Interval Estimation

Luc Demortier

Laboratory of Experimental High Energy Physics
The Rockefeller University, New York, NY 10065, USA

Monday 10th June, 2013

Abstract

This is a review of interval construction methods, both frequentist and Bayesian. The frequentist side covers Neyman's construction, test inversion, pivoting, asymptotic approximations, the bootstrap, and various techniques for handling nuisance parameters. On the Bayesian side the discussion includes highest posterior density regions, equal-tailed intervals, upper and lower limits, likelihood regions, and lowest posterior loss regions. Many examples are given to illustrate the concepts. The review ends with a discussion of the role of intervals in search procedures in high energy physics.

1 Introduction

Point estimation procedures provide very concise summaries of what the data have to say about a given parameter of interest: Each summary consists of a single number, representing what is in some sense the most likely value of the parameter, given the observed data. The downside of this conciseness is that it does not come with a characterisation of the reliability of the estimate. This is what interval estimates attempt to remedy: Instead of a single numerical estimate, two numerical limits are provided, plus a level of confidence about the true value of the parameter of interest lying between these limits. If there is more than one parameter of interest, intervals are replaced by multi-dimensional regions. It is also possible for an interval construction method to yield a union of disjoint intervals or regions, and this may be sensible in some contexts (see example 4 below).

Not surprisingly, the correct interpretation of the confidence level of an interval or region depends strongly on the statistical paradigm one is operating in, *Bayesian* or *frequentist*. Furthermore, specification of a desired confidence level does not uniquely determine an interval construction. Other desiderata enter into play, for example interval length, behaviour under reparameterisation, effect of physical boundaries, systematic uncertainties, etc. After briefly reviewing such interval characterisations in section 2, we describe frequentist constructions in section 3 and Bayesian ones in section 4. Following the two sections on methodology, a graphical comparison using a problem involving a physical boundary is provided in section 5. Next, the role of interval construction in search procedures is addressed in section 6, in particular in relation to the issues of *coverage* and *measurement sensitivity*. A short summary of the chapter is given in section 7.

2 Characterisation of interval constructions

Confidence level is the primary characteristic of an interval construction, but its meaning is radically different in the Bayesian and frequentist approaches to statistical inference. In the Bayesian approach, the final result of a measurement is the posterior distribution of the parameter of interest, and interval estimation is one method among others for summarising the information contained in this distribution. The confidence level associated with a Bayesian interval is the integral of the posterior over that interval and is also called *credibility.* It represents the probability for the parameter of interest to lie somewhere inside the interval, given one's prior beliefs and the observed data.

On the other hand, the frequentist approach does not associate probability distributions with constants of nature and therefore requires a different concept to quantify the reliability of interval estimates. This is the concept of *coverage*, which characterises how an interval construction procedure behaves over large numbers of replications of the measurement under consideration. Coverage answers the question: "If N new data sets are collected under the same conditions as the actually observed one, and the same measurement is performed each time, what fraction of these measurements will yield a confidence interval that contains the true value of the parameter of interest, as $N \to \infty$?" It should be noted that a desired coverage cannot always be achieved exactly, for example when the observable is discrete (e.g. a number of events), or when systematic uncertainties are present.

Even though the credibility and coverage interpretations of a confidence level belong to different statistical paradigms, it is often instructive to investigate the credibility of frequentist intervals and the coverage of Bayesian intervals. This is because Bayesian inferences fully condition on the observed data, whereas frequentist ones take both observed and unobserved data into account. Thus one could question the *relevance* of a frequentist result for the data at hand, and this can be clarified by studying its posterior credibility with a well-motivated, proper Bayesian prior. Similarly, one could question the *replicability* of a Bayesian result, and this can be investigated with a well-defined ensemble of measurements (real or simulated). An interesting result in this regard is that when a proper prior is used, the prior-averaged coverage of a Bayesian interval construction equals its nominal credibility, thus guaranteeing replicability in some average sense. When a proper, *evidence-based prior* is not available, coverage may still provide useful guidance in choosing a so-called *objective prior* [KW96].

As already indicated, the desired confidence level of an interval estimate does not uniquely specify how to construct such an estimate. There are many possibilities, and for choosing among them it is useful to examine other interesting properties:

- *Interval length:* For a given confidence level, short intervals are more informative than long ones, at least when they cover the true value of the parameter. A frequentist concept that may be useful in this regard is that of *expected length*, which is the interval length averaged over the ensemble of all possible observations and viewed as a function of the parameter of interest. It can be shown that the expected length of a confidence interval is equal to the probability of including a false value of the parameter in the interval, integrated over all false values [Pra61]. Since the expected length involves an ensemble average, it is not a Bayesian criterion. However, given a Bayesian posterior distribution, a popular interval construction is that known as *highest posterior density* (HPD), which yields the shortest interval of a given credibility (see section 4).

- *Equivariance under parameter transformations:* When measuring a quantity such as a particle mass θ, the result may be used by theorists to draw inferences about another quantity $\eta = f(\theta)$, where f can be an arbitrarily complicated function. Suppose now that we apply the *same* interval construction procedure to both parameters, obtaining $[\theta_1, \theta_2]$ and $[\eta_1, \eta_2]$, respectively. It would be useful to have $[\eta_1, \eta_2] = [f(\theta_1), f(\theta_2)]$, but this is generally not true; For example, the shortest intervals in θ do not necessarily map onto the shortest intervals in η.

- *Behaviour with respect to systematic uncertainties:* Systematic uncertainties are modeled with the help of nuisance parameters, that is, parameters that are of no direct interest to the experimenter but must be known in order to draw inferences about the parameter of interest. Examples include calibration constants, energy scales, and detection efficiencies. Nuisance parameters are constrained by auxiliary measurements or Bayesian priors, which determine the distribution of associated systematic uncertainties. Typically one expects the length of an interval for the parameter of interest to increase with the width of that distribution.

- *Effect of physical boundaries:* When the parameter space has boundaries imposed by physical constraints, some interval constructions may yield intervals that lie partially or completely in the unphysical region for some subset of observations. The physical part of these intervals is then either unreasonably narrow or empty, a highly undesirable situation. Examples where this may happen are measurements of efficiencies and acceptances, where the true value is constrained to lie between 0 and 1, and particle masses, where it must be positive. It is also possible for a parameter boundary to have special physical significance. In a search for new particles, for example, the production rate is constrained to positive values. The value zero, however, has special significance since it corresponds to the background-only hypothesis (no new particles). Whether interval estimation is the appropriate type of inference in such situations, as opposed to e.g. hypothesis testing, is an issue that needs to be carefully thought out.

- *Relation to point estimate:* When measuring a property of a system known to exist (e.g. the mass of the top quark), one usually reports both an interval and a point estimate, and it is desirable that the latter be contained in the former. However, intervals and point estimates provide different types of inference and there is no unique relationship between them. One can try to introduce such a relationship[1], but this does not necessarily yield optimal procedures. On the other hand, there are some natural associations, as between equal-tailed intervals and medians, and between likelihood-ratio-ordered intervals and maximum-likelihood estimates, but these associations are not exclusive. Furthermore, one should not expect an interval to be *centered* on the

[1] The Hodges–Lehmann estimator, for example, is defined as the limit of an interval construction as the confidence level goes to zero [HL63]

associated point estimate; this depends on the probability distribution of the observation(s), on the presence of physical boundaries, systematic uncertainties, etc. Finally, it is sometimes meaningful to report an interval without point estimate, for example when a new physics process has not been observed and one wishes to provide an upper limit on its production rate.

- *Generality:* Is the interval construction procedure general enough that it can be applied to any problem, regardless of its complexity?

Needless to say, there does not exist a single interval construction method that adequately addresses all the above characterisations in all the problems encountered in practice. It is nevertheless useful to keep these characterisations in mind, and perhaps to prioritise them when searching for an optimal method in a specific situation.

3 Frequentist methods

The basic frequentist interval construction is due to Neyman [Ney37]. This procedure is very general, can be applied to multi-dimensional problems and also provides a method for the elimination of *nuisance parameters*[2]. We therefore start this section with a discussion of this construction. Subsequently, sections 3.2 through 3.4 present simpler methods. In less simple, more realistic situations, *bootstrapping methods* as described in section 3.5 can be used. These are particularly well suited to particle physics, where observations ("events") are independent and identically distributed. As a matter of fact, parametric bootstrap methods are already being used by physicists every time they substitute a point estimate for a parameter in a model in order to generate so-called pseudo-data. It is therefore important to understand what can be expected from the bootstrap in terms of some of the interval properties listed in the introduction. We close this discussion of frequentist intervals with comments on the handling of nuisance parameters, together with a detailed case study, in section 3.6.

3.1 Neyman's construction

The Neyman construction is illustrated in figure 1 for the case of estimating a one-dimensional continuous parameter θ from observations whose distribution depends only on θ. The first step is to choose a point estimator $\hat{\theta}$ of θ, to make a graph of θ versus $\hat{\theta}$ and to plot the probability density distribution (pdf) of $\hat{\theta}$ for several values of θ. In figure 1(a) this has been done for $\theta = 1$, 2, and 3. For each value of θ considered in step 1, step 2 consists in selecting an interval of $\hat{\theta}$ values that has a fixed integrated probability, for example 68%. Finally, at step 3 the interval boundaries are connected across θ values to obtain the so-called confidence belt. Once data are collected, the observed value $\hat{\theta}_{obs}$ of the estimator of θ is computed, and the confidence belt is used to derive the corresponding interval $[\theta_1, \theta_2]$ for θ (figure 1(b)).

To see why this procedure works, consider that if $\theta_{\rm true}$ is the true value of θ, there is by construction a 68% probability for the point $(\hat{\theta}_{\rm obs}, \theta_{\rm true})$ to be inside the confidence belt, and

[2] "Eliminating nuisance parameters" is statistics terminology for "incorporating the effect of systematic uncertainties."

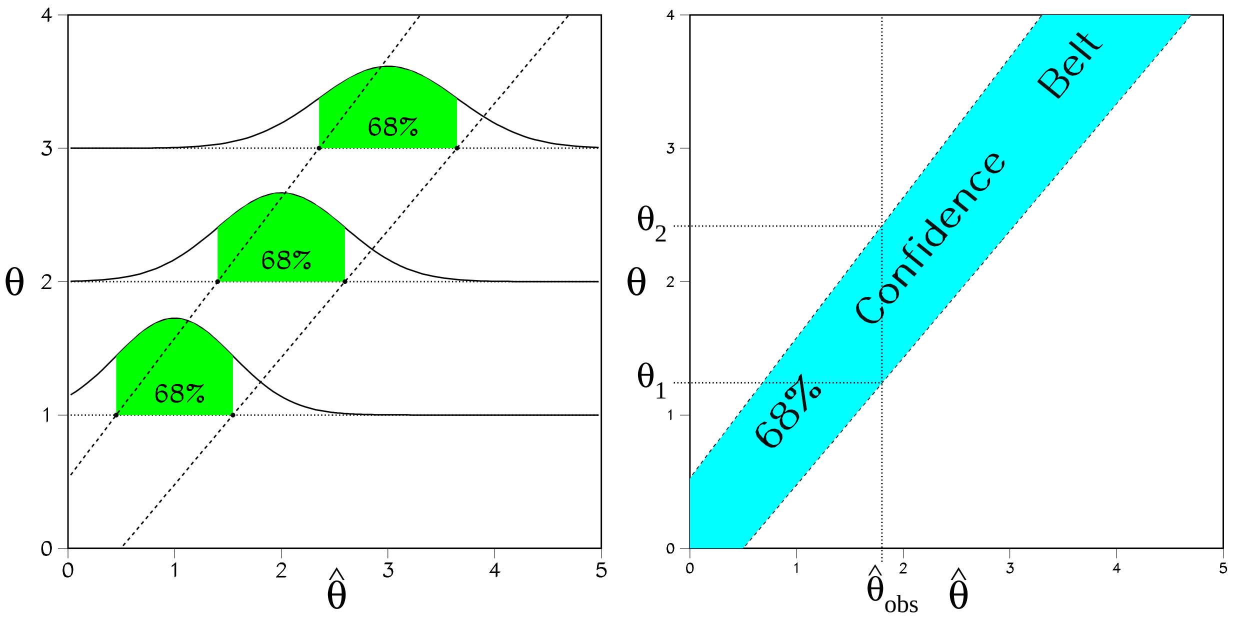

Figure 1: (a) Neyman construction of a 68% confidence interval on a parameter θ. (b) Example use of this construction (see text). Only the part of the construction that falls inside the first quadrant is shown.

only when this happens will θ_{true} be inside the interval $[\theta_1, \theta_2]$ corresponding to $\hat{\theta}_{\text{obs}}$. There is therefore a 68% chance that the reported interval will contain θ_{true}, and this holds regardless of the value of θ_{true}.

