GENERALIZED FREQUENTIST METHODS FOR CALCULATING P-VALUES AND CONFIDENCE INTERVALS

LUC DEMORTIER

The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA

E-mail: luc@fnal.gov

Generalized frequentism addresses problems that are not exactly solvable using conventional frequentism. Such problems include the calculation of p-values and confidence intervals when nuisance parameters are present, or when interest is focused on a complicated function of the parameters of the model under consideration. Although generalized frequentist methods are based on *exact* probability statements, they do not necessarily yield coverage in the conventional sense. However, simulation studies indicate that these methods tend to overcover, and often surpass other available methods in terms of test power or interval length.

1 Introduction

An often challenging component of frequentist calculations is the elimination of nuisance parameters. There seems to be no method that is generally applicable and at the same time theoretically guaranteed to preserve exact coverage in all cases. However, a couple of likelihood-based methods are known to behave reasonably well in many situations. In the first method, called profiling, the likelihood is *maximized* with respect to the nuisance parameters. The second method, marginalization, *integrates* the likelihood over these parameters. Whichever technique is chosen, its coverage properties for the problem at hand must then be verified a posteriori.

This paper aims to present a third approach, known as generalized frequentism.[1,2] Its strategy is to *extend* the conventional definitions of p-values and confidence intervals in such a way that statistical problems involving nuisance parameters can be solved "exactly", i.e. using exact probability statements. The resulting *generalized* p-values and confidence intervals tend to behave well with respect to the usual frequentist definitions, hence their interest.

2 Generalized p-Values

Let X be a random variable with density $f(x \,|\, \theta, \nu)$, where θ is the parameter of interest and ν is a nuisance parameter. We are interested in testing:

$$H_0 : \theta \;\leq\; \theta_0 \text{ versus } H_1 : \theta \;>\; \theta_0.$$

The usual way of solving this problem is to find a *test statistic* $T(X)$, defined as a function of the data X which does not depend on unknown parameters, whose distribution is free of unknown nuisance parameters, and which is stochastically increasing with θ, i.e. such that the probability $\mathbb{P}\mathrm{r}(T(X) \geq t \,|\, \theta)$ increases with θ for all t. One then calculates the p-value:

$$p \;=\; \mathbb{P}\mathrm{r}\big[T(X) \;\geq\; T(x) \,|\, H_0\big],$$

where x is the observed value of X. A small p-value indicates that the observed x does not support H_0.

There are many problems for which test statistics as defined above simply do not exist. In these cases a solution can be found by extending the definition of test statistic to that of a *generalized test variable*, which is a function $T(X, x, \theta, \nu)$ of the random variable X, its observed value x (treated as a constant), and the parameters θ and ν, such that the following requirements are satisfied:

1. $T(x, x, \theta, \nu)$ does not depend on θ or ν;
2. The distribution of $T(X, x, \theta_0, \nu)$ under H_0 is free of ν;
3. Given x and ν, $\mathbb{P}\mathrm{r}\big[T(X, x, \theta, \nu) \;\geq\; t \,|\, \theta\big]$ is a monotonic function of θ.

The generalized p-value based on $T(X, x, \theta, \nu)$ is defined similarly to a conventional p-value:

$$p \;=\; \mathbb{P}\mathrm{r}\big[T(X, x, \theta, \nu) \;\geq\; T(x, x, \theta, \nu) \,|\, H_0\big].$$

We emphasize that in this probability statement, only X is considered as a random variable, whereas the observed value x is held constant. Because of the way $T(X, x, \theta, \nu)$ is defined, this p-value is free of unknown parameters and allows the desired interpretation that small p corresponds to lack of support for H_0. However, although p is based on an exact probability statement, the coverage probability $\mathbb{P}\mathrm{r}(p \;\leq\; \alpha)$ may depend on nuisance parameters and needs to be checked explicitly.

There exists no general method that will systematically yield all possible generalized test variables for a given problem. However, an easy and useful recipe is available.[3,4] To formulate it we consider a slightly more general problem involving k unknown parameters $\alpha_1, \alpha_2, \ldots, \alpha_k$, and where the parameter of interest θ is a function of the α_i. We make the following assumptions:

1. There exists a set of observable statistics, $(X_1, X_2, \ldots, X_k)$, that is equal in number to the number of unknown parameters α_i.
2. There exists a set of invertible pivots[a], $(V_1, V_2, \ldots, V_k)$, relating the statistics (X_i) to the unknown parameters (α_i).

