Generalized Frequentist Methods for Computing p-Values and Confidence Intervals

Luc Demortier
The Rockefeller University

PHYSTAT 2005

Oxford, September 12–15, 2005

Outline of talk:

- Motivation and General Principles;
- p-Values;
- Confidence Intervals;
- Summary.

Motivation

A major challenge in the frequentist approach to hypothesis testing and interval estimation is the elimination of nuisance parameters. There seems to be no general method that is theoretically guaranteed to preserve exact coverage in all cases, but a couple of methods are known to have good asymptotic properties:

1. PROFILING
   Calculations are based on a likelihood that is *maximized* with respect to the nuisance parameters, separately for each value of the parameter(s) of interest;

2. MARGINALIZING
   Calculations are based on a likelihood that is *integrated* over the nuisance parameters, with or without a "prior".

Whichever method is chosen, the practitioner must then, a posteriori, verify its coverage properties for the problem at hand.

The purpose of this talk is to present a third general method for eliminating nuisance parameters.

Generalized Inference

The aim of generalized inference 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.

Caveats:

1. This is a very limited aim... No claim is made that a new statistical paradigm is being introduced.
2. "Exact" probability statements are not necessarily useful in practice, but they are theoretically valid.
3. Even though generalized p-values and generalized confidence intervals are based on *extensions* of the usual frequentist definitions, they tend to behave well with respect to the usual definitions, hence their interest.

Some 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).

2. Samaradasa Weerahandi, "Generalized confidence intervals," J. Amer. Statist. Assoc. **88**, 899 (1993).

3. Samaradasa Weerahandi, "Exact statistical methods for data analysis," Springer-Verlag New York, Inc., 1995 (XIV+328pp).

4. 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.

5. Samaradasa Weerahandi, "Generalized inference in repeated measures: exact methods in MANOVA and mixed models," John Wiley & Sons, Hoboken, New Jersey, 2004 (XV+352pp).

6. Jan Hannig, Hari Iyer, and Paul L. Patterson, "On fiducial generalized confidence intervals," Colorado State University Department of Statistics Technical Report 2004/12.

7. Jan Hannig, "On multidimensional fiducial generalized confidence intervals," Colorado State University Department of Statistics Technical Report 2005/1.

Conventional p-Values

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

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

Recall the usual definition of a p-value:

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

where x is the observed value of X and $T(X)$ is a *test statistic*, i.e. a function of the data X with the following properties:

1. $T(X)$ does not depend on unknown parameters;
2. The distribution of $T(X)$ does not depend on unknown nuisance parameters;
3. The probability $\mathrm{Pr}(T(X) \geq t \,|\, \theta)$ increases with θ.

A small p-value indicates that the observed x does not support H_0.

Generalized p-Values

p-Values are generalized by extending the definition of test statistic:

A generalized test variable $T(X, x, \theta, \nu)$ is a function of the random variable X, its observed value x, 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, $\Pr\big[T(X, x, \theta, \nu) \geq t \,|\, \theta\big]$ is a monotonic function of θ.

The generalized p-value based on $T(X, x, \theta, \nu)$ is now defined in the usual way:

$$p \;=\; \Pr\big[T(X, x, \theta, \nu) \;\geq\; T(x, x, \theta, \nu) \,|\, H_0\big].$$

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 $\Pr(p \;\leq\; \alpha)$ itself may depend on nuisance parameters and needs to be checked explicitly.

The Substitution Method for Generalized p-Values

Generalized p-values were introduced in 1989, but for a long time the construction of generalized test variables relied on intuition and guesswork. A general recipe was introduced in 2002.

Assumptions:

1. There is a set of observable statistics $(X_1, X_2, \ldots, X_k)$ with known distributions, that is equal in number to the number of unknown parameters of the problem, say $(\alpha_1, \alpha_2, \ldots, \alpha_k)$; In many applications a set of minimal sufficient statistics will serve this purpose;
2. Through a set of random variables $(V_1, V_2, \ldots, V_k)$ having distributions free of unknown parameters, the statistics X_i are related to the unknown parameters.

Recipe:

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

Example of the Substitution Method

- Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a sample drawn from a Gaussian distribution with mean μ and width σ. We are interested in $\theta \equiv \sigma/\mu$.
- The sample mean and variance, $\bar{Y}$ and S^2 respectively, are a set of minimal sufficient statistics for μ and σ; they correspond to the X_i in the recipe.
- The random variables

$$Z \equiv \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad U \equiv \frac{n\,S^2}{\sigma^2}$$

relate the statistics $\bar{Y}$ and S^2 to the parameters μ and σ, and have distributions free of unknown parameters:

$$Z \sim N(0,1) \quad \text{and} \quad U \sim \chi^2_{n-1}.$$

They correspond to the V_i in the recipe.

- Finally, the recipe says (1) to write θ in terms of Z, U, $\bar{Y}$, and S^2, (2) to replace $\bar{Y}$, S by their observed values, and (3) to subtract the result from θ:

$$\theta \equiv \frac{\sigma}{\mu} = \frac{\sqrt{n}\,S/\sqrt{U}}{\bar{Y} - S\,Z/\sqrt{U}} \longrightarrow T \equiv \theta - \frac{\sqrt{n}\,s/\sqrt{U}}{\bar{y} - s\,Z/\sqrt{U}} = \theta - \frac{\sigma}{\bar{y}S/s + \mu - \bar{Y}}.$$

Generalized Inference with Discrete Distributions

One of the assumptions of the substitution method is the existence of pivotal quantities, i.e. random variables that depend on unknown parameters but whose distribution does not.

For continuous distributions this is usually not a problem, as the cumulative distribution function (cdf) is itself an exact pivot, with a uniform distribution.

For discrete distributions however, the cdf is not an exact pivot. This is unfortunate in HEP, where we often deal with Poisson and binomial statistics.

To solve this problem we have adopted the following procedure:

1. Randomize the observed, discrete data. For example, when observing a discrete number of events from a Poisson distribution, we add or subtract a uniform random number between 0 and 1.
2. Apply the generalized frequentist method to the randomized observation and its distribution.
3. When interpreting or checking the properties of the result, "ignore" the components related to the randomization.

This is in line with our intent to use generalized frequentist methods as a mere toolbox for deriving useful frequentist results.