The Neyman construction requires four ingredients: an estimator $\hat{\theta}$ of the parameter of interest θ, a reference ensemble, an ordering rule, and a confidence level. We now take a look at each of these ingredients individually.

3.1.1 Ingredient 1: the estimator

The estimator is the quantity plotted along the abscissa in the Neyman construction plot. Suppose for example that we collect n independent measurements x_i of the mean θ of a Gaussian distribution with known standard deviation. Then clearly we should use the average $\bar{x}$ of the x_i as an estimate of θ, since $\bar{x}$ is a sufficient statistic[3] for θ. On the other hand, if θ is constrained to be positive, then it might make more sense to use $\hat{\theta} = \max\{0, \bar{x}\}$ instead of $\hat{\theta} = \bar{x}$. These two estimators lead to intervals with very different properties. We will come back to this example in section 5.

It should be pointed out that the original formulation of Neyman's construction does not require a choice of point estimator. It proceeds directly from the distribution of the full data sample at each parameter value. Thus, if the sample contains n measurements, step 2 of the construction consists in delimiting an n-dimensional region of sample space with total integrated probability equal to the desired confidence level. This is clearly a non-trivial operation. Fortunately, in the vast majority of practical cases the reduction of the observed

[3] A statistic $T(X)$ is sufficient for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ. In a sense, $T(X)$ captures all the information about θ contained in the sample.

sample to a point estimate is a simplifying step that captures all the relevant information.

3.1.2 Ingredient 2: the reference ensemble

This refers to the probability distribution of the point estimator under replication of the measurement. In order to specify these replications, one must decide which random and non-random aspects of the measurement are relevant to the inference of interest. We give two examples to illustrate this point.

Example 1 (Efficiency estimation) First consider the measurement of an efficiency ϵ. A useful point estimator of ϵ is the ratio $\hat{\epsilon} \equiv k/n$ of the number k of events of interest ("successes") over the total number n of events collected. However, the distribution of this estimator depends on how the data were collected. If we took data until our total sample reached a certain size, then k will follow a binomial distribution. If we took data until we found a pre-specified number of events of interest, then the total number of events collected will have a negative binomial distribution. These two data collection schemes differ by their stopping rule, a non-random aspect of the measurement that affects inferences about ϵ. Note that these schemes also imply very different prior opinions about ϵ: In the binomial case one leaves open the possibility that ϵ could be zero, whereas in the negative binomial case ϵ is a priori believed to be non-zero. ■

Example 2 (Mass of a short-lived particle) For an example where a random aspect of the observation affects inferences, consider the measurement of the mass of a short-lived particle whose decay mode determines the measurement resolution. We only have one observation of the particle. Should we then refer our measurement to an ensemble that includes all possible decay modes, or only the decay mode actually observed? For simplicity assume that the estimator $\hat{\theta}$ of the mass follows a Gaussian distribution with mean θ and standard deviation σ, and that there is a probability p_h that the particle decays hadronically, in which case $\sigma \equiv \sigma_h$; otherwise the particle decays leptonically and $\sigma \equiv \sigma_\ell < \sigma_h$. Thus if we decide to condition on the observed decay mode, the distribution of $\hat{\theta}$ is Gaussian with mean θ and width σ_h or σ_ℓ. If we don't condition, the distribution of $\hat{\theta}$ is a mixture of two Gaussians:

$$f_\theta(\hat{\theta}) \;=\; p_h \frac{e^{-\frac{1}{2}\left(\frac{\hat{\theta}-\theta}{\sigma_h}\right)^2}}{\sqrt{2\pi}\,\sigma_h} \;+\; (1-p_h)\frac{e^{-\frac{1}{2}\left(\frac{\hat{\theta}-\theta}{\sigma_\ell}\right)^2}}{\sqrt{2\pi}\,\sigma_\ell}\,. \tag{1}$$

By ignoring the decay-mode information we can actually expect a more precise measurement. Indeed, if we report our measurement in the form $\hat{\theta} \pm \delta$, then δ equals σ_h for hadronic decays and σ_ℓ for leptonic decays. When the decay mode is ignored, δ is the solution of

$$\int_{\theta-\delta}^{\theta+\delta} f_\theta(\hat{\theta})\,\mathrm{d}\hat{\theta} \;=\; p_h \,\mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma_h}\right) \;+\; (1-p_h)\,\mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma_\ell}\right) \;=\; 0.68\,. \tag{2}$$

For a numerical example, take $p_h = 0.5$, $\sigma_h = 10$ and $\sigma_\ell = 1$ (in arbitrary units). The expected interval length with known decay mode is then $2[p_h\sigma_h + (1-p_h)\sigma_\ell] = 11.0$. When the decay mode is ignored, the expected interval length is $2\delta \approx 9.50$, noticeably smaller. To understand this feature, imagine repeating the measurement a large number of times. In the conditional analysis the coverage of the interval is 68% both within the subensemble of hadronic decays and within the subensemble of leptonic decays. On the other hand, in the unconditional analysis the coverage is $\mathrm{erf}(\delta/(\sqrt{2}\,\sigma_h)) \approx 36\%$ for hadronic decays and $\mathrm{erf}(\delta/(\sqrt{2}\,\sigma_\ell)) \approx 100\%$ for leptonic decays, correctly averaging to 68% over all decays combined. Qualitatively, by shifting some coverage probability from the hadronic decays to the higher-precision leptonic ones, the unconditional construction is able to reduce the expected interval length.

The above problem is an adaptation to high-energy physics of a famous example in the statistics literature [Cox58; Bon88], used to discuss the merits of conditioning versus power (or interval length). In spite of the loss of expected precision, most physicists would agree to condition on the observed decay mode. ■

3.1.3 Ingredient 3: the ordering rule

The ordering rule is the rule we use to decide which $\hat{\theta}$ values to include in the interval at step 2 of the construction. The only constraint on that interval is that it must contain 68% of the $\hat{\theta}$ distribution (or whatever confidence level is desired for the overall construction). For example, we could start with the $\hat{\theta}$ value that has the largest probability density and then keep adding values with lower and lower probability density until we cover 68% of the distribution. Another possibility is to start with $\hat{\theta} = -\infty$ and add increasing values of $\hat{\theta}$, again until we reach 68%. Of course, in order to obtain a smooth confidence belt at the end, we should choose the ordering rule consistently from one θ value to the next. This is what endows the resulting intervals with their inferential meaning: an ordering rule is a rule that orders parameter values according to their perceived compatibility with the observed data. Below we list the most common ordering rules, all assuming that we are interested in an $(1-\alpha)$-level confidence set $C_{1-\alpha}$ for a parameter θ. We use a point estimator $\hat{\theta}$ whose observed value in the data at hand is $\hat{\theta}_{\mathrm{obs}}$; the cumulative distribution of $\hat{\theta}$ is $F_\theta(\hat{\theta})$ and its density is $f_\theta(\hat{\theta})$.

- *Lower-limit ordering:* $C_{1-\alpha} = \{\theta:\ F_\theta(\hat{\theta}_{\mathrm{obs}}) \le 1-\alpha\}$
  $C_{1-\alpha}$ is the set of θ values for which $\hat{\theta}_{\mathrm{obs}}$ is smaller than or equal to the $100(1-\alpha)^{\mathrm{th}}$ percentile of F_θ. If, as is usually the case, $\hat{\theta}$ is stochastically increasing with θ[4], then the parameter value θ_{low} with $F_{\theta_{low}}(\hat{\theta}_{\mathrm{obs}}) = 1-\alpha$ is the lower limit of $C_{1-\alpha}$.

- *Upper-limit ordering:* $C_{1-\alpha} = \{\theta:\ F_\theta(\hat{\theta}_{\mathrm{obs}}) \ge \alpha\}$
  $C_{1-\alpha}$ is the set of θ values for which $\hat{\theta}_{\mathrm{obs}}$ is larger than or equal to the $100\alpha^{\mathrm{th}}$ percentile of F_θ. The parameter value θ_{up} with $F_{\theta_{up}}(\hat{\theta}_{\mathrm{obs}}) = \alpha$ is the upper limit of the set $C_{1-\alpha}$.

- *Equal-tails ordering:* $C_{1-\alpha} = \{\theta:\ \frac{\alpha}{2} \le F_\theta(\hat{\theta}_{\mathrm{obs}}) \le 1-\frac{\alpha}{2}\}$
  $C_{1-\alpha}$ is the set of θ values for which $\hat{\theta}_{\mathrm{obs}}$ falls between the $100(\frac{\alpha}{2})^{\mathrm{th}}$ and $100(1-\frac{\alpha}{2})^{\mathrm{th}}$ percentiles of F_θ. The previous definitions of lower and upper limits show that equal-tailed intervals must have the form $C_{1-\alpha} = [\theta_1, \theta_2]$, where the boundaries are themselves confidence limits: $]-\infty, \theta_1]$ and $[\theta_2, +\infty[$ are both $(\frac{\alpha}{2})$ CL intervals. Furthermore, θ_1

[4] The random variable $\hat{\theta}$ is said to be stochastically increasing with the parameter θ if $\theta_1 < \theta_2$ implies $F_{\theta_1}(\hat{\theta}) > F_{\theta_2}(\hat{\theta})$. In words, the bulk of the distribution of $\hat{\theta}$ tracks changes in θ.

and θ_2 can be solved from $F_{\theta_1}(\hat{\theta}_{\rm obs}) = 1-\frac{\alpha}{2}$ and $F_{\theta_2}(\hat{\theta}_{\rm obs}) = \frac{\alpha}{2}$. The relationship between equal-tailed intervals and confidence limits leads to the following interpretation of the former in terms of "plausibility" [ET93, p. 157]: Values of θ smaller than θ_1 are implausible because they result in probability less than $\alpha/2$ of obtaining a $\hat{\theta}$ value at least as *large* as observed, and values of θ larger than θ_2 are implausible because they result in probability less than $\alpha/2$ of obtaining a $\hat{\theta}$ value at least as *small* as observed.

- *Probability-density ordering:* $C_{1-\alpha} = \{\theta :\ f_\theta(\hat{\theta}_{\rm obs}) \geq k_{1-\alpha}(\theta)\}$
  $C_{1-\alpha}$ is the set of θ values for which $\hat{\theta}_{\rm obs}$ falls within the $100(1-\alpha)\%$ most probable region of f_θ. The cutoff $k_{1-\alpha}(\theta)$ is determined by the coverage requirement, namely that $\int_{C_{1-\alpha}} f_\theta(\hat{\theta})\,{\rm d}\hat{\theta} = 1-\alpha$; this requirement can introduce a θ dependence in $k_{1-\alpha}$, but no $\hat{\theta}$ dependence.

- *Likelihood-ratio ordering:* $C_{1-\alpha} = \{\theta :\ f_\theta(\hat{\theta}_{\rm obs})/[\max_{\theta'} f_{\theta'}(\hat{\theta}_{\rm obs})] \geq k'_{1-\alpha}(\theta)\}$
  $C_{1-\alpha}$ is the set of θ values for which $\hat{\theta}_{\rm obs}$ falls in the region of sampling probability $1-\alpha$ where the likelihood ratio in favour of θ is larger than anywhere outside the region ($k'_{1-\alpha}(\theta)$ is fixed by the coverage requirement). Note that the maximisation in the denominator of the likelihood ratio must be restricted to the physical region of θ-space [FC98].

In contrast with the equal-tails, upper-limit, and lower-limit ordering rules, the probability-density and likelihood-ratio rules do not always produce simple intervals: In complex problems they may yield confidence sets that are unions of disjoint intervals.

Table 1 summarises these ordering rules together with their defining equations, and shows the result of applying them to a measurement of the lifetime θ of an exponential decay. In this case the probability density of the estimated lifetime $\hat{\theta}$ is $f_\theta(\hat{\theta}) = \exp(-\hat{\theta}/\theta)/\theta$, and the cumulative distribution is $F_\theta(\hat{\theta}) = 1 - \exp(-\hat{\theta}/\theta)$.