The recipe is then as follows:

1. By writing the parameter of interest, θ, in terms of the parameters α_i, express θ in terms of the statistics X_i and the pivots V_i.
2. Replace the X_i by their observed values x_i and subtract the result from θ.

For a simple application of this recipe, consider a sample $\{Y_1, Y_2, \ldots, Y_n\}$ drawn from Gauss(μ, σ), a Gaussian distribution with mean μ and width σ, both unknown. We are interested in the ratio $\theta \equiv \sigma/\mu$. The sample mean and standard deviation are a set of minimal sufficient statistics for μ and σ:

$$X_1 \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{and} \quad X_2 \equiv \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\bigl(Y_i - X_1\bigr)^2}.$$

The random variables

$$V_1 \equiv \frac{X_1 - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad V_2 \equiv \frac{n\, X_2^2}{\sigma^2}$$

relate the statistics (X_1, X_2) to (μ, σ), and have distributions free of unknown parameters:

$$V_1 \sim \text{Gauss}(0, 1) \quad \text{and} \quad V_2 \sim \chi^2_{n-1}.$$

Applying the recipe yields a generalized test variable, which can be written in terms of (V_1, V_2) or (X_1, X_2):

$$T \equiv \theta - \frac{\sqrt{n}\, x_2}{x_1\sqrt{V_2} - x_2\, V_1} = \theta - \frac{\sigma}{x_1 X_2/x_2 + \mu - X_1}.$$

The first expression for T shows that its distribution under H_0 is free of unknown parameters (the observed values x_1 and x_2 being treated as constants), whereas the second expression shows that the observed value of T is zero. The property of stochastic monotonicity is somewhat harder to verify.

2.1 Application to Poisson Significance Tests

For a slightly more complex application we turn to a common problem in high-energy physics. Consider a Poisson process consisting of a background with strength b superimposed on a signal with strength s:

$$f_N(n; b+s) = \frac{(b+s)^n}{n!}\, e^{-b-s}.$$

The nuisance parameter b is determined from a Gaussian measurement x:

$$f_X(x; b) = \frac{e^{-\frac{1}{2}\left(\frac{x-b}{\Delta b}\right)^2}}{\sqrt{2\pi}\,\Delta b}.$$

It is assumed that $b \geq 0$ but that, due to resolution effects, x can take both positive and negative values. We are interested in testing $H_0:\ s = 0$ versus $H_1:\ s > 0$. This problem has two parameters, b and s, two statistics, N and X, and two pivots:

$$V_1 = \frac{X - b}{\Delta b} \quad \text{and} \quad V_2 = F_N(N; b+s),$$

where $F_N(N; b+s)$ is the cumulative Poisson distribution with mean $b+s$. The pivot V_1 has a Gaussian distribution with mean 0 and width 1. Due to the discreteness of the Poisson distribution however, V_2 is only an approximate pivot. This can be remedied by introducing a uniform random variable U between 0 and 1, and replacing N by $Y \equiv N + U$ for the purpose of applying the recipe of section 2. This is nothing more than a mathematical artifice that provides us with an invertible pivot involving N. Indeed, the cumulative distribution of Y, say $F_Y^+(y, b+s)$, is an invertible pivot with a uniform distribution between 0 and 1. Let $G^+(Y, V)$ be the inverse of that pivot, i.e. $G^+(y, V) = \mu$ if and only if $V = F_Y^+(y, \mu)$. The generalized test variable is then:

$$T = s + \bigl(x - V_1\,\Delta b\bigr) - G^+(n, V_2),$$

and the generalized p-value is:

$$p = \mathbb{P}\mathrm{r}(T \geq 0 \,|\, s = 0).$$

From the definition of T it can be seen that this p-value is simply the probability for the difference between a Gauss$(x, \Delta b)$ and a Gamma$(n, 1)$ random variable to be positive. Analytically, the p-value equals the tail area of a convolution between these

[a] Pivots are random variables V_i that depend on the data X_j and the parameters α_k, but whose joint distribution is free of unknown parameters. They are called invertible if, for fixed values of the X_j, the mapping $(\alpha_k) \to (V_i)$ is invertible.