Randomizing Poisson Statistics (1)

Let N and U be two random variables with the following distributions:

$$\begin{aligned} N &\sim \mathrm{Poisson}(\mu) \\ U &\sim \mathcal{U}_{[0,1[} \end{aligned}$$

One way to generate a continuous statistic from N is to work with $Y^{+} \equiv N + U$, whose distribution is given by:

$$F_{Y^{+}}(y \,|\, \mu) \;=\; \sum_{i=0}^{\lfloor y \rfloor - 1} \frac{\mu^{i}\, e^{-\mu}}{i!} \;+\; \big(y \,-\, \lfloor y \rfloor\big)\, \frac{\mu^{\lfloor y \rfloor}\, e^{-\mu}}{\lfloor y \rfloor !},$$

where $\lfloor y \rfloor$ is the largest integer smaller than or equal to y. Alternatively, one could work with $Y^{-} \equiv N - U$, whose distribution is:

$$F_{Y^{-}}(y \,|\, \mu) \;=\; \sum_{i=0}^{\lceil y \rceil} \frac{\mu^{i}\, e^{-\mu}}{i!} \;-\; \big(\lceil y \rceil \,-\, y\big)\, \frac{\mu^{\lceil y \rceil}\, e^{-\mu}}{\lceil y \rceil !},$$

where $\lceil y \rceil$ is the smallest integer larger than or equal to y.

Randomizing Poisson Statistics (2)

In the remainder of this talk it will be more convenient to use the notation:

$$
\begin{aligned}
G_y^{(+)}(\mu) &\equiv F_{Y^+}(y \,|\, \mu) \\
G_y^{(-)}(\mu) &\equiv F_{Y^-}(y \,|\, \mu).
\end{aligned}
$$

Then, if y is a positive integer and V a uniform random number between 0 and 1, one has the interesting property that:

$$
\begin{aligned}
G_y^{(+)-1}(V) &\sim \mathrm{Gamma}(y, 1) & (y \geq 1) \\
G_y^{(-)-1}(V) &\sim \mathrm{Gamma}(y+1, 1) & (y \geq 0)
\end{aligned}
$$

Poisson Signal Significance (1)

Consider a Poisson process consisting of a background with strength b superimposed on a signal with unknown strength s:

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

where the background rate b is determined from a Gaussian measurement x:

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

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:

$$H_0: \ s = 0 \qquad \text{vs.} \qquad H_1: \ s > 0.$$

The background strength b is a nuisance parameter.

Poisson Signal Significance (2)

This problem has two parameters, b and s, two statistics, N and X, and after randomizing N with $Y \equiv N + U$, two pivots:

$$\begin{aligned} V &= G_Y^{(+)}(b+s) \quad &\sim \mathcal{U}_{[0,1]}, \\ W &= \frac{X-b}{\Delta b} \quad &\sim N(0,1). \end{aligned}$$

The generalized test variable is then:

$$T = s + \bigl(x - W\,\Delta b\bigr) - G_y^{(+)-1}(V),$$

and the p-value is:

$$p = \Pr(T \geq 0 \,|\, s = 0).$$

For integer y this is the probability for the difference between an $N(x, \Delta b)$ and a Gamma$(y, 1)$ random variable to be positive. Analytically it corresponds to the tail area of a convolution between these random variables:

$$p = \int_0^{+\infty} dt\, \frac{t^{y-1}\, e^{-t}}{\Gamma(y)}\, \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-t}{\sqrt{2}\,\Delta b}\right)\right].$$

Poisson Signal Significance: Other Methods (1)

It is instructive to compare the generalized p-value just derived with two other methods:

1. The Prior-Predictive p-Value

This is a very common technique in HEP. The information about b is essentially treated as a subjective prior in a Bayesian calculation of the p-value. Using Monte Carlo integration, this results in the following algorithm:

1. Generate $b \sim N(x, \Delta b)$;
2. If $b < 0$, repeat step 1;
3. Generate $n \sim \mathrm{Poisson}(b)$.

The p-value is then approximated by the fraction of generated n values that are at least as large as the one observed. Analytically, the prior-predictive p-value is given by:

$$p = \sum_{n=y}^{\infty} \int_0^{+\infty} db\ \frac{b^n\, e^{-b}}{n!}\ \frac{e^{-\frac{1}{2}\left(\frac{b-x}{\Delta b}\right)^2}}{\frac{1}{2}\left[1+\mathrm{erf}\left(\frac{x}{\sqrt{2}\,\Delta b}\right)\right]} = \int_0^{+\infty} dt\ \frac{t^{y-1}\, e^{-t}}{\Gamma(y)}\ \frac{1+\mathrm{erf}\left(\frac{x-t}{\sqrt{2}\,\Delta b}\right)}{1+\mathrm{erf}\left(\frac{x}{\sqrt{2}\,\Delta b}\right)}.$$

Poisson Signal Significance: Other Methods (2)

2. The Likelihood-Ratio p-Value

Here one assumes that the background information comes from a genuine measurement, so that a joint likelihood can be defined:

$$\mathcal{L}(b, s \,|\, y, x) \;=\; \frac{(b+s)^y \, e^{-b-s}}{y!} \, \frac{e^{-\frac{1}{2}\left(\frac{x-b}{\Delta b}\right)^2}}{\sqrt{2\pi}\,\Delta b}.$$

The likelihood ratio statistic is:

