have joint CDF

$$H(\cdot, \ldots, \cdot; \theta)$$

with $\theta$ unknown and for which $T_1$ is sufficient with CDF $G(t_1, \theta)$. Also assume there is a statistic $T_2$ whose distribution does not depend on $\theta$, and that there is a one to one relationship between $(T_1, T_2)$ and $X_1, \ldots, X_n$.

If $T_1$ independent of $T_2$, and if $x_1, \ldots, x_n$ is an observation of $X_1, \ldots, X_n$ with $T_2(x_1, \ldots, x_n) = t_2$, then for $t'_1$, an independent realization from $G(\cdot, \theta_0)$, with $\theta_0$ any fixed value for the parameter,

the “realization” $x'_1, \ldots, x'_n$ corresponding to the pair $(t'_1, t_2)$,

behaves as an observation from the CDF

$$H(\cdot, \ldots, \cdot; \theta_0).$$
Conditions of theorem satisfied in all location-scale families where dimension of minimal sufficient $T_1$ is fixed (exponential families, where $T_1$ is equivariant; $T_2$ invariant, taken as the vector of residues so independence follows). Inverse-Gaussian may be shown to meet conditions of the theorem with slight variation that does not affect; namely that one can establish a two-to-one correspondence between the $x$-sample(s) and $(t_1, t_2)$ with samples differing only in order in, say $x_1$ and $x_2$. For details, see Gracia-Medrano and O'Reilly (2004).

Power of randomized procedures, like the one mentioned or the one using half sample; or others using external randomization, has been reported repeatedly very low.
Recently in Nguyen and Ding (2003) an apparent different proposal to deal the two parameter inverse-Gaussian. Proposal connected to earlier work reducing the case of unknown parameters to that of known parameters by a transformation, based on sequential application of Rao-Blackwell CDF estimate (O’Reilly and Quesenberry, 1973). Transformation maps original \( x \)-sample to \( n - 2 \) independent uniform random variables in \((0,1)\); in case of inverse-Gaussian, and into \( n - k \) \( U(0,1) \) independent r.v.’s in case of a \( k \) dimensional minimal sufficient statistic.

The “catch”, of no apparent randomization, is that if handed the original observations, already in ascending order, to apply the transformation, a random permutation of the sample is required, as to have, as data, \( iid \) observations.

In Johnson (2004) a random realization \( \tilde{\theta} \) from the posterior is taken as the fixed value of parameter in building a chi-square test (and same limit as in the simple case is obtained).
3. COMPUTER INTENSIVE METHODS

In location-scale, it has long been known EDF tests are independent of parameters. One simply simulates many samples of size $n$ from parent having fixed arbitrary values for the parameter and with those many samples, compute each time the EDF test. These values form a simulated sample from the distribution of the EDF test.

This done with a simple algorithm. Sample fed into a computer program, that generates, say 10,000 samples, computing the corresponding values for the test, finding proportion of simulated values that exceed the value of the statistic computed with the initially observed sample. This is the “exact” $p$-value.
A “boom” in applications of **parametric bootstrap** in goodness-of-fit, in finding approximate p-values. One estimates parameters, substitute in parent distribution and then simulate many samples of size n, as if estimates were parameter values. With each bootstrap sample, test statistic is evaluated and, as in case of location-scale, observed value of the statistic compared with the simulated values. Procedure yields an “approximate p-value” (proportion of simulated values exceeding the observed value).

Many papers deal with this procedure. In each, showing that **asymptotically**, one is evaluating a correct p-value. Many authors working in this area but this is not a survey note, we refer to Henze and Klar (2002) and Gurtler and Henze (2000) for a continuous and discrete case. If family is location-scale, method exact.
A more recent effort to compute the p-value of an EDF test, from its conditional distribution given the minimal sufficient $T$ has been used. May be applied in discrete cases. Introduced in O’Reilly and Gracia-Medrano (2006) with a simulation using the Rao-Blackwell estimate of the CDF. Made explicit in inverse-Gaussian.

For cases where there is not known Rao-Blackwell expression, as in the $\text{gamma}(\alpha, \beta)$ simulation has been used mantaining what is underlying; i.e. the conditional distribution of the sample, requiring a more elaborate simulation Gibbs (Lockhart et.al., 2006).
Outside usual stream of goodness-of-fit, we learned of Lindqvist and Taraldsen’s (2005) paper, where idea is in finding conditional expectations given a sufficient statistic $T$, with a “weighted” Monte Carlo. Authors follow and extend Engen and Lillegård’s (1997) results to produce conditionally independent samples given $T$.

Since setting in these articles as in previous, aimed to compute exact p-values, are related to Durbin’s theorem mentioned in Section 2, we stated a corollary, now intended to the generation of conditionally independent samples.
Corollary (Durbin, 1961)

Assume $X_1, \ldots, X_n$ have joint CDF

$$H(\cdot, \ldots, \cdot; \theta)$$

with $\theta$ unknown and for which $T_1$ is sufficient with CDF $G(t_1, \theta)$. Also assume there is a statistic $T_2$ whose distribution does not depend on $\theta$, and that there is a one to one relationship between $(T_1, T_2)$ and $X_1, \ldots, X_n$.