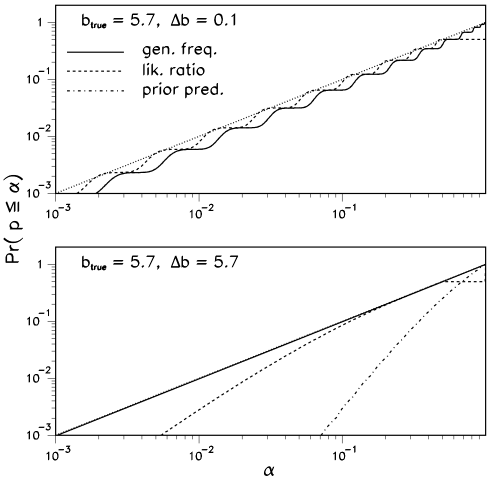

Figure 1. Comparative coverage of p-values. The dotted line represents exact coverage. In the top plot, the coverage of the prior-predictive p-value is indistinguishable from that of the generalized frequentist p-value. In the bottom plot, the coverage of the generalized frequentist p-value is indistinguishable from exact coverage.

random variables; for $n > 0$ it is given by:

$$p \;=\; \int_0^{+\infty} dt\, \frac{t^{n-1}\, e^{-t}}{\Gamma(n)}\; \frac{1+\mathrm{erf}\big(\frac{x-t}{\sqrt{2}\,\Delta b}\big)}{2},$$

and we define p to be 1 when $n = 0$. It is instructive to compare this p-value with two other methods. The first one is quite popular in high-energy physics, and consists in calculating the p-value assuming a fixed value for the nuisance parameter b, and then to average this p-value over $f_X(x;b)$, considered as a prior distribution for b. This yields the so-called "prior-predictive p-value" p_{pp}, which, for $n > 0$, is:

$$p_{pp} \;=\; \int_0^{+\infty} dt\; \frac{t^{n-1}\, e^{-t}}{\Gamma(n)}\; \frac{1+\mathrm{erf}\big(\frac{x-t}{\sqrt{2}\,\Delta b}\big)}{1+\mathrm{erf}\big(\frac{x}{\sqrt{2}\,\Delta b}\big)}.$$

The second method starts from the likelihood ratio statistic:

$$\lambda \;=\; \frac{\sup\limits_{s=0,\; b\geq 0} f_N(n; b+s)\, f_X(x;b)}{\sup\limits_{s\geq 0,\; b\geq 0} f_N(n; b+s)\, f_X(x;b)}$$

For large values of b, the distribution of $-2\ln\lambda$ under H_0 is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$, i.e. it assigns half a unit of probability to the singleton $\{-2\ln\lambda = 0\}$, whereas the other half is distributed as a chisquared with one degree of freedom over $0 < -2\ln\lambda < +\infty$. We then *define* the likelihood ratio p-value as the appropriate tail area of this distribution. For small values of b this is obviously an approximation, but not a bad one, in the sense that the frequentist validity of the p-value appears to be maintained: $\mathbb{P}\mathrm{r}(p \leq \alpha) \leq \alpha$. Figure 1 compares the coverage probability $\mathbb{P}\mathrm{r}(p \leq \alpha)$ of the three p-values just discussed, as a function of the significance level α, for a simple numerical example. The coverage calculation fluctuates both n and x. For small values of the background uncertainty Δb, the likelihood ratio p-value is somewhat better than the other two, but for large Δb the generalized frequentist p-value is clearly superior.

3 Generalized Confidence Intervals

A standard method for constructing confidence intervals is based on pivots. Let $Q(X,\theta)$ be a pivot for a random variable X with distribution $F_X(x;\theta)$, and let S_α be a subset of the sample space of Q such that

$$\mathbb{P}\mathrm{r}(Q(X,\theta) \in S_\alpha) \;=\; \alpha.$$

Note that the probability in this equation is unambiguously determined since the distribution of Q does not depend on unknown parameters. Given an observed value x for X, a $100\alpha\%$ confidence interval for θ is then:

$$C_\alpha \;=\; \{\theta :\;\; Q(x,\theta) \in S_\alpha\}$$

In problems for which a conventional pivot is not available, one can try to construct a *generalized* pivot, i.e. a function $Q(X, x, \theta, \nu)$ of the random variable X, its observed value x, the parameter of interest θ, and the nuisance parameter ν, such that the following requirements are satisfied:

1. $Q(x, x, \theta, \nu)$ does not depend on ν;
2. The distribution of $Q(X, x, \theta, \nu)$ is free of (θ, ν).

Generalized confidence intervals can then be defined similarly to conventional ones, but using $Q(X, x, \theta, \nu)$ instead of $Q(X, \theta)$.