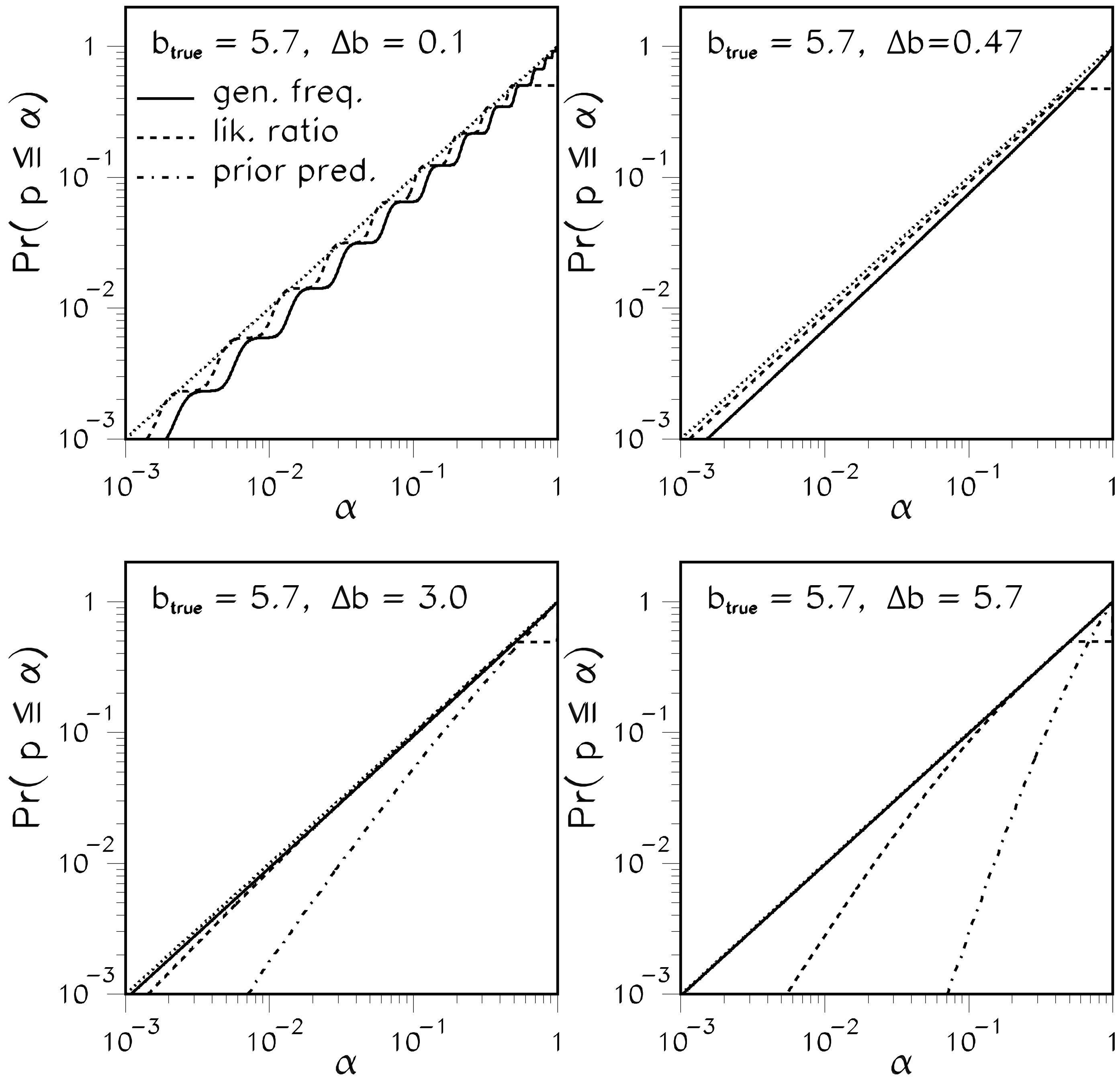
$$\lambda \;=\; \frac{\sup_{\substack{b\geq 0\\ s=0}} \;\mathcal{L}(b, s \,|\, y, x)}{\sup_{\substack{b\geq 0\\ s\geq 0}} \;\mathcal{L}(b, s \,|\, y, x)}.$$

It can be shown that for large values of b, the quantity $-2\ln\lambda$ is distributed as $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$. For small b however, the distribution of $-2\ln\lambda$ depends on b. In that case, a general way of eliminating the b dependence while maintaining frequentist coverage is to calculate the supremum p-value:

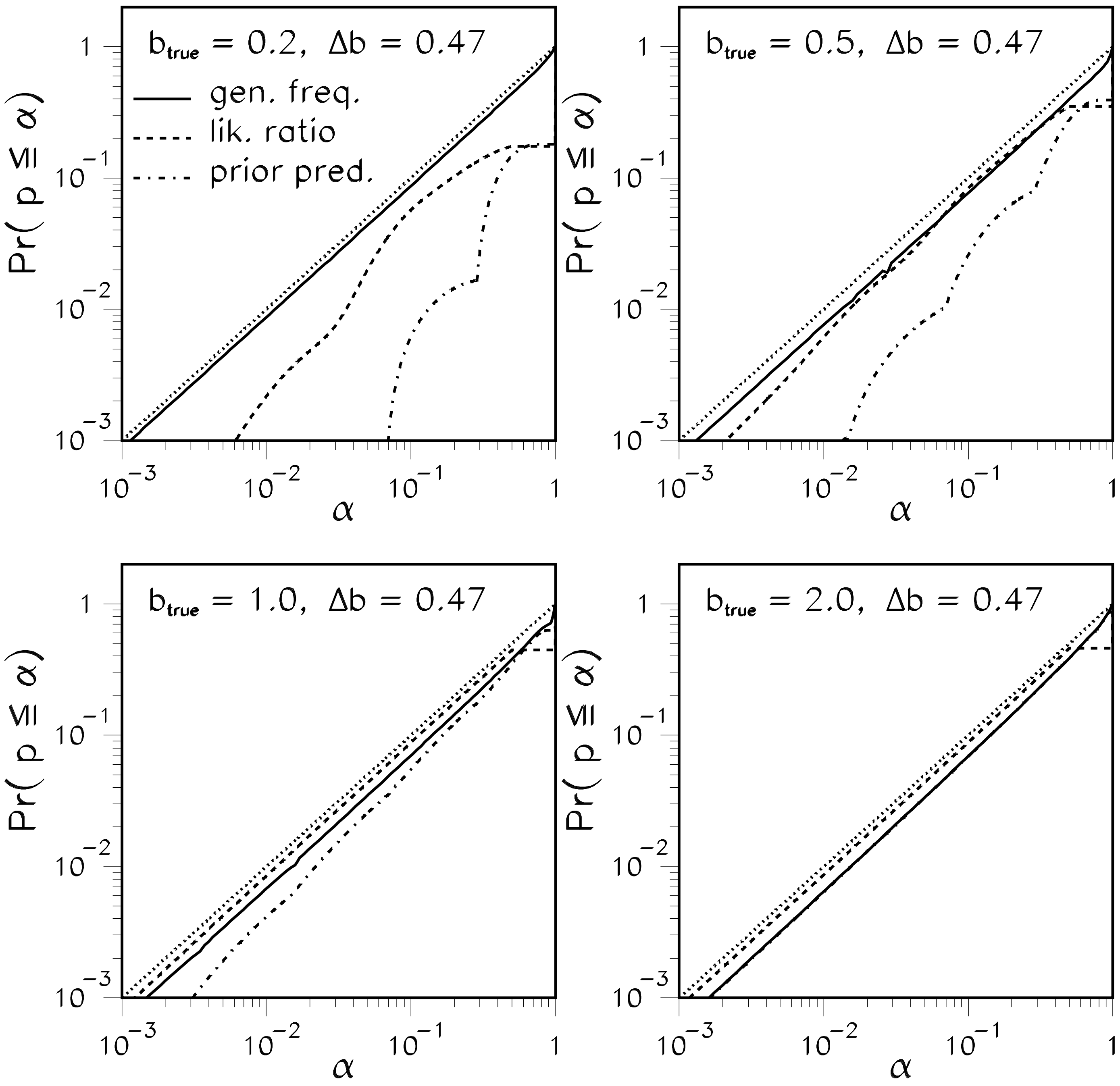
$$p_{\text{sup}} \;=\; \sup_{b\geq 0} \;\Pr(\lambda \leq \lambda_0 \,|\, s=0)$$

