Significance testing without truth

William Perkins, Mark Tygert, and Rachel Ward

October 2, 2012

Abstract

A popular approach to significance testing proposes to decide whether the given hypothesized statistical model is likely to be true (or false). Statistical decision theory provides a basis for this approach by requiring every significance test to make a decision about the truth of the hypothesis/model under consideration. Unfortunately, many interesting and useful models are obviously false (that is, not exactly true) even before considering any data. Fortunately, in practice a significance test need only gauge the consistency (or inconsistency) of the observed data with the assumed hypothesis/model — without enquiring as to whether the assumption is likely to be true (or false), or whether some alternative is likely to be true (or false). In this practical formulation, a significance test rejects a hypothesis/model only if the observed data is highly improbable when calculating the probability while assuming the hypothesis being tested; the significance test only gauges whether the observed data likely invalidates the assumed hypothesis, and cannot decide that the assumption — however unmistakably false — is likely to be false a priori, without any data.

Essentially, all models are wrong, but some are useful. — G. E. P. Box

Contents

1   Introduction

As pointed out in the above quotation of G. E. P. Box, many interesting models are false (that is, not exactly true), yet are useful nonetheless. Significance testing helps measure the usefulness of a model. Testing the validity of using a model for virtually any purpose requires knowing whether observed discrepancies are due to inaccuracies or inadequacies in the model or (on the contrary) could be due to chance arising from necessarily finite sample sizes. Significance tests gauge whether the discrepancy between the model and the observed data is larger than expected random fluctuations; significance tests gauge the size of the unavoidable random fluctuations.

A traditional approach, along with its modern formulation in statistical decision theory, tries to decide whether a hypothesized model is likely to be true (or false). However, in many practical circumstances, a significance test need only gauge the consistency (or inconsistency) of the observed data with the assumed hypothesis/model — without ever enquiring as to whether the assumption is likely to be true (or false), or whether some alternative is likely to be true (or false). In this practical formulation, a significance test rejects a hypothesis/model only if the observed data is highly improbable when calculating the probability while assuming the hypothesis being tested. Whether or not the assumption could be exactly true in reality is irrelevant.

An illustrative example may help clarify. When testing the goodness of fit for the Poisson regression where the distribution of Y given x is the Poisson distribution of mean $\exp(\theta^{(0)} + \theta^{(1)}x + \theta^{(2)}x^2 + \theta^{(3)}x^3)$, the conventional Neyman-Pearson null hypothesis is

$$H_0^{\mathrm{NP}} : \text{there exist real numbers } \theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \theta^{(3)} \text{ such that } y_1, y_2, \ldots, y_n \text{ are independent}$$
$$\text{draws from the Poisson distributions with means } \mu_1, \mu_2, \ldots, \mu_n, \text{ respectively,} \quad (1)$$

where
$$\ln(\mu_k) = \theta^{(0)} + \theta^{(1)}x_k + \theta^{(2)}(x_k)^2 + \theta^{(3)}(x_k)^3 \quad (2)$$

for $k = 1, 2, \ldots, n$, and the observations (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) are ordered pairs of scalars (real numbers paired with nonnegative integers). A related but perhaps simpler null hypothesis is

$$H_0 : y_1, y_2, \ldots, y_n \text{ are independent draws from the Poisson distributions}$$
$$\text{with means } \hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n, \text{ respectively,} \quad (3)$$

where
$$\ln(\hat{\mu}_k) = \hat{\theta}^{(0)} + \hat{\theta}^{(1)}x_k + \hat{\theta}^{(2)}(x_k)^2 + \hat{\theta}^{(3)}(x_k)^3 \quad (4)$$

for $k = 1, 2, \ldots, n$, with $\hat{\theta}$ being a maximum-likelihood estimate. Needless to say, even if the observed data really does arise from Poisson distributions whose means are exponentials of a cubic polynomial, the particular values $\hat{\theta}^{(0)}$, $\hat{\theta}^{(1)}$, $\hat{\theta}^{(2)}$, $\hat{\theta}^{(3)}$ of the parameters of the fitted polynomial will almost surely not be exactly equal to the true values. Even though the estimated values of the parameters may not be exactly correct, it still makes good sense to enquire as to whether the fitted cubic polynomial is consistent with the data up to random fluctuations inherent in using a finite amount of observed data.