As with p-values, there is no systematic method for generating all possible generalized pivots for a problem, but a simple recipe is available.[3,4] It is based on the same assumptions as those listed in section 2, and the recipe itself is almost identical to the one used to obtain generalized test variables. The only difference is step 2, which becomes:

2. Replace the X_i by their observed values x_i.

In other words, given a generalized test variable $T(X, x, \theta, \nu)$, the corresponding generalized pivot is obtained as $Q(X, x, \theta, \nu) = \theta - T(X, x, \theta, \nu)$.

3.1 *Application to Poisson Upper Limits*

Suppose that we observe a Poisson event count X_1 with mean $b + \epsilon\sigma$, where b is a background, ϵ a sensitivity factor, and σ a cross section of interest:

$$X_1 \sim \text{Poisson}(b + \epsilon\,\sigma).$$

Information about b and ϵ are assumed to come from two auxiliary measurements:

$$X_2 \sim \text{Poisson}(c\,b), \qquad X_3 \sim \text{Poisson}(\tau\,\epsilon),$$

where c and τ are known constants. Applying the above recipe yields the following generalized pivot for σ:

$$Q = \frac{\tau\left[G^-(x_1, V_1) - G^-(x_2, V_2)/c\right]}{G^-(x_3, V_3)},$$

where, similarly to the G^+ introduced in section 2.1, G^- is the inverse of the pivot defined by the cumulative distribution of $X - U$, X being a Poisson variate and U a uniform one.[b] The V_i quantities are independent uniform random variables, and the x_i are the observed values of the corresponding X_i.

Suppose now that we wish to calculate upper limits on σ. It is straightforward to verify that the "observed" value of Q is the parameter of interest σ. Therefore, upper limits on σ are obtained by calculating the corresponding quantiles of the distribution of Q. A numerical example of the coverage of these upper limits is shown in Figure 2, together with a reference Bayes calculation. There is slight undercoverage at high σ values.

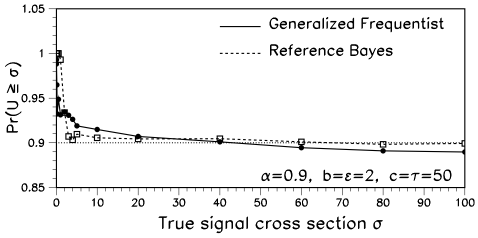

Figure 2. Coverage of upper limits U on the cross section of a signal process, as a function of the true value σ of that cross section. The nominal uncertainties on the background b and the efficiency ϵ are 10%. Solid: generalized frequentist; dashes: reference Bayes.

4 Summary

Generalized frequentist methods allow one to calculate significances and confidence intervals in a wide variety of situations involving nuisance parameters.

In problems with continuous sample spaces, these methods are based on exact probability statements but do not have a conventional frequency interpretation. Nevertheless, their conventional frequentist properties appear to be very good. In fact, Hannig *et al.*[4] have shown that under some general conditions, generalized confidence intervals for scalar or vector parameters have proper frequentist coverage, at least asymptotically.

Although the current literature on generalized frequentism does not appear to treat problems with discrete sample spaces, we have described how these can be solved by introducing a randomization scheme.

Using a simple Poisson example, we have shown that generalized frequentist methods compare favorably to other methods of eliminating nuisance parameters, such as likelihood ratio and Bayes.

References

1. Kam-Wah Tsui and Samaradasa Weerahandi, "Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters," J. Amer. Statist. Assoc. **84**, 602 (1989). Erratum: ibid. **86**, 256 (1991).
2. Samaradasa Weerahandi, "Generalized confidence intervals," J. Amer. Statist. Assoc. **88**, 899 (1993). Erratum: ibid. **89**, 726 (1994).
3. Hari K. Iyer and Paul D. Patterson, "A recipe for constructing generalized pivotal quantities and generalized confidence intervals," Colorado State University Department of Statistics Technical Report 2002/10; also at `http://www.stat.colostate.edu/research/2002_10.pdf`.
4. Jan Hannig, Hari Iyer, and Paul L. Patterson, "On fiducial generalized confidence intervals," Colorado State University Department of Statistics Technical Report 2004/12; also at `http://www.stat.colostate.edu/~hari/fiducial/fgpq.pdf`.

[b] When applying generalized frequentist methods to discrete distributions, the results depend slightly on the randomization scheme. The use of G^+ in section 2.1 was dictated by the desire to maintain coverage, even though $G^+(x, V)$ is not defined when $x = 0$. In section 3.1 it seems more important to use a function that *is* defined at $x = 0$, which is the case for $G^-(x, V)$.