If $-2\ln\lambda$ is stochastically increasing with b, then $p_{\text{sup}} = \lim_{b\to\infty} p$. We will assume that this is true in the following.

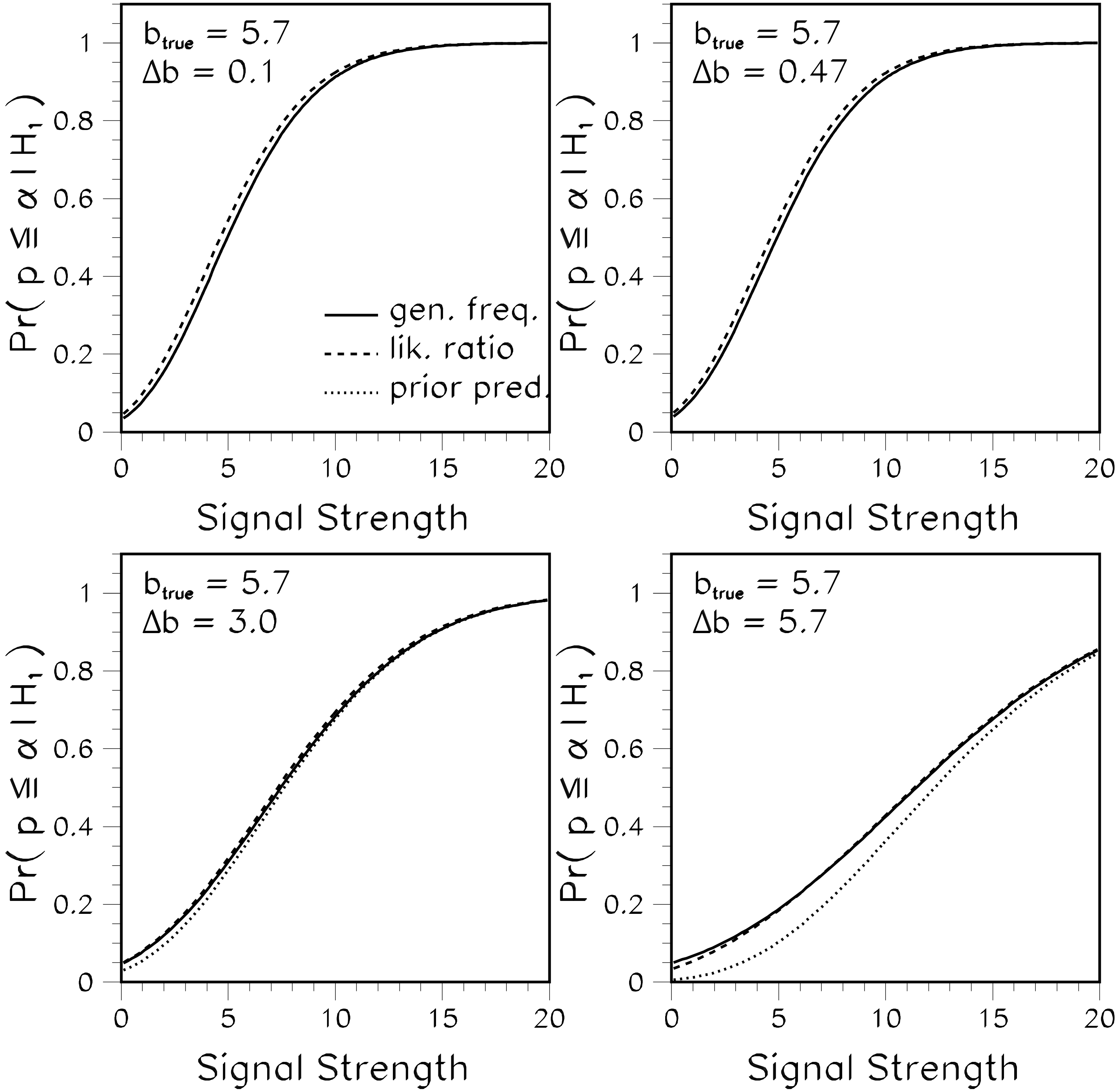
Coverage of p-Values (1)