In fact, since subsequent use of the model usually involves the particular fitted polynomial — whose specification includes the observed parameter estimates — analyzing the model including the estimated values of the parameters makes more sense than trying to decide whether the data really did come from Poisson distributions whose means are exponentials of some unspecified cubic polynomial. For instance, any plot of the fit (such as a plot of the means of the Poisson distributions) must use the estimated values of the parameters, and any statistical interpretation of the plot should also depend explicitly on the estimates; a significance test can gauge the consistency of the plotted fit with the observed data, without ever asking whether the plotted fit is the truth (it is almost surely not identical to the underlying reality) and without making some decision about an abstract family of polynomials which may or may not include both the plotted fit and the underlying reality.

A popular measure of divergence from the null hypothesis is the log–likelihood-ratio

$$g^2 = 2 \sum_{k=1}^{n} y_k \ln(y_k/\hat{\mu}_k). \tag{5}$$

A P-value (see, for example, Section 3 below) quantifies whether this divergence is larger than expected from random fluctuations inherent in using only n data points. It is not obvious how to calculate an exact P-value for H_0^{NP} from (1) and (2), which refers to cubic polynomials with undetermined coefficients. In contrast, H_0 from (3) and (4) refers explicitly to the particular fitted value $\hat{\theta}$; H_0 concerns the particular fit displayed in a plot, and is natural for the statistical interpretation of such a plot.

Thus, when calculating significance, the assumed model should include the particular values of any parameters estimated from the observed data. Such parameters are known as "nuisance" parameters. As illustrated with H_0 from (3) and (4), the assumed hypothesis will be "simple" in the Neyman-Pearson sense, but will depend on the observed values of the parameters — that is, the hypothesis will be "data-dependent"; the hypothesis will be "random." Including the particular values of the parameters estimated from the observed data replaces the "composite" hypothesis of the conventional Neyman-Pearson formulation with a "simple" data-dependent hypothesis. As discussed in Section 4 below, fully conditional tests also incorporate the observed values of the parameters, but make the extra assumption that all possible realizations of the experiment — observed or hypothetical — generate the same observed values of the parameters. The device of a "simple data-dependent hypothesis" such as H_0 incorporates the observed values explicitly without the extra assumption.

For most purposes, a parameterized model is not really operational — that is, suitable for making precise predictions — until its specification is completed via the inclusion of estimates for any nuisance parameters. The results of the significance tests considered below depend on the quality of both the models and the parameter estimators. However, the results are relatively insensitive to the particular observed realizations of the parameter estimators (that is, to the parameter estimates) unless specifically designed to quantify the quality of the parameter estimates. To quantify the quality of the parameter estimates, we recommend testing separately the goodness of fit of the parameter estimates, using confidence intervals, confidence distributions, parametric bootstrapping, or significance tests within parametric models, whose statistical power is focused against alternatives within the parametric family constituting the model (for further discussion of the latter, see Section 5 below).

The remainder of the present article has the following structure: Section 2 very briefly discusses Bayesian-frequentist hybrids, referring for details to the definitive work of Gelman (2003). Section 3 defines P-values — also known as "attained significance levels" — which quantify the consistency of the observed data with the assumed models. Section 4 details several approaches to testing the goodness of fit for distributional profile. Section 5 discusses testing the goodness of fit for various properties beyond just distributional profile.