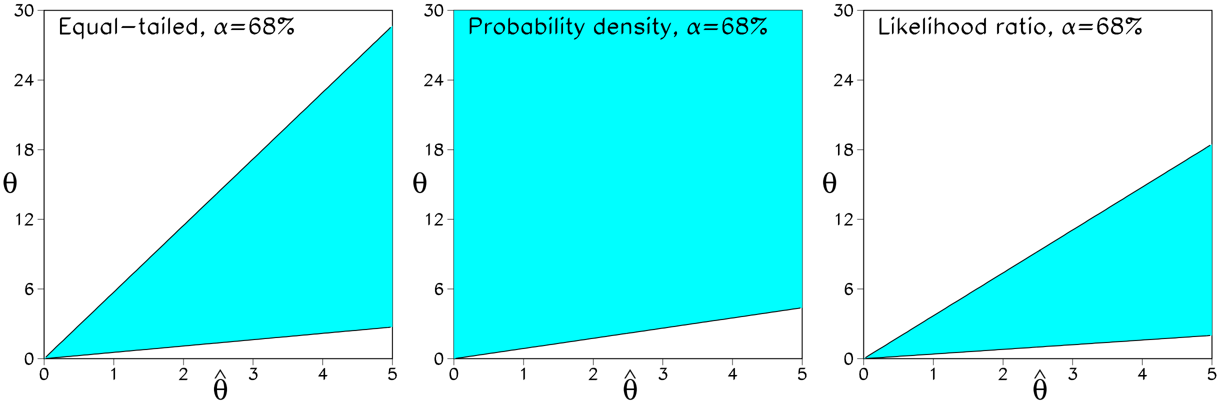

Figure 2: Confidence belts for an exponential lifetime, with $1-\alpha = 68\%$ *and three different ordering rules: (a) equal-tailed; (b) probability density; (c) likelihood ratio.*

It is interesting to note that the boundaries of the exponential intervals increase linearly with the observation $\hat{\theta}_{\rm obs}$. Since $\hat{\theta}$ is an unbiased estimator of θ, the expected lengths of these intervals are trivial to obtain: For $1-\alpha = 68\%$ they are $5.19\,\theta$, $2.59\,\theta$, and $3.29\,\theta$ for the

Table 1: Common ordering rules used in constructing frequentist intervals with confidence level $1-\alpha$. The first two columns give the name and defining equation for each rule, the third column shows the solution of the defining equation for the measurement of the lifetime θ of an exponential decay, and the last column applies the solution to the case $1-\alpha = 68\%$. The quantity $1-\alpha'$ at the bottom of the third column is the unique number between 0 and α that satisfies the equation $(1-\alpha'+1-\alpha)\ln(1-\alpha'+1-\alpha) = (1-\alpha')\ln(1-\alpha')$. For $1-\alpha = 0.68$, $1-\alpha' \approx 0.0829$. For the exponential-decay example, $k_{1-\alpha}(\theta) = \alpha/\theta$ and $k'_{1-\alpha}(\theta) = -e\,(1-\alpha')\ln(1-\alpha')$.

		Exponential decay example	
Ordering rule	Defining equation	General solution	Case $1-\alpha = 68\%$
Lower limit:	$F_{\theta_{low}}(\hat\theta_{\rm obs}) = 1-\alpha$	$\theta_{low} = \frac{-\hat\theta_{\rm obs}}{\ln\alpha}$	$[0.88\,\hat\theta_{\rm obs},\ +\infty[$
Upper limit:	$F_{\theta_{up}}(\hat\theta_{\rm obs}) = \alpha$	$\theta_{up} = \frac{-\hat\theta_{\rm obs}}{\ln(1-\alpha)}$	$[0,\ 2.59\,\hat\theta_{\rm obs}]$
Equal tails:	$\begin{cases} F_{\theta_1}(\hat\theta_{\rm obs}) = 1-\frac{\alpha}{2} \\ F_{\theta_2}(\hat\theta_{\rm obs}) = \frac{\alpha}{2} \end{cases}$	$\begin{cases} \theta_1 = \frac{-\hat\theta_{\rm obs}}{\ln\left(\frac{\alpha}{2}\right)} \\ \theta_2 = \frac{-\hat\theta_{\rm obs}}{\ln\left(1-\frac{\alpha}{2}\right)} \end{cases}$	$[0.55\,\hat\theta_{\rm obs},\ 5.74\,\hat\theta_{\rm obs}]$
Prob. density:	$f_\theta(\hat\theta_{\rm obs}) \geq k_{1-\alpha}(\theta)$	Same as lower limit	
Likelihood ratio:	$\frac{f_\theta(\hat\theta_{\rm obs})}{\max_{\theta'} f_{\theta'}(\hat\theta_{\rm obs})} \geq k'_{1-\alpha}(\theta)$	$\begin{cases} \theta_1 = \frac{-\hat\theta_{\rm obs}}{\ln(1-\alpha')} \\ \theta_2 = \frac{-\hat\theta_{\rm obs}}{\ln(1-\alpha'+1-\alpha)} \end{cases}$	$[0.40\,\hat\theta_{\rm obs},\ 3.70\,\hat\theta_{\rm obs}]$

equal-tailed, upper-limit, and likelihood-ratio intervals, respectively. Some confidence belts for this measurement are plotted in figure 2.

3.1.4 Ingredient 4: the confidence level

The confidence level labels a family of intervals; some conventional values are 68%, 90%, and 95%. It is very important to understand that a confidence level does *not* characterise single intervals; it only characterises families of intervals. The following example illustrates this.

Example 3 (Mass of a new elementary particle) Suppose we wish to measure the mass θ of a new elementary particle, and assume for simplicity that our measurement x of this mass has a Gaussian distribution with unit variance. Thus, even though θ must be positive for physics reasons, measurement resolution effects can cause x to be negative. Before performing our measurement we decide that we will report a 68% CL likelihood-ratio ordered interval. However, a colleague of ours prefers to report an 84% CL upper limit. The measurement is then performed and yields $x = 0$, leading both of us to report the same numerical interval $[0.0, 0.99]$! This demonstrates that the same numerical interval can have two very different coverages (confidence levels), depending on which ensemble it is considered to belong to. ■

In the above example the frequentist coverages of both interval procedures agree exactly with their respective confidence levels. As mentioned in section 2, this agreement is not always possible, especially when the observations are discrete or when nuisance parameters are present. In general an interval construction is considered valid from the frequentist point of view if its coverage, as a function of the true value of the parameter of interest, is everywhere equal to, or larger than, the stated confidence level. If this is not the case one says that the construction *undercovers*. One way to verify the coverage characteristics of a given interval procedure is to plot the interval boundaries as a function of the observation. This yields a confidence belt, as in figure 1(b). For each value of the parameter θ, integrating the probability density of the observation between the confidence belt boundaries yields the coverage at that particular θ value.

3.2 Test inversion

As indicated in the previous subsection, the inferential core of Neyman's construction is the ordering rule; the rest is just a geometrical embedding to enforce the coverage constraint. The ordering rule itself can be viewed as the construction of a test for each physical value of the parameter; each such test has a different acceptance region in sample space. Therefore, if we have a proper frequentist test to start with, we can dispense with the rest of Neyman's construction and proceed directly to defining the confidence interval as the set of parameter values for which the acceptance region contains the observation. This is known as the *test-inversion method*. To fix the notation, suppose we are interested in a parameter $\theta \in \Theta$, and that for each allowed value θ_0 of θ we can construct a size α test of

$$H_0 : \ \theta \equiv \ \theta_0 \quad \text{versus} \quad H_1 : \ \theta \ < \ \theta_0 \,. \tag{3}$$

Consider then the set $C_{1-\alpha}$ of all the θ_0 values for which H_0 is accepted. This set depends on the observations and is therefore random. We have:

$$P\big[\theta_0 \in C_{1-\alpha}; \theta = \theta_0\big] \ = \ P\big[H_0 \text{ is accepted} \mid H_0\big] \ = \ 1 - \alpha \,. \tag{4}$$

Hence $C_{1-\alpha}$ is an $(1-\alpha)$ CL set for θ. To picture the shape of this set, note that if H_0 is rejected for a given θ_0, then, because of the form of H_1, all values of θ larger than θ_0 will also be rejected. Therefore the set $C_{1-\alpha}$ of accepted θ values will have an upper boundary θ_{up}. For the simple example of a Gaussian data point x with mean θ and known standard deviation σ, one can test equation (3) with the statistic $y \equiv \theta_0 - x$. Under H_0 this statistic has a Gaussian distribution with mean zero and standard deviation σ, and large observed values $y_{\rm obs}$ of y constitute evidence against H_0 in the direction of H_1. The p-value of test (3) is therefore

$$p(\theta_0) = \int_{y_{\rm obs}}^{\infty} \frac{e^{-\frac{1}{2}\left(\frac{y}{\sigma}\right)^2}}{\sqrt{2\pi}\,\sigma}\,{\rm d}y = \frac{1}{2}\Big[1 - {\rm erf}\Big(\frac{y_{\rm obs}}{\sqrt{2}\,\sigma}\Big)\Big] = \frac{1}{2}\Big[1 - {\rm erf}\Big(\frac{\theta_0 - x_{\rm obs}}{\sqrt{2}\,\sigma}\Big)\Big] \,. \tag{5}$$

Values of θ for which $p(\theta) \geq \alpha$ are accepted by the test and included in the confidence interval. The upper limit of the interval is the solution of $p(\theta_{up}) = \alpha$, which is $\theta_{up} = x_{\rm obs} + z_{1-\alpha}\,\sigma$, where $z_{1-\alpha} \equiv \sqrt{2}\,{\rm erf}^{-1}(1-2\alpha)$ is the $(1-\alpha)$-quantile of the standard normal distribution (a Gaussian with zero mean and unit standard deviation).

If one is interested in an $(1-\alpha)$ CL lower limit θ_{low} on θ, the test to consider is

$$H_0 : \ \theta \equiv \ \theta_0 \quad \text{versus} \quad H_1' : \ \theta \ > \ \theta_0 \,, \tag{6}$$

again with size α. For the Gaussian example the result is $\theta_{low} = x_{\rm obs} - z_{1-\alpha}\,\sigma$.

An $(1-\alpha)$ CL two-sided interval for θ can be obtained by computing lower and upper limits at the $(1-\frac{\alpha}{2})$ CL, or by inverting a size α two-sided test:

$$H_0 :\ \theta \equiv\ \theta_0 \qquad \text{versus} \qquad H_1'' :\ \theta\ \neq\ \theta_0\,. \tag{7}$$

In this case, an appropriate test statistic for the Gaussian example is $y = |x - \theta_0|$, which has a folded Gaussian distribution. By solving the appropriate p-value equation as before, one obtains the $[\theta_{low}; \theta_{up}]$ with $\theta_{low} = x_{\rm obs} - z_{1-\frac{\alpha}{2}}\,\sigma$ and $\theta_{up} = x_{\rm obs} + z_{1-\frac{\alpha}{2}}\,\sigma$.

It should be clear from the above discussion that the construction of confidence intervals by this method requires the inversion of a *family* of tests rather than of a single test. Thus the method will not work if one has a nice test for a special value of the parameter of interest, but the test does not generalise to other values. It may also happen that inversion of a family of tests results in a union of disjoint intervals rather than a single interval. In general one can expect the properties of a family of tests to be reflected in the properties of the resulting intervals: Conservative tests lead to wide intervals, and powerful tests to narrow intervals.

3.3 Pivoting

A pivot is a function $Q(\theta, \mathbf{x})$ of both the observation $\mathbf{x}$ and the parameter θ whose distribution does not depend on any unknown parameters (not even on θ). Because of this special property, it is in principle possible to find, for any $\alpha \in [0, 1]$, constants $a(\alpha)$ and $b(\alpha)$ such that $P[a(\alpha) \leq Q(\theta, \mathbf{x}) \leq b(\alpha)] = 1 - \alpha$ for all θ. Therefore, the set of observations X such that $a(\alpha) \leq Q(\theta, \mathbf{x}) \leq b(\alpha)$ can be interpreted as the acceptance region of a size α test of the hypothesis that θ is the true value. Since this acceptance region is by construction valid for testing any θ, inverting the test leads to an $(1-\alpha)$ CL set for θ. This is the *pivoting method* for constructing confidence sets. In general there is no guarantee that such sets will be simple intervals, or that they will be optimal in any sense, but the simplicity of the construction makes it worth trying in situations that allow it. Furthermore, the concept of pivot is crucial to the development of the frequentist theory of confidence intervals beyond the setting where exact solutions can be found. This will become clear in section 3.4 on asymptotic approximations and section 3.5 on the bootstrap.