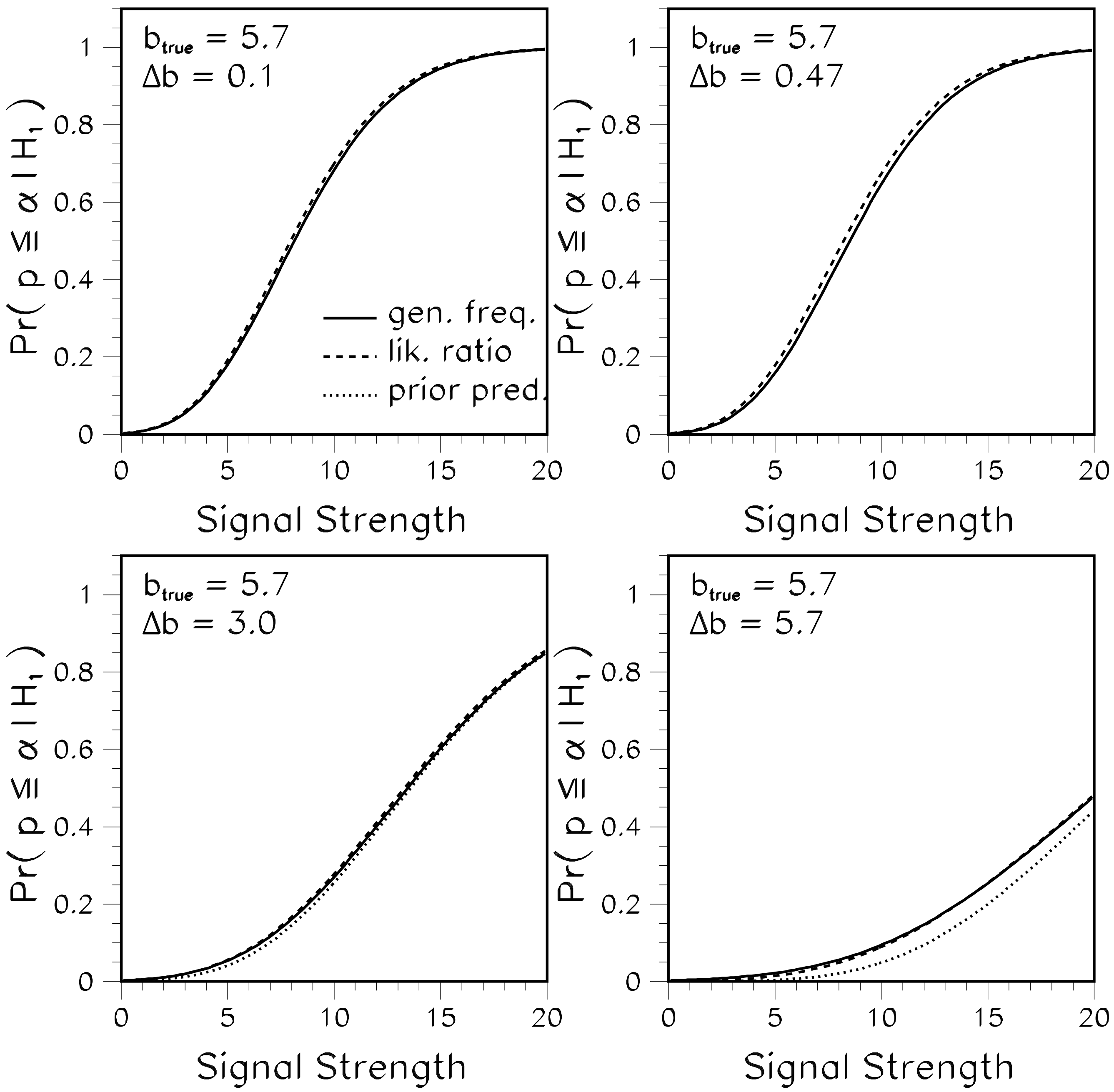
Coverage of p-Values (2)



Power of p-Values for $\alpha = 0.05$



Power of p-Values for $\alpha = 0.0027$



Conventional Confidence Intervals

A standard method for constructing confidence intervals is based on pivots:

Let $X \sim F(x \,|\, \theta)$. A random variable $Q(X, \theta)$ is a pivot if the distribution of $Q(X, \theta)$ is the same for all values of θ.

To construct an α-confidence level interval C_α for θ we need to find a subset S_α of the sample space of Q, such that:

$$\Pr(Q(X, \theta) \in S_\alpha) \;=\; \alpha.$$

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, the confidence interval for θ is then:

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

Example of Conventional Confidence Interval (1)

For a simple example we return to the problem of a Poisson process whose mean is the sum of a background and a signal component:

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

where we assume b to be known exactly. After randomizing with $Y = N - U$, the following is an exact pivot:

$$Q(Y, s) \;\equiv\; G_Y^{(-)}(b+s).$$

Let t be a constant such that:

$$\mathrm{Pr}\Big[G_Y^{(-)}(b+s) \geq t\Big] \;=\; \alpha.$$

Since $G_Y^{(-)}(b+s)$ has a uniform distribution between 0 and 1, the solution of this equation is $t = 1 - \alpha$, and the confidence interval for s is:

$$C_\alpha \;=\; \{s : \;\; G_y^{(-)}(b+s) \geq 1-\alpha\} \;=\; \{s : \;\; s \leq G_y^{(-)-1}(1-\alpha) - b\}.$$

For integer y this gives the usual Poisson upper limits on the signal strength s.

Example of Conventional Confidence Interval (2)