Cox (2006) details many advantages of interpreting significance as gauging the consistency of an assumption/hypothesis with observed data, rather than as making decisions about the actual truth of the assumption. However, significance testing is meaningless without any observations, unlike purely Bayesian methods, which can produce results without any data, courtesy of the prior (the prior is the statistician's system of a priori beliefs, accumulated from prior experience, law, morality, religion, etc., without reference to the observed data). Significance tests are deficient in this respect. Those interested in what is to be considered true in reality and in making decisions more generally should use Bayesian and sequential (including multilevel) procedures. Significance testing simply gauges the consistency of models with observed data; generally significance testing alone cannot handle the truth.

2 Bayesian versus frequentist

Traditionally, significance testing is frequentist. However, there exist Bayesian-frequentist hybrids known as "Bayesian P-values"; Gelman (2003) sets forth a particularly appealing formulation. Bayesian P-values test the consistency of the observed data with the model *used together with a prior* for nuisance parameters. In contrast, the P-values discussed in the present paper test the consistency of the observed data with the model *used together with a parameter estimator*. In the Bayesian formulation, a P-value depends explicitly on the choice of prior; in the formulation of the present paper, a P-value depends explicitly on the choice of parameter estimator. Thus, when there are nuisance parameters, the two types of P-values test slightly different hypotheses and provide slightly different information; each type is ideal for its own set-up. Of course, if there are no nuisance parameters, then Bayesian P-values and the P-values discussed below are the same.

3 P-values

A P-value for a hypothesis H_0 is a statistic such that, if the P-value is very small, then we can be confident that the observed data is inconsistent with assuming H_0. The P-value associated with a measure of divergence from a hypothesis H_0 is the probability that $D \geq d$, where d is the divergence between the observed and the expected (with the expectation following H_0 for the observations), and D is the divergence between the simulated and the expected (with the expectation following H_0 for the simulations, and with the simulations performed assuming H_0). When taking the probability that $D \geq d$, we view D as a random variable, while viewing d as fixed, not random. For example, when testing the goodness of fit for the model of i.i.d. draws from a probability distribution $p_0(\theta)$, where θ is a nuisance parameter that must be estimated from the data, that is, from observations $x_1, x_2, \ldots, x_n$,

we use the null hypothesis

$$H_0 : x_1, x_2, \ldots, x_n \text{ are i.i.d. draws from } p_0(\hat{\theta}), \text{where } \hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n). \qquad (6)$$

The P-value for H_0 associated with a divergence δ is the probability that $D \geq d$, where

- $d = \delta(\hat{p}, p_0(\hat{\theta}))$,

- $\hat{p}$ is the empirical distribution of $x_1, x_2, \ldots, x_n$,

- $\hat{\theta}$ is the parameter estimate obtained from the observed draws $x_1, x_2, \ldots, x_n$,

- $D = \delta(\hat{P}, p_0(\hat{\Theta}))$,

- $\hat{P}$ is the empirical distribution of i.i.d. draws $X_1, X_2, \ldots, X_n$ from $p_0(\hat{\theta})$, and

- $\hat{\Theta}$ is the parameter estimate obtained from the simulated draws $X_1, X_2, \ldots, X_n$.

If the P-value is very small, then we can be confident that the observed data is inconsistent with assuming H_0. Examples of divergences include χ^2 (for categorical data) and the maximum absolute difference between cumulative distribution functions (for real-valued data).

Remark 3.1. To compute the P-value assessing the consistency of the experimental data with assuming H_0, we can use Monte-Carlo simulations (very similar to those used by Clauset et al. (2009)). First, we estimate the parameter θ from the n given experimental draws, obtaining $\hat{\theta}$, and calculate the divergence between the empirical distribution and $p_0(\hat{\theta})$. We then run many simulations. To conduct a single simulation, we perform the following three-step procedure:

1. we generate n i.i.d. draws according to the model distribution $p_0(\hat{\theta})$, where $\hat{\theta}$ is the estimate calculated from the experimental data,

2. we estimate the parameter θ from the data generated in Step 1, obtaining a new estimate $\tilde{\theta}$, and

3. we calculate the divergence between the empirical distribution of the data generated in Step 1 and $p_0(\tilde{\theta})$, where $\tilde{\theta}$ is the estimate calculated in Step 2 from the data generated in Step 1.