3.3.1 Gaussian means and standard deviations

To illustrate the pivoting method, consider the example of n measurements x_i from a Gaussian distribution with mean θ and standard deviation σ. This example has a particularly rich pivot structure. Suppose first that θ is unknown and σ known. In this case the quantity

$$Q_1(\theta, \mathbf{x})\ \equiv\ \frac{\bar{x} - \theta}{\sigma/\sqrt{n}}\,, \qquad \text{where} \qquad \bar{x}\ \equiv\ \frac{1}{n}\sum_{i=1}^{n} x_i\,, \tag{8}$$

has a standard normal distribution and is therefore a pivot. Thus, writing z_γ for the corresponding γ-quantile, we have:

$$\begin{aligned} 1-\alpha &= P\Big[z_{\frac{\alpha}{2}} \leq Q_1(\theta, \mathbf{x}) \leq z_{1-\frac{\alpha}{2}} \,\Big|\, \theta, \sigma\Big], & (9)\\ &= P\Big[z_{\frac{\alpha}{2}} \leq \frac{\bar{x}-\theta}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}} \,\Big|\, \theta, \sigma\Big], & (10)\\ &= P\Big[\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \,\Big|\, \theta, \sigma\Big], & (11)\\ &= P\Big[\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \theta \leq \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \,\Big|\, \theta, \sigma\Big], & (12) \end{aligned}$$

where in the last line we used the symmetry of the unit Gaussian distribution to write $z_\gamma = -z_{1-\gamma}$. This result shows that we have obtained a symmetric confidence interval for θ. Setting for example $1-\alpha = 68\%$ yields $z_{1-\frac{\alpha}{2}} = z_{0.84} = 1$, and the confidence interval after observing $\bar{x} = \bar{x}_{obs}$ is simply $\bar{x}_{obs} \pm \sigma/\sqrt{n}$.

A pivot is not necessarily unique or optimal. Consider for instance the case where θ is known and σ unknown. In principle we could use pivot (8) again, this time to construct confidence intervals for the variance σ^2. Solving

$$z_{\frac{\alpha}{2}} \leq \frac{\bar{x}_{obs}-\theta}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}} \tag{13}$$

for σ^2 yields an $(1-\alpha)$ CL lower limit

$$\sigma^2 \;\geq\; \frac{n(\bar{x}_{obs}-\theta)^2}{\chi^2_{1,1-\alpha}}, \tag{14}$$

where $\chi^2_{n,1-\alpha}$ is the $(1-\alpha)$-quantile of a χ^2 for n degrees of freedom, and we used the relation $z_{1-\alpha} = \sqrt{\chi^2_{1,1-2\alpha}}$.

We can then take advantage of the fact that the interval between two lower limits, one at the $(1-\frac{\alpha}{2})$ CL and the other at the $(\frac{\alpha}{2})$ CL, is itself an $(1-\alpha)$ CL equal-tailed two-sided interval, to obtain

$$\frac{n(\bar{x}_{obs}-\theta)^2}{\chi^2_{1,1-\frac{\alpha}{2}}} \;\leq\; \sigma^2 \;\leq\; \frac{n(\bar{x}_{obs}-\theta)^2}{\chi^2_{1,\frac{\alpha}{2}}} \quad \text{with confidence } 1-\alpha\ . \tag{15}$$

However, this interval is not optimal because it is based on a rather poor estimator of σ^2, namely $n\,(\bar{x}-\theta)^2$, which has a variance of $2\sigma^4$. In contrast, the usual estimator

$$S^2_{\theta,n} \;\equiv\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\theta)^2 \tag{16}$$

has variance $2\sigma^4/n$. Helpfully, the quantity $n\,S^2_{\theta,n}/\sigma^2$ is pivotal with a χ^2_n distribution and can therefore serve to construct intervals for σ. We have:

$$P\Big[\chi^2_{n,\frac{\alpha}{2}} \leq \sum_{i=1}^{n}\Big(\frac{x_i-\theta}{\sigma}\Big)^2 \leq \chi^2_{n,1-\frac{\alpha}{2}} \,\Big|\, \theta, \sigma\Big] \;=\; 1-\alpha\,, \tag{17}$$

so that

$$\frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\chi^2_{n,1-\frac{\alpha}{2}}} \;\leq\; \sigma^2 \;\leq\; \frac{\sum_{i=1}^{n}(x_i-\theta)^2}{\chi^2_{n,\frac{\alpha}{2}}} \quad \text{with confidence } 1-\alpha\ . \tag{18}$$

Note that for $n = 1$ this interval coincides with that given in (15). However, at larger values of n, the interval (18) has smaller expected length.

Finally, we consider the case where both θ and σ are unknown. If we are interested in σ^2, we can use

$$S_n^2 \;\equiv\; \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{19}$$

as estimator, and construct intervals based on the fact that

$$Q_2(\sigma, \mathbf{x}) \;\equiv\; (n-1)\, S_n^2/\sigma^2 \tag{20}$$

is a pivot following a χ^2 distribution function with $n-1$ degrees of freedom. If on the other hand θ is of interest, a new pivot can be constructed by taking the ratio of $Q_1(\theta, \mathbf{x})$ and $\sqrt{Q_2(\sigma, \mathbf{x})/(n-1)}$:

$$Q_3(\theta, \sigma, \mathbf{x}) \;=\; \frac{\sqrt{n}(\bar{x}-\theta)/\sigma}{\sqrt{S_n^2/\sigma^2}} \;=\; \frac{\bar{x}-\theta}{S_n/\sqrt{n}}\,. \tag{21}$$

This ratio is distributed as a central t variate[5] for $n-1$ degrees of freedom. With $t_{n-1,1-\alpha}$ the $(1-\alpha)$-quantile of Student's t_{n-1} distribution and $(\bar{x}_{obs}, s_n)$ the observed value of $(\bar{x}, S_n)$, the following is a $(1-\alpha)$ CL equal-tailed, symmetric interval for θ:

$$\bar{x} - \frac{s_n}{\sqrt{n}}\, t_{n-1,1-\frac{\alpha}{2}} \;\leq\; \theta \;\leq\; \bar{x} + \frac{s_n}{\sqrt{n}}\, t_{n-1,1-\frac{\alpha}{2}}\,. \tag{22}$$

For $1-\alpha = 68\%$ one finds $t_{1,1-\frac{\alpha}{2}} = 1.82$ and $t_{19,1-\frac{\alpha}{2}} = 1.02$. As n becomes large, the t_{n-1} quantiles converge to the corresponding unit Gaussian quantiles.

3.3.2 Exponential lifetimes

Although the previous subsection is limited to problems involving the Gaussian distribution, the pivoting method can be applied to many other situations. In fact, a pivot that is often available is the *cumulative distribution* of the data (cdf), viewed as a function of the data and the parameters. For continuous data this pivot is uniformly distributed. We illustrate this idea with the construction of confidence intervals on the lifetime τ associated with an exponential decay. The cdf of the measurement t, and hence the pivot, is $Q(\tau, t) = 1 - e^{-t/\tau}$. Since this is a uniform pivot, an $(1-\alpha)$ CL interval for τ is given by

$$\frac{\alpha}{2} \;\leq\; 1 - e^{-t/\tau} \;\leq\; 1 - \frac{\alpha}{2} \tag{23}$$

or

$$\frac{t}{-\ln\left(\frac{\alpha}{2}\right)} \;\leq\; \tau \;\leq\; \frac{t}{-\ln\left(1-\frac{\alpha}{2}\right)}\,. \tag{24}$$

For $1-\alpha = 68\%$ this yields $\tau \in [0.55t, 5.74t]$, which is the equal-tailed interval listed in table 1.

[5] A variate is a random variable.

3.3.3 Binomial efficiencies

For discrete data the cdf is no longer an exact pivot, but it can still be used to construct confidence sets. As an example we consider the measurement of an efficiency ϵ based on the observation of x successes out of n trials. The cdf is binomial, but it will be convenient to express it in terms of a Beta cdf $B(x; a, b)$,

$$\begin{aligned} P[K \leq x; \epsilon, n] &= \sum_{k=0}^{x} \binom{n}{k} \epsilon^k (1-\epsilon)^{n-k} \\ &= \int_0^{1-\epsilon} \frac{t^{n-x-1}(1-t)^x}{B(x+1, n-x)}\,\mathrm{d}t \;=\; B(1-\epsilon; n-x, x+1)\,, \end{aligned} \tag{25}$$

where $B(a,b) \equiv \Gamma(a)\Gamma(b)/\Gamma(a+b)$. The second equality can be derived by integration by parts. Although the cdf of a binomial variable K is not an exact pivot, it can be made exact by introducing a random variable U that is uniform on the interval $[0,1[$, and considering the cdf of the sum $K+U$. This is a continuous cdf and therefore an exact uniform pivot, so that the following inequalities define an $(1-\alpha)$ CL interval for ϵ:

$$\frac{\alpha}{2} \;\leq\; P[K+U \leq x; \epsilon, n] \;\leq\; 1 - \frac{\alpha}{2}\,. \tag{26}$$

Since U is unobserved, solving for ϵ requires a worst-case analysis, in which the random variable U is replaced by a constant such that the inequalities in formula (26) hold regardless of the value of U. For the lower limit this means that U should be replaced by 0:

$$\frac{\alpha}{2} \;\leq\; P[K+U \leq x; \epsilon, n] \;\leq\; P[K \leq x; \epsilon, n] \;=\; P\big[B_{n-x,x+1} \leq 1-\epsilon\big]\,, \tag{27}$$

where we used equation (25) and $B_{a,b}$ is a random variable with a Beta(a,b) distribution. Writing $B_{a,b,\alpha}$ for the (α)-quantile of this distribution, the above result implies that

$$1 - \epsilon \;\geq\; B_{n-x,x+1,\frac{\alpha}{2}}\,, \tag{28}$$

which yields the upper limit ϵ_{up} of the desired interval for ϵ:

$$\epsilon \;\leq\; 1 - B_{n-x,x+1,\frac{\alpha}{2}} \;=\; B_{x+1,n-x,1-\frac{\alpha}{2}} \;\equiv\; \epsilon_{up}\,. \tag{29}$$

This expression is undefined for $x = n$, in which case we set $\epsilon_{up} = 1$. In order to guarantee the upper inequality in (26), we must replace the unobserved random variable U by the constant 1. Similar manipulations as above yield the lower limit ϵ_{low} of the interval:

$$\epsilon \;\geq\; B_{x,n-x+1,\frac{\alpha}{2}} \;\equiv\; \epsilon_{low}\,. \tag{30}$$

For $x = 0$ we set $\epsilon_{low} = 0$. The interval $[\epsilon_{low}, \epsilon_{up}]$, with endpoints given in equations (29) and (30), is known as an $(1-\alpha)$ CL Clopper–Pearson interval for the efficiency ϵ [CP34]. This interval is easy to code into a computer program, using the incomplete beta function [Pre+07]. It can also be computed from tables of Snedecor's F distribution and the following relationship between Beta and F quantiles:

$$B_{a,b,\gamma} \;=\; \left[1 + \frac{b}{a}\,F_{2b,2a,1-\gamma}\right]^{-1}. \tag{31}$$

Because of the worst-case analysis involving the random variable U, Clopper–Pearson intervals are conservative (they overcover). Overcoverage is generally unavoidable in discrete sample spaces. An in-depth comparison with other constructions can be found in [CHT10].

3.3.4 Poisson means

Another frequent application in physics is the computation of an upper limit on the expected mean θ of the number of events of a new signal at a collider experiment. The observation is a number of events n, which is assumed to be Poisson distributed with mean $\theta + \nu$, where ν is a known background contamination. The cdf is again discrete and therefore not an exact pivot, but we can make it exact by adding a uniform variate U on $[0, 1[$ to the Poisson variate N. Thus the set of θ satisfying $\alpha \leq P[N + U \leq n; \theta + \nu]$ is an exact $(1 - \alpha)$ CL interval. Repeating the worst-case analysis argument of the previous section, we replace U by the constant 0 to obtain a conservative interval. This is a one-sided interval bounded by an upper limit, as we now show. First note that

$$\alpha \;\leq\; P[N \leq n; \theta + \nu] \;=\; \sum_{k=0}^{n} \frac{(\theta + \nu)^k \, e^{-\theta-\nu}}{k!} \;=\; \int_{\theta+\nu}^{+\infty} \frac{t^n \, e^{-t}}{\Gamma(n+1)} \mathrm{d}t\,, \tag{32}$$

where the last equality can be proved by integration by parts. After substituting $t = z/2$ in the integrand on the right, one recognises this as the cdf of a χ^2 variate $\chi^2_{2(n+1)}$ for $2(n + 1)$ degrees of freedom. With some rearrangement the above inequality can therefore be rewritten as

$$P[\chi^2_{2(n+1)} \leq 2(\theta + \nu)] \;\leq\; 1 - \alpha\,, \tag{33}$$

which implies that $2(\theta + \nu)$ is smaller than the $(1 - \alpha)$-quantile of the variate $\chi^2_{2(n+1)}$. Hence the following is an $(1 - \alpha)$ CL upper limit on θ:

$$\theta \;\leq\; \frac{1}{2}\chi^2_{2(n+1),1-\alpha} - \nu. \tag{34}$$

This result was first reported by Garwood in 1936 [Gar36]. For an $(1 - \alpha)$ CL two-sided interval, similar calculations yield

$$\left[\frac{1}{2}\chi^2_{2n,\frac{\alpha}{2}} - \nu,\; \frac{1}{2}\chi^2_{2(n+1),1-\frac{\alpha}{2}} - \nu\right]\,. \tag{35}$$

For $n = 0$ the lower limit of the interval is $-\nu$. Some numerical examples of equations (34) and (35), for $\nu = 0$, are shown in table 2. Also shown there are *Feldman–Cousins intervals*, which are based on a likelihood-ratio ordering rule [FC98]. A significant advantage of Feldman–Cousins intervals is that they are never unphysical, regardless of how large the background contamination ν is. This is not the case for Garwood intervals. The frequentist coverage of 68% CL Garwood central intervals is plotted in figure 3 (again with $\nu = 0$). Viewed as a function of the true Poisson mean, the coverage is highly discontinuous due to the discreteness of the Poisson distribution.