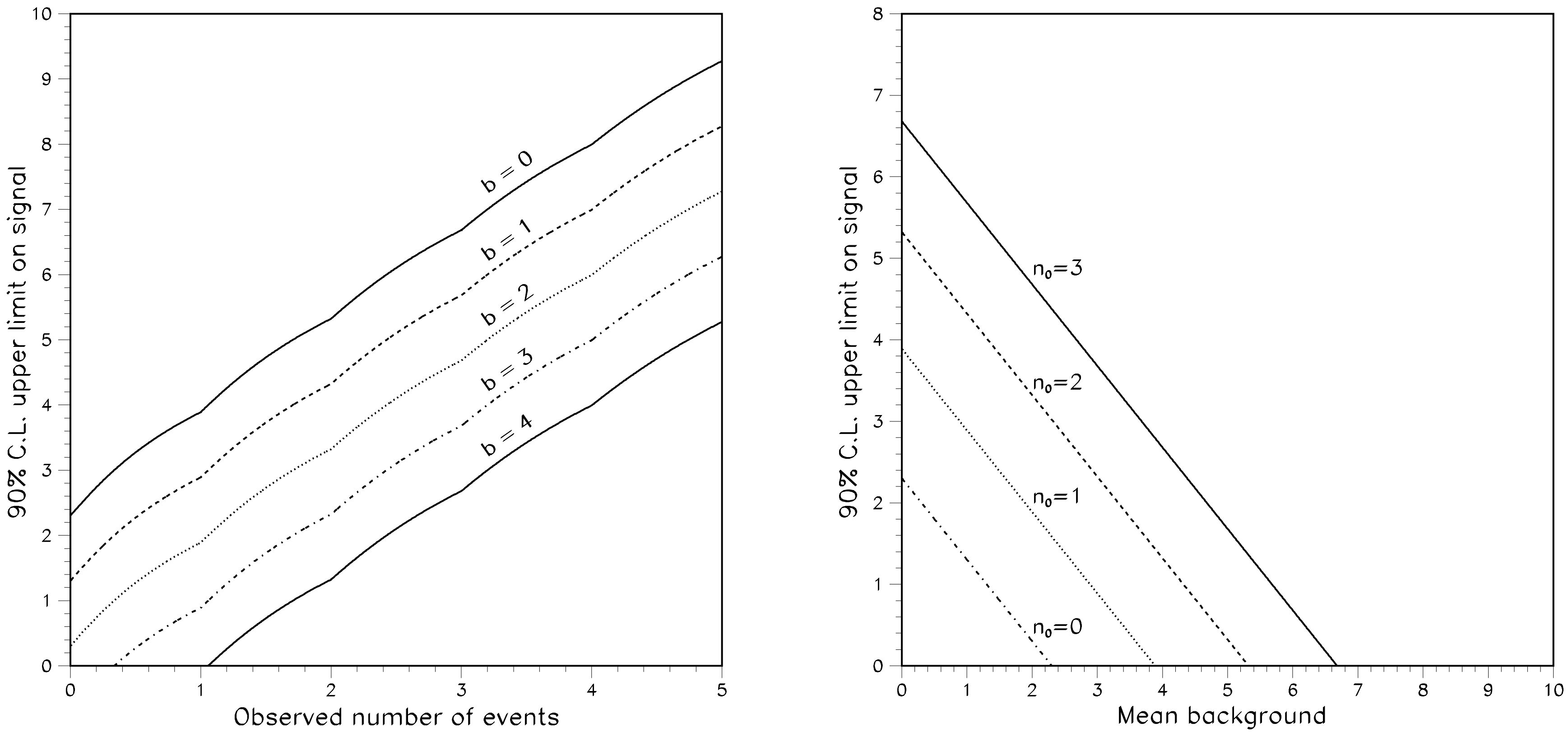


Figure 1: Conventional frequentist upper limits on the rate of a Poisson process, as a function of the observed number of events (left) and the mean background (right).

Generalized Confidence Intervals

Confidence intervals are generalized by extending the definition of pivot:

A generalized pivot $Q(X, x, \theta, \nu)$ is a function 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. The observed value $Q(x, x, \theta, \nu)$ does not depend on ν;
2. The distribution of $Q(X, x, \theta, \nu)$ is free of θ and ν.

The generalized confidence interval based on $Q(X, x, \theta, \nu)$ is now defined in the usual way. Let S_α be a subset of the sample space of Q such that:

$$\mathbf{Pr}\Big[Q(X, x, \theta, \nu) \in S_\alpha\Big] \;=\; \alpha,$$

then:

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

is a $100\alpha\%$ generalized confidence interval for θ.

The Substitution Method for Generalized Pivots

This is very similar to the substitution method for generalized p-values:

Assumptions:

1. There is a set of observable statistics $(X_1, X_2, \ldots, X_k)$ with known distributions, that is equal in number to the number of unknown parameters of the problem, say $(\alpha_1, \alpha_2, \ldots, \alpha_k)$; In many applications a set of minimal sufficient statistics will serve this purpose;
2. Through a set of random variables $(V_1, V_2, \ldots, V_k)$ having distributions free of unknown parameters, the statistics X_i are related to the unknown parameters.

Recipe:

1. By writing the parameter of interest, θ, in terms of the parameters α_i, express θ in terms of the sufficient statistics X_i and the random variables V_i.
2. Replace the statistics X_i by their observed values x_i.

Application to Poisson Proc. without Nuisance Par. (1)

Back to the example of a Poisson process whose mean is a sum of an *un*known signal and known background component. The conventional pivot for that problem is:

$$Q_{\text{conv.}} \;=\; G_Y^{(-)}(b+s).$$

A simple calculation with the substitution method yields the following generalized pivot:

$$Q_{\text{gen.}} \;=\; G_y^{(-)-1}\Big[G_Y^{(-)}(b+s)\Big] \;-\; b.$$

It is easy to verify that upper limits calculated with $Q_{\text{gen.}}$ are identical to those calculated with $Q_{\text{conv.}}$. However, the substitution method does not by itself allow one to find *all possible* generalized pivots. In fact, for the above Poisson problem an equally valid generalized pivot is given by:

$$Q'_{\text{gen.}} \;=\; \Big|G_y^{(-)-1}\Big[G_Y^{(-)}(b+s)\Big] \;-\; b\Big|.$$

Upper limits calculated with $Q'_{\text{gen.}}$ are never zero and do not systematically decrease with the mean background, as do limits calculated with $Q_{\text{conv.}}$.

Application to Poisson Proc. without Nuisance Par. (2)