After conducting many such simulations, we may estimate the P-value for assuming H_0 as the fraction of the divergences calculated in Step 3 that are greater than or equal to the divergence calculated from the empirical data. The accuracy of the estimated P-value is inversely proportional to the square root of the number of simulations conducted; for details, see Remark 3.2 below.

Remark 3.2. The standard error of the estimate from Remark 3.1 for an exact P-value P is $\sqrt{P(1-P)/\ell}$, where ℓ is the number of Monte-Carlo simulations conducted to produce the estimate. Indeed, each simulation has probability P of producing a divergence that is greater than or equal to the divergence corresponding to an exact P-value of P. Since the simulations are all independent, the number of the ℓ simulations that produce divergences

greater than or equal to that corresponding to P-value P follows the binomial distribution with ℓ trials and probability P of success in each trial. The standard deviation of the number of simulations whose divergences are greater than or equal to that corresponding to P-value P is therefore $\sqrt{\ell P(1-P)}$, and so the standard deviation of the *fraction* of the simulations producing such divergences is $\sqrt{P(1-P)/\ell}$. Of course, the fraction itself is the Monte-Carlo estimate of the exact P-value (we use this estimate in place of the unknown P when calculating the standard error $\sqrt{P(1-P)/\ell}$).

4 Goodness of fit for distributional profile

Given observations x_1, x_2, ..., x_n, we can test the goodness of fit for the model of i.i.d. draws from a probability distribution $p_0(\theta)$, where θ is a nuisance parameter, via the null hypothesis

$$H_0 : x_1, x_2, \ldots, x_n \text{ are i.i.d. draws from } p_0(\hat{\theta})$$
$$\text{for the particular observed value of } \hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n). \quad (7)$$

The Neyman-Pearson formulation considers instead the null hypothesis

$$H_0^{\mathrm{NP}} : \text{there exists a value of } \theta \text{ such that } x_1, x_2, \ldots, x_n \text{ are i.i.d. draws from } p_0(\theta). \quad (8)$$

The fully conditional null hypothesis is

$$H_0^{\mathrm{FC}} : x_1, x_2, \ldots, x_n \text{ are i.i.d. draws from } p_0(\hat{\theta})$$
$$\text{and } \hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n) \text{ takes the same value in all possible realizations.} \quad (9)$$

That is, whereas H_0 supposes that the particular observed realization of the experiment happened to produce a parameter estimate $\hat{\theta}$ that is consistent with having drawn the data from $p_0(\hat{\theta})$, H_0^{FC} assumes that every possible realization of the experiment — observed or hypothetical — produces exactly the same parameter estimate. Few experimental apparatus constrain the parameter estimate to always take the same (a priori unknown) value during repetitions of the experiment, as H_0^{FC} assumes. Assuming H_0^{FC} amounts to conditioning on a statistic that is minimally sufficient for estimating θ; computing the associated P-values is not always trivial. Furthermore, the assumption that H_0^{FC} is true seems to be more extreme, a more substantial departure from H_0^{NP}, than H_0. Finally, testing the significance of assuming H_0 would seem to be more apropos in practice for applications in which the experimental design does not enforce that repeated experiments always yield the same value for $p_0(\hat{\theta})$. We cannot recommend the use of H_0^{FC} in general. Unfortunately, H_0^{NP} also presents problems....

If the probability distributions are discrete, there is no obvious means for defining an exact P-value for H_0^{NP} when H_0^{NP} is false; moreover, any P-value for H_0^{NP} when H_0^{NP} is true would depend on the correct value of the parameter θ, and the observed data does not determine this value exactly. The situation may be more favorable when measuring discrepancies with divergences that are "approximately ancillary" with respect to θ, but quantifying "approximately" seems to be problematic except in the limit of large numbers of

draws. (Some divergences are asymptotically ancillary in the limit of large numbers of draws, but this is not especially helpful, as any asymptotically consistent estimator $\hat{\theta}$ converges to the correct value in the limit of large numbers of draws; θ is almost surely known exactly in the limit of large numbers of draws, so there is no benefit to being independent of θ in that limit.) Section 3 of Robins and Wasserman (2000) reviews these and related issues.