Table 2: Frequentist interval constructions for the mean of a Poisson distribution when N events are observed: 95% CL upper limits (column 2), 68% CL equal-tailed intervals (column 3), and 95%- and 68% CL Feldman–Cousins intervals (columns 4 and 5, from [FC98]).

	Garwood		Feldman–Cousins	
	Upper limit	Equal-tailed interval		
N	95% CL	68% CL	95% CL	68% CL
0	3.00	[0.00, 1.84]	[0.00, 3.09]	[0.00, 1.29]
1	4.74	[0.17, 3.30]	[0.05, 5.14]	[0.37, 2.75]
2	6.30	[0.71, 4.64]	[0.36, 6.72]	[0.74, 4.25]
3	7.75	[1.37, 5.92]	[0.82, 8.25]	[1.10, 5.30]
4	9.15	[2.09, 7.16]	[1.37, 9.76]	[2.34, 6.78]
5	10.51	[2.84, 8.38]	[1.84, 11.26]	[2.75, 7.81]
6	11.84	[3.62, 9.58]	[2.21, 12.75]	[3.82, 9.28]
7	13.15	[4.43, 10.77]	[2.58, 13.81]	[4.25, 10.30]
8	14.45	[5.23, 11.95]	[2.94, 15.29]	[5.30, 11.32]
9	15.71	[6.06, 13.11]	[4.36, 16.77]	[6.33, 12.79]
10	16.96	[6.89, 14.27]	[4.75, 17.82]	[6.78, 13.81]

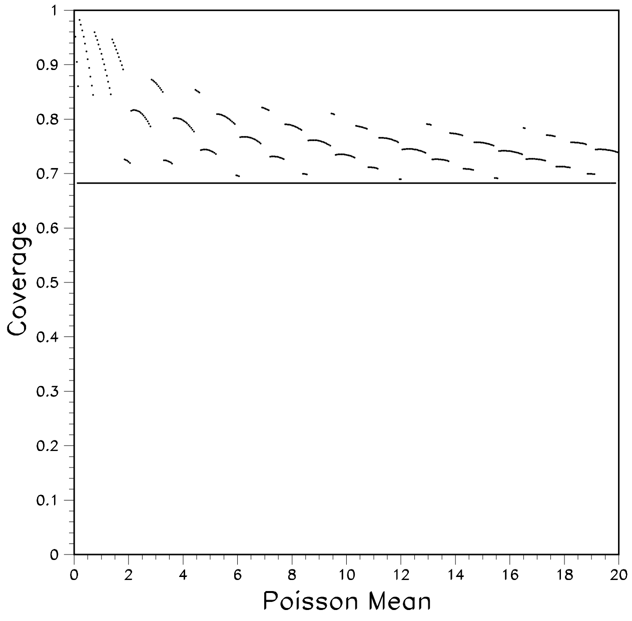

Figure 3: Frequentist coverage of 68% CL Garwood central intervals for the mean of a Poisson distribution. The coverage is evaluated in increments of 0.1 *in the Poisson mean, and the nominal coverage of the construction is indicated by the solid horizontal line.*

3.4 Asymptotic approximations

In section 3.1.3 we mentioned the likelihood-ratio ordering rule as an option for the construction of Neyman confidence sets. Given data x, a parameter of interest θ and its maximum-likelihood estimate (MLE) $\hat{\theta} = \hat{\theta}(x)$, this rule includes in the confidence set any θ value that is not rejected by an α-level test based on the likelihood ratio $\lambda(x;\theta) \equiv L(x;\theta)/L(x;\hat{\theta})$. For large samples it turns out that the confidence level constraint is easy to implement thanks to Wilks's theorem [Wil38]. The latter states that, under standard regularity conditions, $-2\ln\lambda(x;\theta)$ is asymptotically distributed as a χ^2 variate for d degrees of freedom, where d equals the dimensionality of θ (in the terminology of section 3.3, $-2\ln\lambda(x;\theta)$ is an asymptotic pivot). This provides a simple way to construct a $(1-\alpha)$ CL interval, by taking the set of θ values for which

$$-2\ln\lambda(x;\theta) \;\leq\; \chi^2_{d,1-\alpha}\,, \tag{36}$$

where $\chi^2_{d,1-\alpha}$ is the $(1-\alpha)$-quantile of a χ^2 distribution for d degrees of freedom. Thus, if θ is one-dimensional, use $\chi^2_{1,0.68} \approx 1$ for a 68% CL interval, $\chi^2_{1,0.95} \approx 4.00$ for a 95% CL interval, etc.

If nuisance parameters are present, collectively labeled ν, the same result applies provided the likelihood ratio is defined by

$$\lambda(x;\theta) \;\equiv\; \frac{L(x;\theta,\hat{\hat{\nu}}(\theta))}{L(x;\hat{\theta},\hat{\nu})}\,, \tag{37}$$

where $\hat{\hat{\nu}}(\theta)$ is the profile likelihood estimate of ν, that is, its MLE evaluated at a fixed value of θ, and $\hat{\nu}$ is the global MLE of ν, without constraining to a fixed θ.

For one-dimensional θ it is often helpful to plot a graph of $-2\ln\lambda(x;\theta)$ versus θ, since this allows the interval to be determined at various confidence levels, and to assess the "Gaussianity" of the problem, e.g. whether 95% CL intervals have twice the length of 68% CL intervals. In addition, it may happen that confidence sets obtained by this method consist of two or more disjoint intervals, in which case a plot is most useful. For a two-dimensional parameter vector $\boldsymbol{\theta}$ one can plot contours of $-2\ln\lambda(x;\boldsymbol{\theta})$ in the plane of $\boldsymbol{\theta}$ values. Then for example, the contour corresponding to $-2\ln\lambda(x;\boldsymbol{\theta}) = \chi^2_{2,0.68} \approx 2.30$ encloses a 68% confidence region for $\boldsymbol{\theta}$, and for 95% confidence one should use $\chi^2_{2,0.95} \approx 6.18$. In high energy physics these constructions are typically done with the help of the routine *Minos* in the MINUIT program package [JR75]. A general treatment of likelihood asymptotics for high energy physics can be found in [Cow+11].

3.5 Bootstrapping

The confidence interval constructions we have examined so far all assume that it is possible to write down explicitly the probability distribution of the data in analytical form, including its dependence on the parameter of interest and on nuisance parameters. Unfortunately this is not always the case in high energy physics. A good example is the measurement of the top quark mass, where the dependence of data distributions on the parameter of interest is buried deeply in complex Monte Carlo simulations of physics processes and detector responses. The bootstrap method provides a powerful way to circumvent this difficulty. It is a bridge between exact methods, which cannot be used in complex physics analyses, and asymptotic methods, which lack coverage accuracy in finite samples. There are two ideas at the core of bootstrap methods [DHY03]: the *plug-in principle* and *resampling.* The plug-in principle sounds rather obvious as it states that in order to estimate a quantity of interest, one should replace the unknown data cdf F by an estimate $\hat{F}$. If nothing is known a priori about F, and all we have is a data sample $x_1, x_2, \ldots, x_n$, then F can be estimated by the *empirical* distribution of the data, which assigns probability $1/n$ to each data point:

$$\hat{F}(x) \; = \; \frac{\#\{x_i \leq x\}}{n}\,. \tag{38}$$

Another possibility is that F has a known functional form that depends on some unknown parameter ψ, in which case it could be estimated by substituting the maximum-likelihood estimate (MLE) $\hat{\psi}$ for ψ.

Suppose now that we are interested in estimating a quantity θ, which could be something as simple as the mean of a population characteristic or as complex as the mass of the top quark. The true value of θ is defined as the result of applying the appropriate estimating procedure to the true distribution, which we write as $\theta = \theta(F)$, whereas the plug-in estimate of θ is obtained by applying the same procedure to the estimated distribution, $\hat{\theta} = \theta(\hat{F})$. For example when θ is a mean and $\hat{F}$ is an empirical distribution we have:

$$\theta = \theta(F) = \int x \, \mathrm{d}F(x) \quad \text{and} \quad \hat{\theta} = \theta(\hat{F}) = \int x \, \mathrm{d}\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i\,. \tag{39}$$

Thus, quantities of interest should be viewed as functionals of distributions, or as outcomes of procedures or algorithms applied to distributions.

The second core idea of the bootstrap is resampling, whereby difficult analytical calculations are replaced by simulations. Resampling comes in two versions, *parametric* and *non-parametric*. In the parametric version it is assumed that the distribution F of the data is known up to some parameter ψ. An estimate of ψ, typically the MLE, is then substituted in the expression for F in order to allow the generation of random data samples. In non-parametric resampling no assumption is made about the form of F. Instead, the data $x_1, \ldots, x_n$ themselves are used to approximate statistical fluctuations according to F. This is done by *resampling with replacement* from the set $\{x_1, \ldots, x_n\}$: for each resample, n data points are successively selected at random from this set, and each selected data point is "put back" in the set before selecting the next one. Thus, some of the original data points will appear more than once in a resampled data set, and some will not appear at all. A (parametric or non-parametric) resampled data set is often called a *bootstrap sample.*

There exists a bewildering array of bootstrap methods for computing confidence intervals [CB00]. These methods can be broadly classified in three categories: pivotal, non-pivotal, and test inversion. Section 3.5.1 discusses the bootstrap-t interval as an example of pivotal methods, and makes the important point that the best way to improve the theoretical coverage accuracy of an interval is to bootstrap a pivot (or an asymptotic pivot). Unfortunately theoretical coverage accuracy is not everything, and other important considerations lead to the definition of the non-pivotal percentile intervals in section 3.5.2, first in a "simple" version and then an improved, "automatic" version that incorporates a test-inversion technique. Finally, section 3.5.3 describes a calibration procedure that can improve the coverage accuracy of any confidence interval construction.