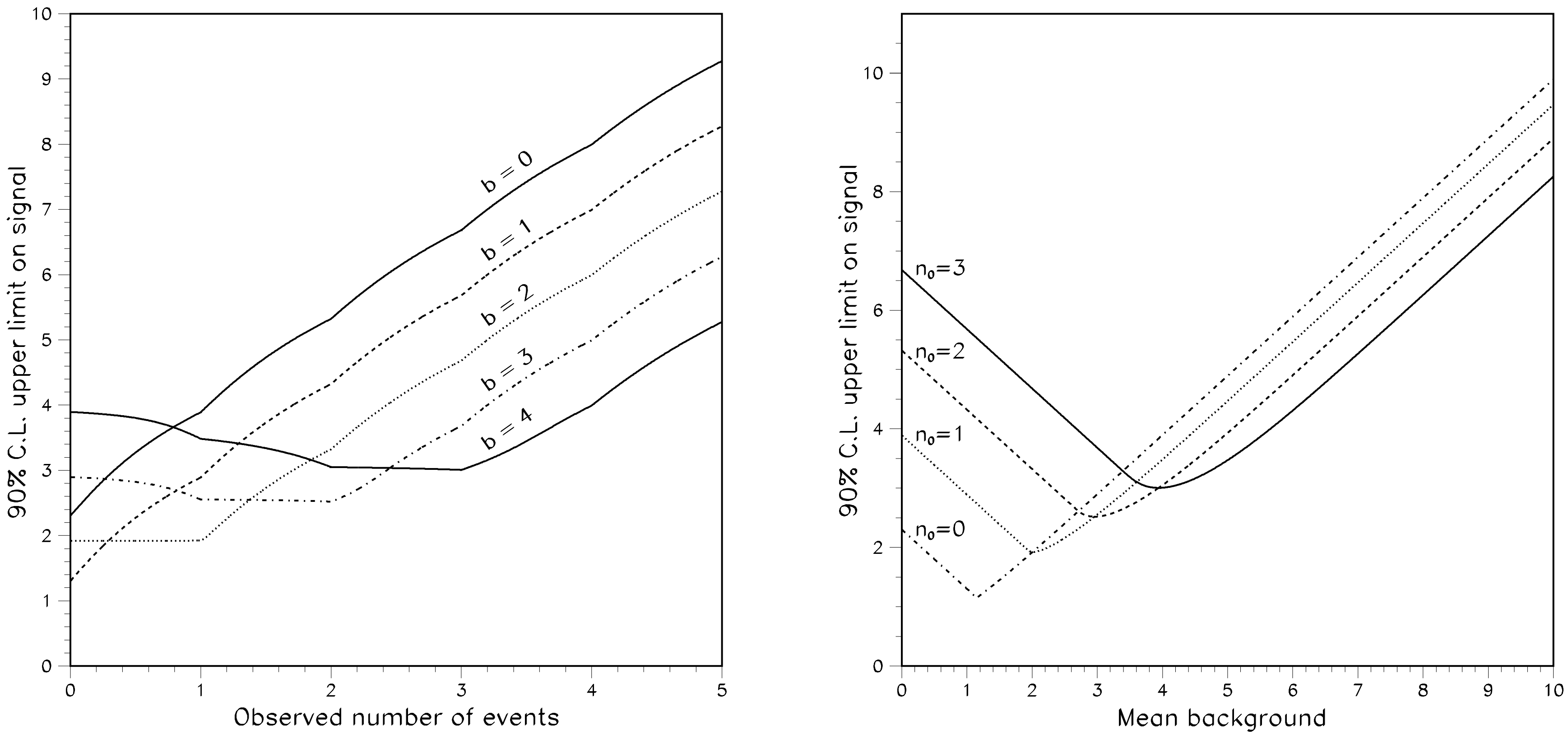


Figure 2: Generalized frequentist upper limits on the rate of a Poisson process, as a function of the observed number of events (left) and the mean background (right).

Application to Poisson Proc. with Nuisance Par. (1)

Suppose now that we observe:

$$L \;\sim\; \mathrm{Poisson}(b \,+\, \epsilon\,\sigma),$$

and wish to set an upper limit on the signal cross section σ. Information about the background b and effective efficiency ϵ comes from two auxiliary measurements:

$$M \;\sim\; \mathrm{Poisson}(c\,b), \qquad N \;\sim\; \mathrm{Poisson}(\tau\,\epsilon),$$

where c and τ are known constants. To solve this problem we again randomize, using $X = L - T_1$, $Y = M - T_2$, and $Z = N - T_3$, the T_i being uniform random variables between 0 and 1. We then have the conventional pivots:

$$U \;=\; G_X^{(-)}(b + \epsilon\sigma), \qquad V \;=\; G_Y^{(-)}(c\,b), \qquad W \;=\; G_Z^{(-)}(\tau\,\epsilon),$$

and the substitution method yields the following generalized pivot for the signal strength s:

$$Q \;=\; \frac{\tau\left[G_x^{(-)-1}(U) \;-\; G_y^{(-)-1}(V)/c\right]}{G_z^{(-)-1}(W)}$$

Application to Poisson Proc. with Nuisance Par. (2)

Substituting the definitions of U, V, and W in the expression for the generalized pivot Q leads to:

$$Q(X, Y, Z, x, y, z, \sigma, b, \epsilon) \;=\; \frac{\tau\left[G_x^{(-)-1}(G_X^{(-)}(b + \epsilon\sigma)) \;-\; G_y^{(-)-1}(G_Y^{(-)}(c\,b))/c\right]}{G_z^{(-)-1}(G_Z^{(-)}(\tau\,\epsilon))},$$

showing that the "observed" value of Q is the parameter of interest σ. Therefore, upper limits on σ correspond to the quantiles of the distribution of Q. A simple Monte Carlo algorithm to calculate this distribution is as follows:

1. Generate $R_1 \sim \mathrm{Gamma}(x + 1, 1)$;
2. Generate $R_2 \sim \mathrm{Gamma}(y + 1, 1)$;
3. Generate $R_3 \sim \mathrm{Gamma}(z + 1, 1)$;
4. Calculate $Q \;=\; \tau\,(R_1 \;-\; R_2/c)/R_3$.

Application to Poisson Proc. with Nuisance Par. (3)