If $T_1$ independent of $T_2$, and if $x_1, \ldots, x_n$ is an observation of $X_1, \ldots, X_n$ with $T_1(x_1, \ldots, x_n) = t_1$, then for $x'_1, \ldots, x'_n$, an independent realization obtained from the parent but fixing $\theta_0$ arbitrarily, and if $t'_2$ is the corresponding value for $T_2$ from the $x'$-sample,

the “realization” $x^*_1, \ldots, x^*_n$ corresponding to the pair $(t_1, t'_2)$,

behaves as an observation from the CDF

$$H(\cdot, \ldots, \cdot; \theta).$$

This result, is not to be used once, but many times generating “look alike samples”.
The proof of result relays on the 1:1 correspondence of the \(x\)-sample and \((t_1, t_2)\). Dependance on \(\theta\) is through \(t_1\). From the \(x'\)-sample, after identifying its \(t'_2\), the corresponding \(t'_1\), which is the part related to \(\theta_0\), is **changed to the original** \(t_1\), so dependance is on \(\theta\), as original sample. Corresponding to new pair \((t_1, t'_2)\), namely \(x^*_1, \ldots, x^*_n\) has the property of behaving like a sample from the parent with parameter \(\theta\) yielding \(t_1\).

With Corollary, straight-forward to show that in group models result follows immediately. In Lindqvist and Taraldsen’s paper (2005), the method of Engen and Lillegård (1997), to produce an \(x^*\)-sample is shown to be correct if the model **admits a “pivotal”** in the classical sense, a sufficient condition found by them. With \(T_1\) equivariant sufficient and \(T_2\) the invariant residues, condition of pivotal corresponds to family being **a group model**.
In non-group models, within the Natural Exponential Family (NEF), if conditional distribution of as many as \((n - k)\) terms is non-degenerate, result may be applied. This is the case of the interesting truncated exponential family introduced in Lindqvist et.al., (2003) where not being a group model (not having a pivotal), used as counter example for Engen and Lillegård’s method; not yielding correct \(x^*\)-samples. Use of Gibbs or the Rao-Blackwell expression allows generation of correct “look-alike” samples.

Corollary may be seen to be absolutely up to date, referring to problems posed now; like generation of conditionally iid samples.
We want to mention that **Rosenblatt’s multivariate transformation** (Rosenblatt, 1952), which was an extension of the probability integral transformation (PIT) to a continuous multivariate CDF. When applied in a conditional setting, allowed the construction, under certain circumstances, of $n - k$ iid uniforms out of a sample of $n$ iid observations from a parent with unknown $k$-dimensional $\theta$, allowing a substitution of composite goodness-of-fit problems for a simple one (see O’Reilly and Quesenberry, 1973).

In the same way as the inverse of the PIT would allow generation of an observation of a random variable with arbitrary distribution function, **Rosenblatt’s transformation applied inversely** and within a conditional setting, is in the heart of the procedure to obtain conditional samples; the look-alike samples.
4. COMMENTS

In the remarkable paper by Durbin, there is something not directly relevant to the present discussion, but that gives us an idea of what were the real worries then, regarding computing. In his Section 6, he suggests the elimination, by systematic subsampling of part of the original ordered sample. He explains:

"Most of the work to carry out the goodness of fit tests described in Section 3 is likely to occur in making the probability integral transformation of the $x's$ to the $u's$; indeed when the sample is large the amount of labour required can be prohibitively great".

In the present days thinking of not using all available information would simply not be right. As an interesting note we observe that taking a subsample from the previously ordered sample (say each $s$-th term) as suggested by Durbin, would certainly yield less power than using all sample but would not violate the likelihood principle.
We believe some seminal ideas were placed properly in statistical thought many years ago, and some new discoveries are somehow related or are updated views of these valuable results. I certainly feel very lucky for having read Professor Durbin’s article.

I acknowledge collaborations with:

José María González B., Leticia Gracia-Medrano and Raúl Rueda from UNAM; and Richard Lockhart and Michael Stephens from Simon Fraser University, Canada.

Thanks

e-mail federico@sigma.iimas.unam.mx

(Next four slides are References, then one output)


Inverse-Gaussian fit.

The sample itself is given below (n=25):

1.01000 1.11000 1.13000 1.15000 1.16000 1.17000
1.17000 1.20000 1.52000 1.54000 1.54000 1.57000
1.64000 1.73000 1.79000 2.09000 2.09000 2.57000
2.75000 2.93000 3.19000 3.54000 3.57000 5.11000
5.62000

(Precipitation Jug Bridge Maryland, O'Reilly-Rueda 1992, also Henze and Klar 2002)

Uses R-B PIT, estimates of mu and lambda:
2.155600 8.081986

A squared and W: 7.672372E-001 1.224187E-001

number of *samples is 10,000

Time in seconds 36.5200

the conditional p-values are : .0538 .0666
Output from program **Gamma** that generates *samples. With same data set as used in inverse-Gaussian:

estimates of alfa and beta: 3.9885 .5405

100 burned samples, 20 inter-sample burned samples and 10,000 final look-alikes.

time is 13:19:14:85

number of total samples generated is 210,080

Observed Anderson Darling is : 1.006337

The p-value sits between : .011 .011

time is 13:45:11:00 so about 16 min Observe that the 1/2 minute for 10,000 in inverse-Gaussian would have been about 10 min to get the roughly 200,000 total samples here, but simulator in gamma is certainly slower (iterates over sample). Lognormal fit also done, gave p-values around 0.03.