Remark 4.1. Romano (1988), Henze (1996), Bickel et al. (2006), and others have shown that the P-values for H_0 converge in distribution to the uniform distribution over $[0, 1]$ in the limit of large numbers of draws, when H_0^{NP} is true. In particular, Romano (1988) and Henze (1996) prove this convergence for a wide class of divergence measures.

Remark 4.2. The surveys of Agresti (1992) and Agresti (2001) discuss exact P-values for contingency-tables/cross-tabulations, including criticism of fully conditional P-values. Gelman (2003) provides further criticism of fully conditional P-values. Ward (2012) numerically evaluates the different types of P-values for an application in population genetics. Section 4 of Bayarri and Berger (2004) and the references it cites discuss the menagerie of alternative P-values proposed recently.

5   Goodness of fit for various properties

For comparative purposes, we first review the null hypothesis of the previous section for testing the goodness of fit for distributional profile, namely

$$H_0 : x_1, x_2, \ldots, x_n \text{ are i.i.d. draws from } p_0(\hat{\theta}), \text{where } \hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n), \qquad (10)$$

with θ being the nuisance parameter. The measure of discrepancy for H_0 is usually taken to be a divergence between the empirical distribution $\hat{p}$ and the model $p_0(\hat{\theta})$ (in the continuous case in one dimension, a common characterization of the empirical distribution is the empirical cumulative distribution function; in the discrete case, a common characterization of the empirical distribution is the empirical probability mass function, that is, the set of empirical proportions). One example for p_0 is the Zipf distribution over m bins with parameter θ, a discrete distribution with the probability mass function

$$p_0^{(j)}(\theta) = \frac{C_\theta}{j^\theta} \qquad (11)$$

for $j = 1, 2, 3, \ldots, m$, where the normalization constant is

$$C_\theta = \frac{1}{\sum_{j=1}^m j^{-\theta}} \qquad (12)$$

and θ is a nonnegative real number.

When testing the goodness of fit for parameter estimates, we use the null hypothesis

$$H_0' : x_1, x_2, \ldots, x_n \text{ are i.i.d. draws from } p_0(\phi_0, \hat{\theta}), \text{where } \hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n), \qquad (13)$$

with θ being the nuisance parameter and ϕ being the parameter of interest (and with ϕ_0 being the value of ϕ assumed under the model). Please note that H_0 and H_0' are actually

equivalent, via the identification $p_0(\theta) = p_0(\phi_0, \theta)$. However, the measure of discrepancy for H_0' is usually taken to be a divergence between $\hat\phi$ and ϕ_0 rather than the divergence between $\hat{p}$ and $p_0(\hat\theta)$ that is more natural for H_0. Also, if ϕ is scalar-valued, then confidence intervals, confidence distributions, and parametric bootstrap distributions are more informative than a significance test. A significance test is appropriate if ϕ is vector-valued. One example for p_0 is the sorted Zipf distribution over m bins with θ being the power in the power law and with the maximum-likelihood estimate $\hat\phi$ being a permutation that sorts the bins into rank-order, that is, p_0 is the discrete distribution with the probability mass function

$$p_0^{(j)}(\phi, \theta) = \frac{C_\theta}{(\phi(j))^\theta} \tag{14}$$

for $j = 1, 2, 3, \ldots, m$, where the normalization constant C_θ is defined in (12) with θ being a nonnegative real number, and ϕ is a permutation of the numbers $1, 2, \ldots, m$. The choice for ϕ_0 that is of widest interest in applications is the identity permutation (that is, the "rearrangement" of the bins that does not permute any bins: $\phi_0(j) = j$ for $j = 1, 2, \ldots, m$).