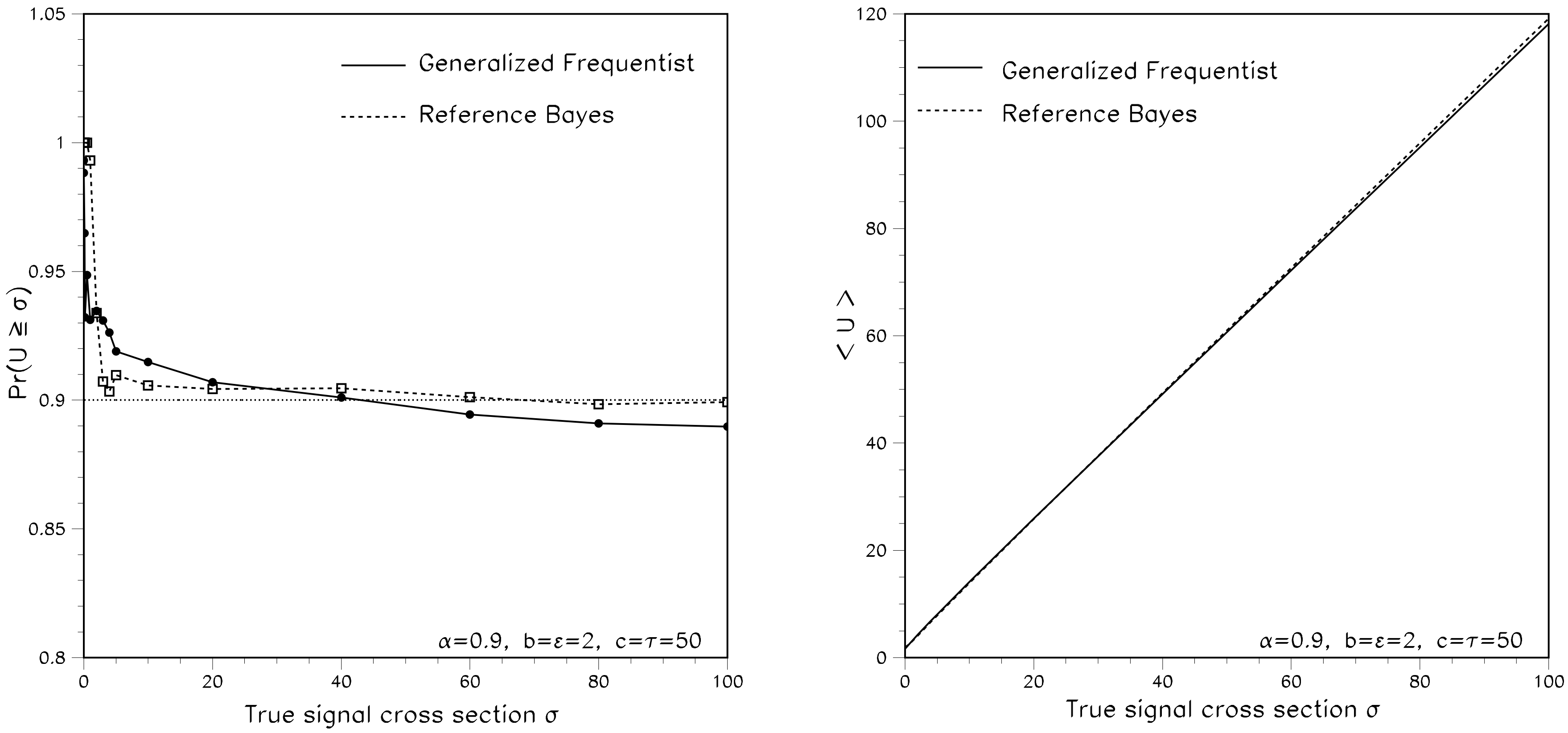


Figure 3: Coverage (left) and size (right) of upper limits on the cross section of a signal process, as a function of the true value of that cross section. The nominal uncertainties on the background b and the efficiency ϵ are 10%. Solid: generalized frequentist; dashes: reference Bayes.

Application to Poisson Proc. with Nuisance Par. (4)

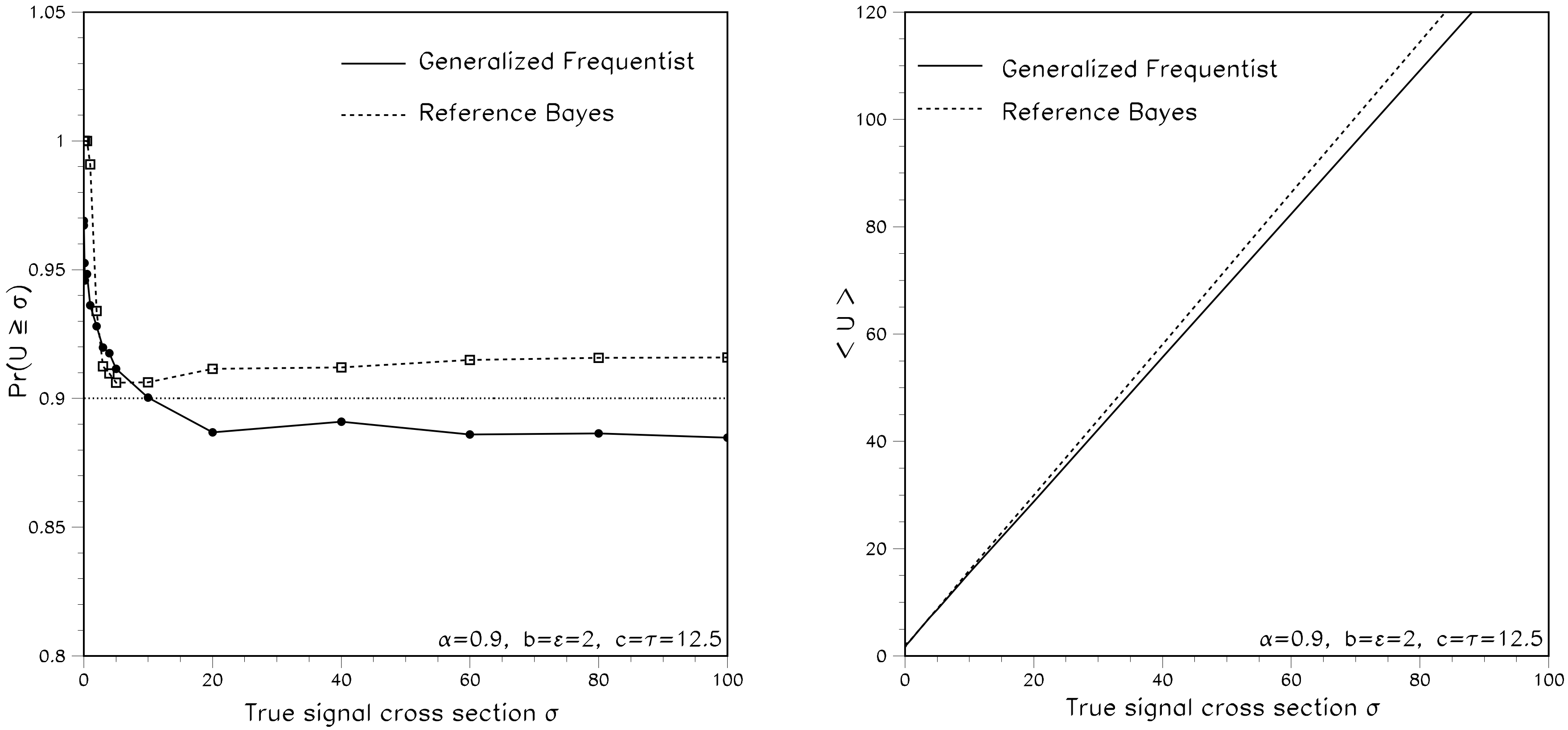


Figure 4: Coverage (left) and size (right) of upper limits on the cross section of a signal process, as a function of the true value of that cross section. The nominal uncertainties on the background b and the efficiency ϵ are 20%. Solid: generalized frequentist; dashes: reference Bayes.

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, it is known that under some general conditions, generalized confidence intervals for scalar or vector parameters have proper frequentist coverage, at least asymptotically.

In problems with discrete sample spaces the methods may no longer be based on exact probability statements, but they still seem to be useful from a conventional frequentist viewpoint.

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.