3.5.1 The bootstrap-t interval

If we have a data set $\{x_1, \ldots, x_n\}$ from which we can derive an estimate $\hat{\theta}$ of the parameter of interest θ, as well as an estimate $\hat{\sigma}$ of the standard deviation of $\hat{\theta}$, then we can form an $(1-\alpha)$ CL *standard interval* for θ,

$$\left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\,\hat{\sigma},\; \hat{\theta} - z_{\frac{\alpha}{2}}\,\hat{\sigma}\right], \tag{40}$$

where z_γ is the γ-quantile of the unit Gaussian distribution. If $\hat{\theta}$ is asymptotically normal, and the estimators $\hat{\theta}$ and $\hat{\sigma}$ are *consistent*, then the asymptotic coverage of the standard interval is $1-\alpha$. In finite samples the actual coverage,

$$P\left[\hat{\theta} - z_{1-\frac{\alpha}{2}}\,\hat{\sigma} \;\leq\; \theta \;\leq\; \hat{\theta} - z_{\frac{\alpha}{2}}\,\hat{\sigma}\right] \;=\; P\left[z_{\frac{\alpha}{2}} \;\leq\; \frac{\hat{\theta}-\theta}{\hat{\sigma}} \;\leq\; z_{1-\frac{\alpha}{2}}\right], \tag{41}$$

typically differs from $1-\alpha$ by a term of order n^{-1}, where n is the sample size. The above expression suggests that one way to reduce this difference would be to correct the $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ coefficients by bootstrapping the quantity

$$t \;\equiv\; \frac{\hat{\theta}-\theta}{\hat{\sigma}}. \tag{42}$$

The idea is to simulate the distribution of t and replace $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ by the corresponding quantiles of this t distribution. Since we do not know the true value of θ, we need to apply the plug-in principle: replace θ by its estimate $\hat{\theta}$, and $\hat{\theta}$ and $\hat{\sigma}$ by their bootstrapped estimates $\hat{\theta}^\star$ and $\hat{\sigma}^\star$ (the $\star$ superscript is a conventional way to indicate a bootstrapped quantity). The following pseudo-code illustrates the calculation:

```
1. Obtain $\hat{\theta} = \theta(\hat{F})$ and $\hat{\sigma} = \sigma(\hat{F})$ from the original data set $\{x_1, \ldots, x_n\}$.
2. For $i = 1$ to $b$:
3.        Generate $\{x^\star_{i1}, \ldots, x^\star_{in}\}$ from $\hat{F}$ to obtain $\hat{F}^\star_i$.
4.        Compute $\hat{\theta}^\star_i = \theta(\hat{F}^\star_i)$ and $\hat{\sigma}^\star_i = \sigma(\hat{F}^\star_i)$.
5.        Set $t^\star_i = \frac{\hat{\theta}^\star_i - \hat{\theta}}{\hat{\sigma}^\star_i}$.
6. Estimate the bootstrap quantiles $t^\star_{[\frac{\alpha}{2}]}$ and $t^\star_{[1-\frac{\alpha}{2}]}$ from the sample of $t^\star_i$.
```

The $(1-\alpha)$ CL bootstrap-t interval for θ is then given by

$$\left[\hat{\theta} - t^\star_{[1-\frac{\alpha}{2}]}\,\hat{\sigma},\; \hat{\theta} - t^\star_{[\frac{\alpha}{2}]}\,\hat{\sigma}\right]. \tag{43}$$

The quantiles $t^\star_{[\gamma]}$ at step 6 can be estimated by taking $t^\star_{[\gamma]} = t^\star_{(k)}$, where $k = \gamma b$ and $t^\star_{(k)}$ is entry number k in the list of sorted bootstrap values $t^\star_{(1)} \leq t^\star_{(2)} \leq \cdots \leq t^\star_{(b)}$. If k is not integer, a linear interpolation can be used:

$$t^\star_{(k)} \;=\; t^\star_{(k')} + (k-k')\big(t^\star_{(k'+1)} - t^\star_{(k')}\big),. \tag{44}$$

Here k' is the largest integer smaller than k. An appropriate value for the number of bootstrap replications b is typically 1000 for confidence interval estimation. Note that, unlike the standard interval (40), the bootstrap-t interval (43) is not necessarily symmetric around $\hat{\theta}$. This asymmetry contributes to the better coverage of the bootstrap-t interval, which typically differs from nominal by a term of order n^{-2}. Although this constitutes a theoretical improvement with respect to the standard interval, it is important to distinguish theoretical from numerical accuracy. In particular, if the estimated standard deviation $\hat{\sigma}$ in equation (42) has itself a large variance, the actual numerical accuracy of the interval (43) may not be much better than that of the standard interval (40). Furthermore, the bootstrap-t interval is primarily designed to work for location parameters such as the mean or median of a sample; it does not work well for parameters such as a standard deviation or correlation coefficient. This is because of the form of the bootstrapped quantity (42), which is a pivot for location parameters but not for scale parameters or correlations. The bootstrap-t interval shares a couple of other disadvantages with the standard interval: It does not respect physical boundaries and is not equivariant under parameter transformations. For all these reasons we now turn to percentile intervals.

3.5.2 Percentile intervals

The endpoints of the standard interval (40) can be reinterpreted in terms of percentiles of the distribution of the bootstrap estimates $\hat{\theta}^{\star}_i$. Indeed, under the conditions of validity of that interval, the $\hat{\theta}^{\star}_i$ are normal with mean $\hat{\theta}$ and standard deviation $\hat{\sigma}$, so that

$$P\left[\hat{\theta}^{\star} \leq \hat{\theta} - z_{1-\frac{\alpha}{2}}\hat{\sigma}\right] \;=\; P\left[\frac{\hat{\theta}^{\star} - \hat{\theta}}{\hat{\sigma}} \leq -z_{1-\frac{\alpha}{2}}\right] \;=\; P\left[\frac{\hat{\theta}^{\star} - \hat{\theta}}{\hat{\sigma}} \leq z_{\frac{\alpha}{2}}\right] \;=\; \frac{\alpha}{2}\,. \tag{45}$$

This suggests the following definition of an $(1-\alpha)$ CL bootstrap interval for θ:

$$\left[\hat{\theta}^{\star}_{[\frac{\alpha}{2}]},\, \hat{\theta}^{\star}_{[1-\frac{\alpha}{2}]}\right], \tag{46}$$

where $\hat{\theta}^{\star}_{[\gamma]}$ is the γ-quantile of the distribution of bootstrap estimates $\hat{\theta}^{\star}_i$. This interval is known as the *simple percentile interval*. Its endpoints are quantiles, making it equivariant under parameter transformations. Thus if $\hat{\theta}$ itself is not distributed according to a Gaussian, but a transformation to a Gaussian exists, the percentile method will be able to take advantage of this in producing an interval with accurate coverage. Another advantage of simple percentile intervals is that they respect physical boundaries on the parameter provided the estimator does so. On the other hand, the construction of this interval is based on the quantity $\hat{\theta}$, which is generally not a pivot, not even asymptotically. Therefore its coverage accuracy, outside of the special case just mentioned, is not better than that of the standard interval, of order n^{-1}.

Several methods have been developed to improve the coverage properties of simple percentile intervals [ET93], some of them requiring non-trivial analytical calculations. Here we focus on one method, known as the *automatic percentile bootstrap* because it does not require such calculations [DR95]. Suppose that $F(\hat{\theta};\theta)$ is the cumulative distribution of the plug-in estimate $\hat{\theta}$. Having observed $\hat{\theta} = \hat{\theta}_{\text{obs}}$, an exact, $(1-\alpha)$ CL equal-tailed interval $[\theta_1, \theta_2]$ for

θ can be obtained by solving the equations

$$F(\hat{\theta}_{\rm obs};\theta_1) \;=\; 1-\frac{\alpha}{2} \quad\text{and}\quad F(\hat{\theta}_{\rm obs};\theta_2) \;=\; \frac{\alpha}{2} \tag{47}$$

(see for example table 1 in section 3.1.3). In the automatic percentile method the solution to these equations is approximated with a bootstrap simulation. Taking the first equation as example, one chooses a starting value for θ_1, bootstraps the corresponding distribution of $\hat{\theta}$, computes its $(1-\frac{\alpha}{2})$-quantile and adjusts θ_1 until that quantile equals $\hat{\theta}_{\rm obs}$. A good starting value for θ_1 would be the output of the simple percentile algorithm. If the cdf F depends on nuisance parameters ν, the latter should be replaced by their profile likelihood estimate $\hat{\hat{\nu}}(\theta)$ when performing the bootstrap (as in section 3.4).

As defined above, the automatic percentile interval is equivariant under reparameterisation, respects physical boundaries provided $\hat{\theta}$ does and has the same coverage accuracy as bootstrap-t intervals, i.e. $\mathcal{O}(n^{-2})$.

3.5.3 Bootstrap calibration

The bootstrap can be used to recalibrate approximate interval constructions. In the case of an upper limit, for example, one first generates a large number of bootstrap samples in order to estimate the *calibration function*, which is the actual coverage $1-\alpha_{\rm true}$ of the upper limit as a function of its nominal coverage $1-\alpha_{\rm nom}$ (the same set of bootstrap samples is used at each $1-\alpha_{\rm nom}$ value). The recalibrated upper limit is then the upper limit computed with the $1-\alpha_{\rm nom}$ corresponding to the desired $1-\alpha_{\rm true}$. However, this is still an approximation since the calibration function was determined by a bootstrap method. In principle one could recalibrate the calibrated limit and obtain an even better result, but such calculations quickly become very complex.

Typical candidates for recalibration are the standard interval (40) and the percentile interval (46). In the latter case the recalibration procedure amounts to a double bootstrap.

3.6 Nuisance parameters

In principle the Neyman construction can be performed when there is more than one parameter; it simply becomes a multi-dimensional construction, and the confidence belt becomes a "hyperbelt". If some parameters are nuisances, they can be eliminated by projecting the final confidence region onto the parameter(s) of interest at the end of the construction. However, there are two difficulties: the conceptual one of designing an ordering rule that minimises the amount of overcoverage introduced by the projection [Pun06], and the more practical one of performing multi-dimensional constructions.

Several simpler, approximate solutions are available. We already discussed two of them: asymptotic approximations in section 3.4 and the bootstrap in section 3.5. A third approach inverts the order of the steps in the multi-dimensional Neyman construction: First eliminate the nuisance parameters ν from the pdf $f(x;\theta,\nu)$ of the data x and then perform a one-dimensional interval construction on the parameter of interest θ. The elimination step can be done by integration over a proper prior distribution $\pi(\nu)$:

$$f(x;\theta,\nu) \;\to\; f^{\dagger}(x;\theta) \;\equiv\; \int f(x;\theta,\nu)\,\pi(\nu)\,{\rm d}\nu\,. \tag{48}$$

Although this is clearly a Bayesian step, nothing prevents one from studying the frequentist properties of intervals derived from $f^\dagger$ [CH92; TC05].

Another possibility is to eliminate the nuisance parameters by profiling the pdf:

$$f(x;\theta,\nu) \;\to\; f^\ddagger(x;\theta) \;\equiv\; f(x;\theta,\hat{\hat{\nu}}(\theta))\,. \tag{49}$$

Here $\hat{\hat{\nu}}(\theta)$ is the profile MLE of ν, which maximises $f(x;\theta,\nu)$ at the observed value of x and at the given value of θ. Even though $\hat{\hat{\nu}}(\theta)$ depends on the data, the interval for θ is constructed under the assumption that, for a given θ, the true value of ν is known and equal to $\hat{\hat{\nu}}(\theta)$. In other words, $f^\ddagger(x;\theta)$ is treated as a properly normalised pdf for x [CL00; SWW09].

It is important to keep in mind that the coverage of the simpler solutions is not guaranteed. It must therefore be checked, at least at a few representative points of parameter space (in both θ and ν).

To illustrate various techniques for handling nuisance parameters we consider a slight generalisation of the background-subtraction problem analysed in section 3.3.4. We have measured a number of events n that follows a Poisson distribution with mean $\theta+\nu$, where θ is a signal of interest and ν a background contamination. Neither θ nor ν is known, but we have an auxiliary measurement of ν in the form of a Poisson-distributed number of events k, with mean $\tau\nu$, where τ is a known constant. The joint probability mass function of n and k is

$$f(n,k;\theta,\nu) \;=\; f_1(n;\theta,\nu)\,f_2(k;\nu) \;=\; \frac{(\theta+\nu)^n e^{-\theta-\nu}}{n!}\,\frac{(\tau\nu)^k e^{-\tau\nu}}{k!}\,. \tag{50}$$

The likelihood ratio for testing a given value of θ is

$$\lambda(n,k;\theta) \;=\; \frac{f(n,k;\theta,\hat{\hat{\nu}}(\theta))}{f(n,k;\hat{\theta},\hat{\nu})}\,, \tag{51}$$

where $(\hat{\theta},\hat{\nu})$ is the MLE of (θ,ν) and $\hat{\hat{\nu}}(\theta)$ is the profile MLE of ν. All these MLEs are constrained to be positive for physical reasons. Assuming we have observed $n = n_{\rm obs}$ and $k = k_{\rm obs}$, from which we can derive estimates $\hat{\theta}_{\rm obs}$, $\hat{\nu}_{\rm obs}$, and $\hat{\hat{\nu}}_{\rm obs}(\theta)$, one can consider the following eight methods for constructing an $(1-\alpha)$ CL interval for θ:

1. *Likelihood-ratio test inversion:* This is an "exact" frequentist method, in the sense that it never undercovers. The interval is defined as the set of θ values for which

$$\min_{\nu}\Big\{P\Big[-2\ln\lambda(N,K;\theta) \le -2\ln\lambda(n_{\rm obs},k_{\rm obs}\,;\,\theta)\;\Big|\;\theta,\nu\Big]\Big\} \;\le\; 1-\alpha\,. \tag{52}$$

The notation $P[E \mid \theta,\nu]$ indicates the probability of event E when $f(n,k \mid \theta,\nu)$ is the true distribution of (N,K). With $q_{1-\alpha}(\theta,\nu)$ the $(1-\alpha)$-quantile of the distribution of $-2\ln\lambda(N,K\,|\,\theta)$, equation (52) is equivalent to:

$$-\,2\ln\lambda(n_{\rm obs},k_{\rm obs}\,;\,\theta) \;\le\; q_{1-\alpha}(\theta) \;\equiv\; \max_{\nu}\, q_{1-\alpha}(\theta,\nu)\,. \tag{53}$$

The minimisation and maximisation in these interval definitions can be viewed as a kind of worst-case analysis that guarantees frequentist coverage, and possibly overcoverage, for all physical values of ν at a given θ.