When testing the goodness of fit for the standard Poisson regression with the distribution of Y given x being the Poisson distribution of mean $\exp\left(\theta^{(0)} + \sum_{j=1}^m \theta^{(j)} x^{(j)}\right)$, we use the null hypothesis

H_0'': $y_1, y_2, \ldots, y_n$ are independent draws from the Poisson distributions with means

$$\exp\left(\hat\theta^{(0)} + \sum_{j=1}^m \hat\theta^{(j)} x_1^{(j)}\right),\ \exp\left(\hat\theta^{(0)} + \sum_{j=1}^m \hat\theta^{(j)} x_2^{(j)}\right),\ \ldots,\ \exp\left(\hat\theta^{(0)} + \sum_{j=1}^m \hat\theta^{(j)} x_n^{(j)}\right),$$

$$\text{respectively,} \quad (15)$$

where θ is the nuisance parameter and $\hat\theta$ is a maximum-likelihood estimate. The measure of discrepancy for H_0'' is usually taken to be the log–likelihood-ratio (also known as the deviance)

$$g^2 = 2\sum_{k=1}^n y_k \ln(y_k/\hat\mu_k), \tag{16}$$

where $\hat\mu_k$ is the mean of the Poisson distribution associated with y_k in H_0'', namely,

$$\hat\mu_k = \exp\left(\hat\theta^{(0)} + \sum_{j=1}^m \hat\theta^{(j)} x_k^{(j)}\right). \tag{17}$$

One example is the cubic polynomial

$$\ln(\mu_k) = \theta^{(0)} + \theta^{(1)} x_k + \theta^{(2)}(x_k)^2 + \theta^{(3)}(x_k)^3 \tag{18}$$

for $k = 1, 2, \ldots, n$, which comes from the choice $m = 3$ and

$$x_k^{(1)} = x_k; \quad x_k^{(2)} = (x_k)^2; \quad x_k^{(3)} = (x_k)^3 \tag{19}$$

for $k = 1, 2, \ldots, n$, given observations as ordered pairs of scalars (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n). Of course, there are similar formulations for other generalized linear models, such as those discussed by McCullagh and Nelder (1989).

Acknowledgements

References

Agresti, A. (1992), A survey of exact inference for contingency tables, *Statist. Sci.*, **7**, 131–153.

Agresti, A. (2001), Exact inference for categorical data: recent advances and continuing controversies, *Stat. Med.*, **20**, 2709–2722.

Bayarri, M. J. and Berger, J. O. (2004), The interplay of Bayesian and frequentist analysis, *Statist. Sci.*, **19**, 58–80.

Bickel, P. J., Ritov, Y., and Stoker, T. M. (2006), Tailor-made tests for goodness of fit to semiparametric hypotheses, *Ann. Statist.*, **34**, 721–741.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009), Power-law distributions in empirical data, *SIAM Review*, **51**, 661–703.

Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge, UK: Cambridge University Press.

Gelman, A. (2003), A Bayesian formulation of exploratory data analysis and goodness-of-fit testing, *Internat. Stat. Rev.*, **71**, 369–382.

Henze, N. (1996), Empirical-distribution-function goodness-of-fit tests for discrete models, *Canad. J. Statist.*, **24**, 81–93.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, London: Chapman and Hall, 2nd ed.

Robins, J. M. and Wasserman, L. (2000), Conditioning, likelihood, and coherence: a review of some foundational concepts, *J. Amer. Statist. Assoc.*, **95**, 1340–1346.

Romano, J. P. (1988), A bootstrap revival of some nonparametric distance tests, *J. Amer. Statist. Assoc.*, **83**, 698–708.

Ward, R. A. (2012), Testing Hardy-Weinberg equilibrium with a simple root-mean-square statistic, Tech. Rep. 12-24, UT-Austin Institute for Computational Engineering and Sciences, available at http://www.ices.utexas.edu/media/reports/2012/1224.pdf.