2. *Naive method:* Given the relative computational complexity of the test-inversion method, it may be worthwhile to compare it to the following simple approach. Under model (50), the MLE of θ is $\hat{\theta}_{\text{obs}} = n_{\text{obs}} - k_{\text{obs}}/\tau$. Ignoring the physical constraint $\hat{\theta}_{\text{obs}} \geq 0$, the variance of this MLE is $\theta + \nu + \nu/\tau$, which can be estimated by $n_{\text{obs}} + k_{\text{obs}}/\tau^2$. Thus an approximate $(1-\alpha)$ CL interval for θ is given by the intersection of
$$\left[\hat{\theta}_{\text{obs}} - z_{1-\frac{\alpha}{2}}\sqrt{n_{\text{obs}} + \frac{k_{\text{obs}}}{\tau^2}},\ \hat{\theta}_{\text{obs}} + z_{1-\frac{\alpha}{2}}\sqrt{n_{\text{obs}} + \frac{k_{\text{obs}}}{\tau^2}}\right], \tag{54}$$
with the physical region $\theta \geq 0$, where z_γ is the γ-quantile of the standard normal distribution.

3. *Asymptotic likelihood-ratio test:* This is the method described in section 3.4: a test inversion as in equation (53), but with $q_{1-\alpha}(\theta)$ approximated by the $(1-\alpha)$-quantile $\chi^2_{1,1-\alpha}$ of a χ^2 distribution for one degree of freedom.

4. *Bayesian elimination:* Here the auxiliary measurement $f_2(k;\nu)$ is replaced by a prior $\pi(\nu)$ for ν. With proper normalisation this is:
$$\pi(\nu) \;=\; \frac{\tau(\tau\nu)^k e^{-\tau\nu}}{\Gamma(k+1)}, \tag{55}$$
and $f_1(n;\theta,\nu)$ is integrated over $\pi(\nu)$ to obtain a distribution of n that depends on θ only:
$$f^\dagger(n;\theta) \;=\; \int f_1(n;\theta,\nu)\,\pi(\nu)\,\mathrm{d}\nu\,. \tag{56}$$
The likelihood ratio is now
$$\lambda^\dagger(n_{\text{obs}};\theta) \;=\; \frac{f^\dagger(n_{\text{obs}};\theta)}{f^\dagger(n_{\text{obs}}\mid\hat{\theta})}, \tag{57}$$
where $\hat{\theta}$ maximises $f^\dagger$ at the observed value n_{obs} of N. One then obtains the $(1-\alpha)$-quantile $q_{\text{Bayes},1-\alpha}(\theta)$ of the distribution of $-2\ln\lambda^\dagger(N;\theta)$ under $f^\dagger(n;\theta)$, and θ values for which $-2\ln\lambda^\dagger(n_{\text{obs}};\theta) \leq q_{\text{Bayes},1-\alpha}(\theta)$ form the desired interval.

5. *Simple percentile:* This is the bootstrap method of section 3.5.2: Intervals are computed from the $\frac{\alpha}{2}$- and $1-\frac{\alpha}{2}$-quantiles of the distribution of the estimator $\hat{\theta} = \max(N - K/\tau, 0)$. This distribution is derived from $f(n,k;\hat{\theta}_{\text{obs}},\hat{\nu}_{\text{obs}})$, where $\hat{\theta}_{\text{obs}}$ and $\hat{\nu}_{\text{obs}}$ are determined from n_{obs} and k_{obs}.

6. *Automatic percentile:* This is the second method described in section 3.5.2. Let $G(\hat{\theta};\theta,\nu)$ be the cumulative distribution of the estimate $\hat{\theta}$. The interval endpoints θ_1 and θ_2 are the solutions of $G(\hat{\theta}_{\text{obs}};\theta_1,\hat{\hat{\nu}}(\theta_1)) = 1-\frac{\alpha}{2}$ and $G(\hat{\theta}_{\text{obs}};\theta_2,\hat{\hat{\nu}}(\theta_2)) = \frac{\alpha}{2}$.

7. *Likelihood-ratio bootstrap:* Again a test inversion as in equation (52), but instead of minimising the likelihood-ratio tail probability with respect to ν, it is evaluated at $\nu = \hat{\nu}_{\text{obs}}$. Thus the confidence region is defined as the set of θ values for which
$$P\Big[-2\ln\lambda(N,K;\theta) \leq -2\ln\lambda(n_{\text{obs}},k_{\text{obs}};\theta)\ \Big|\ \theta,\hat{\nu}_{\text{obs}}\Big] \;\leq\; 1-\alpha\,. \tag{58}$$

If we write $q_{\text{bootstrap},1-\alpha}(\theta, \hat{\nu}_{\text{obs}})$ for the $(1-\alpha)$-quantile of the distribution of $-2\ln\lambda(N, K;\theta)$ under $f(n, k; \theta, \hat{\nu}_{\text{obs}})$, the above inequality is equivalent to

$$-2\ln\lambda(n_{\text{obs}}, k_{\text{obs}};\theta) \;\leq\; q_{\text{bootstrap},1-\alpha}(\theta, \hat{\nu}_{\text{obs}})\,. \tag{59}$$

8. *Likelihood-ratio profile bootstrap:* This is a variation on the previous method, where $\hat{\nu}_{\text{obs}}$ is replaced by $\hat{\hat{\nu}}_{\text{obs}}(\theta)$ in equations (58) and (59). It essentially corresponds to the profile method of equation (49).

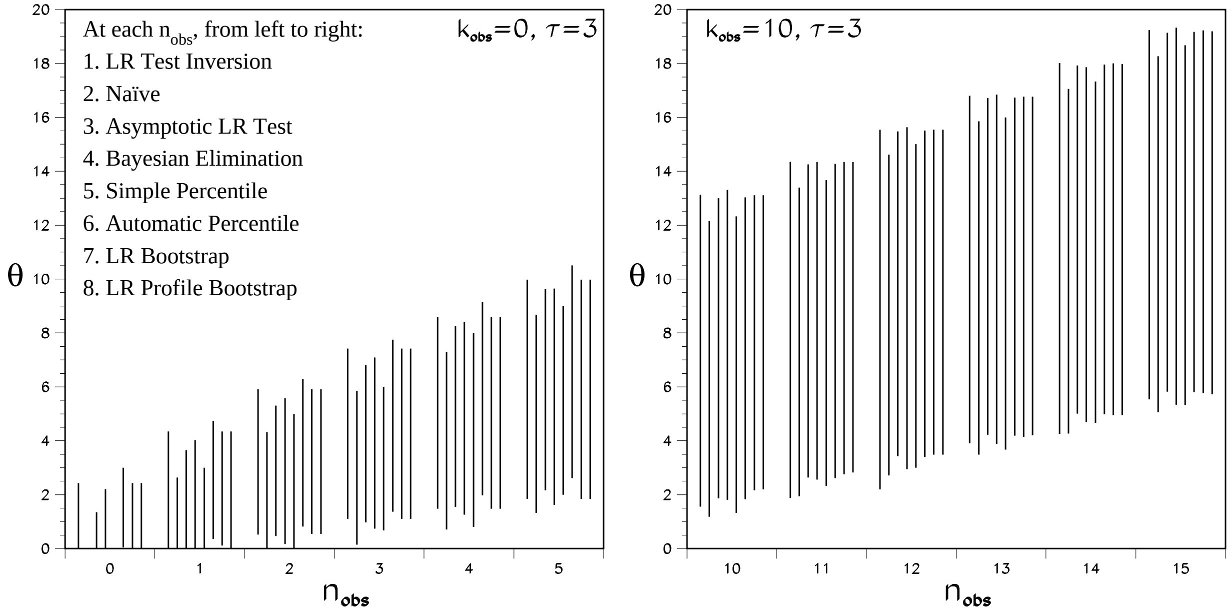

Figure 4: Interval constructions for the signal rate θ in a Poisson signal plus background problem; n_{obs} and k_{obs} are the numbers of events observed in the signal+background and background-only measurements, respectively, and τ is the known ratio of mean backgrounds between the two measurements. (a) The case $k_{\text{obs}} = 0$; (b) the case $k_{\text{obs}} = 10$. At each value of n_{obs}, eight vertical line segments represent 90% CL intervals for θ according to the methods listed in the legend on the left. A missing line segment indicates that the corresponding interval is empty or collapsed to the singleton $\{0\}$.

These interval constructions are compared in figure 4 for the cases $\tau = 3$ with $k_{\text{obs}} = 0$ (a) and $k_{\text{obs}} = 10$ (b), and for several values of n_{obs}. Differences between the methods are particularly pronounced at low n_{obs} and k_{obs}. In particular for $n_{\text{obs}} = k_{\text{obs}} = 0$ the naive and simple percentile intervals have zero length. In the latter case this is due to the use of MLEs to form the bootstrap distribution $f(n, k; \hat{\theta}_{\text{obs}}, \hat{\nu}_{\text{obs}})$, which is degenerate when the parameter estimates are both zero. In principle this could be remedied by choosing different estimators. At low n_{obs} values, the naive interval has the additional problem of extending into the unphysical region, resulting in unreasonably tight constraints on θ. This interval can certainly not be recommended. Among the other constructions, one notes that the asymptotic interval tends to be systematically shorter than the exact one, whereas the likelihood-ratio bootstrap and profile bootstrap intervals are often in good agreement with the latter. The performance of these methods is perhaps more easily judged by examining their frequentist coverage. One can plot the coverage as a function of θ at a fixed value of ν, or make a two-dimensional plot of coverage versus θ and ν, or plot as a function of θ the minimum and maximum coverages obtained when ν varies over a given range. Figure 5 shows the latter option, for $0 \leq \nu \leq 20$. As expected, the exact test-inversion method never undercovers, but it can substantially overcover. The naive, asymptotic, and simple percentile methods all have significant undercoverage at low values of θ. On the other hand, the likelihood-ratio bootstrap and profile bootstrap methods both perform remarkably well. For all methods the coverage tends to improve as θ increases. This conforms with the expectation that these methods should perform well in the large sample limit, which for Poisson processes is attained as the mean θ goes to infinity.

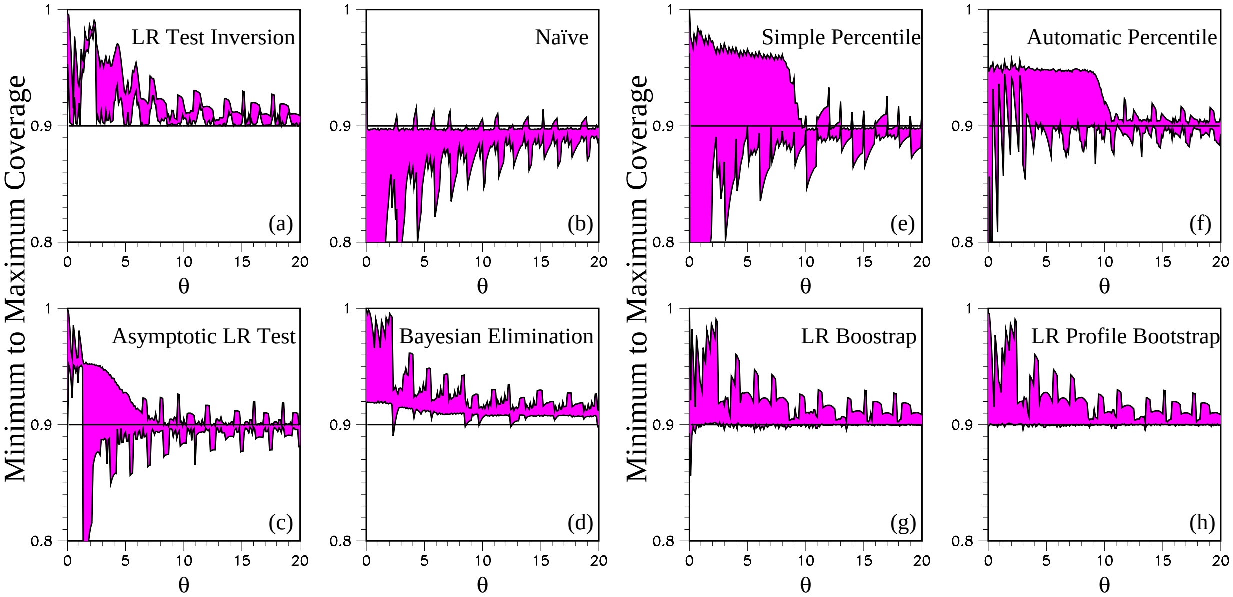

Figure 5: Bands showing the frequentist coverage versus θ of eight interval constructions for the Poisson signal plus background problem discussed in the text. The parameter τ is set to 3. The boundaries of the bands indicate the minimum and maximum coverage obtained by varying ν in the range $[0, 20]$. *The 90% nominal coverage is indicated by a horizontal line in each plot.*

4 Bayesian methods

As already emphasised in the introductory section 1, the output of a Bayesian analysis is *always* the complete posterior distribution for the parameter(s) of interest. However, it is often useful to summarise the posterior by quoting a region with a given probability content. Such a region can be an interval or a union of intervals. Several schemes, or "ordering rules", are available:

- *Highest-posterior-density regions (HPD):* Any parameter value inside such a region has a higher posterior probability density than any parameter value outside the region, guaranteeing that the region will have the smallest possible length (or volume).

Unfortunately this construction is not invariant under reparameterisations, and as example 4 will show, this lack of invariance can result in poor frequentist coverage for some parameter values (of course this will only be of concern to a frequentist or an objective Bayesian).

- *Equal-tailed intervals:* These are intervals that span equal posterior probabilities on each side of the posterior median. For example, a 68% equal-tailed interval extends from the 16th to the 84th posterior percentiles. These intervals are equivariant under one-to-one reparameterisations that are continuous from the left[6]. However, they typically only make sense when the posterior is unimodal[7], and their generalisation to multi-dimensional parameters is non-trivial. Furthermore, if a parameter is constrained to be non-negative, an equal-tailed interval will usually not include the value zero (an exception may occur if the posterior has a substantial probability mass at zero); this may be problematic if zero is a value of special physical significance.

- *Upper and lower limits:* For one-dimensional posterior distributions, these one-sided intervals can be defined using percentiles.

- *Likelihood regions:* These are standard likelihood contours, i.e. regions of parameter values for which the likelihood is larger than for any parameter value outside the region. The size of the region is determined by the desired posterior credibility. Such regions are metric independent and robust with respect to the choice of prior [Was89]. In one-dimensional problems with physical boundaries and unimodal likelihoods this construction yields intervals that have a smooth transition from one-sided to two-sided.

- *Lowest posterior loss regions:* A more foundational approach to Bayesian interval construction starts with a loss structure [BS94]. Suppose that we can in some way quantify the loss $\ell\{\theta_0, \theta\}$ incurred by using the parameter value θ_0 when the true value is θ. After having observed data $\mathbf{x}$, our posterior expected loss is

$$l\{\theta_0; \mathbf{x}\} \;=\; \int \ell\{\theta_0, \theta\}\, p(\theta; \mathbf{x})\, \mathrm{d}\theta\,, \tag{60}$$

where $p(\theta; \mathbf{x})$ is the posterior density. A natural point estimate of θ is then the value that minimises this posterior expected loss, and a natural credible region is the set of θ values for which the posterior expected loss is smaller than for any value outside the set, subject to a credibility constraint. A possible choice of loss function is the quadratic loss — $\ell\{\theta_0, \theta\} = (\theta_0 - \theta)^2$ — which yields the posterior mean as point estimate. Another choice is zero-one loss — $\ell\{\theta_0, \theta\} = 0$ if $|\theta_0 - \theta| \le \epsilon$, and $\ell\{\theta_0, \theta\} = 1$ otherwise, where ϵ is a constant. As ϵ goes to zero this loss function leads to the posterior mode as point estimate and to credibility regions that have highest posterior density. Many more loss functions can be devised, but in the absence of any subjective preference, information-theoretic arguments lead to the concept of *intrinsic discrepancy loss* [Ber05], which is

[6] A reparameterisation $\theta \to \eta(\theta)$ is continuous from the left if $\lim_{\theta \uparrow \theta_0} \eta(\theta) = \eta(\theta_0)$.

[7] A probability density function with a single maximum is called unimodal.

defined as the symmetrised *Kullback–Leibler divergence* between the model indexed by θ_0 and that indexed by θ:

$$\delta\{\theta_0, \theta\} \;=\; \min\Big\{ \; \kappa\{p(\mathbf{x};\theta_0); p(\mathbf{x};\theta)\}, \; \kappa\{p(\mathbf{x};\theta); p(\mathbf{x};\theta_0)\} \; \Big\}\,, \tag{61}$$

with

$$\kappa\{p(\mathbf{x};\theta_0); p(\mathbf{x};\theta)\} \;=\; \int p(\mathbf{x};\theta)\, \ln \frac{p(\mathbf{x};\theta)}{p(\mathbf{x};\theta_0)}\, \mathrm{d}\mathbf{x} \tag{62}$$

(for discrete sample spaces the integral is replaced by a sum). From this definition it follows that the intrinsic discrepancy loss between two models can be interpreted as the minimum expected log-likelihood ratio in favour of the model that generated the data. Credible regions derived from this loss function are labeled “intrinsic”. They enjoy many useful properties, including equivariance under parameter transformation, and they are available in multi-dimensional settings. A one-dimensional example is given in example 4.

Users of Bayesian procedures are generally advised to assess the sensitivity of their result to the choice of prior. Furthermore, if the prior is of the so-called non-informative variety, the behaviour of the result under repeated sampling (i.e. the frequentist coverage) should also be investigated.

In the context of interval construction, it is worth mentioning that non-informative priors can be designed in such a way that the resulting posterior intervals have a frequentist coverage that matches their Bayesian credibility to some order in $1/\sqrt{n}$, n being the sample size. When there are no nuisance parameters and the parameter of interest is one-dimensional, the matching prior to $\mathcal{O}(1/n)$ for one-sided intervals is *Jeffreys' prior*:

$$\pi_J(\theta) \;\propto\; \sqrt{E\Big[-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln L(x;\theta)\Big]}\,. \tag{63}$$

Frequentist coverage is harder to achieve in higher dimensions, but the *Bayesian reference analysis approach* has obtained good results [BS94]. This is an objective Bayesian method based on information-theoretic considerations. In spite of being non-subjective, it provides results with a credibility interpretation: Such results would be obtained by a person whose prior beliefs have minimal effect, relative to the data, on posterior inferences. An application to cross section measurements in high energy physics is described in [DJP10].

The next subsection illustrates Bayesian interval constructions with an example that appears simple and yet can lead to serious difficulties if not handled properly. Sections 4.1 and 4.2 summarise the calculation of Bayesian intervals for binomial efficiencies and Poisson means, thereby complementing the two frequentist interval calculations given in sections 3.3.3 and 3.3.4. The Bayesian calculations are based on Jeffreys' prior, and the resulting intervals are therefore known as Jeffreys intervals.

Example 4 (Measuring track momenta) Consider the measurement of particle transverse momenta in a tracking chamber immersed in a solenoidal magnetic field. A simple model is that for a given particle the charge-signed transverse momentum is the inverse of the radius of curvature ρ of its track, and that the measured curvature radius has a Gaussian distribution with standard deviation σ proportional to the chamber resolution and inversely proportional to the magnetic field strength. Thus, if x is the measured transverse momentum and θ its true value, the likelihood function has the form:

$$L(x;\theta) \;\propto\; e^{-\frac{1}{2}\left(\frac{1/x-1/\theta}{\sigma}\right)^2}\,. \tag{64}$$

A straightforward calculation shows that Jeffreys' prior is proportional to $1/\theta^2$. The properly normalised posterior is therefore

$$p(\theta;x) \;=\; \frac{e^{-\frac{1}{2}\left(\frac{1/x-1/\theta}{\sigma}\right)^2}}{\sqrt{2\pi}\,\sigma\,\theta^2}\,. \tag{65}$$

This posterior is shown in figure 6(a) for the case $\sigma = 1$ and $x = 1$. There are two local maxima, at $\theta_\pm = (-1 \pm \sqrt{1+8x^2\sigma^2})/(4x\sigma^2)$, corresponding to two possible charge assignments to the observed track. As $|x| \to \infty$, the posterior density reaches equal heights at these maxima, reflecting the ambiguity in charge determination at large momenta. However, the posterior mode is a very biased estimate of θ since $|\theta_\pm|$ never exceeds $1/(\sqrt{2}\sigma)$. Highest-posterior-density (HPD) credible regions are shown in figure 6(b): They consist of a single interval at low $|x|$, and of the union of two intervals at large $|x|$. At large $|x|$ the credibility "belt" (i.e. the set of credible regions viewed in (x,θ) space) consists of two horizontal bands that are bounded away from large θ values. As a result, the frequentist coverage of HPD regions is zero at large $|\theta|$! This may surprise in view of the facts that the posterior (65) for the transverse momentum θ can be derived from a Gaussian posterior for the curvature radius ρ via the transformation $\rho \to \theta = 1/\rho$, and that HPD intervals for a Gaussian posterior have exact frequentist coverage. The problem, of course, is that HPD intervals are not equivariant under reparameterisation. This suggests a simple solution, which is to construct an HPD interval for ρ and to invert the endpoints to obtain a credible region for θ, taking care of the case where the ρ interval contains zero. Applying this idea to the 68% credible interval $[1/x - \sigma,\ 1/x + \sigma]$ for ρ leads to the following credible region for θ:

$$\begin{array}{ll} \left[\dfrac{1}{1/x+\sigma},\, \dfrac{1}{1/x-\sigma}\right] & \text{if} \quad |x| < 1/\sigma, \quad \text{and} \\[2ex] \left]-\infty,\, \dfrac{1}{1/x-\sigma}\right] \cup \left[\dfrac{1}{1/x+\sigma},\, +\infty\right[& \text{if} \quad |x| > 1/\sigma. \end{array} \tag{66}$$

This is not an HPD region in the θ parameterisation, but its coverage is 68%, exactly matching its credibility.

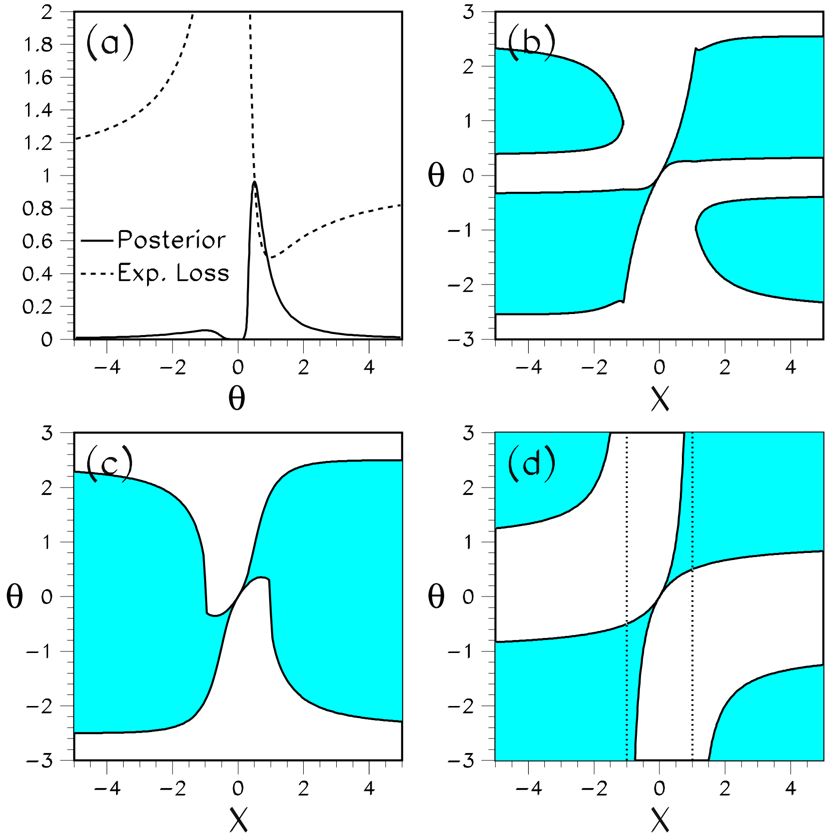

Figure 6: Credible region construction for a transverse momentum with true value θ and observed value x in a tracking chamber immersed in a magnetic field. (a) The posterior density for $x = 1$ (solid line), superimposed on the posterior expected intrinsic loss function (dashes). (b)-(d): The 68% credible intervals (shaded regions) in θ as a function of x, for highest posterior density, equal-tails, and lowest intrinsic loss constructions, respectively. The vertical dotted lines in (d) are asymptotes of the boundaries of the credibility belt. For x values between these lines the credible region is a single interval; outside these lines it is the union of two open intervals (equation (66)).

Figure 6(c) shows the 68% credibility belt for Bayesian equal-tailed intervals. Again, the outer contour of the belt becomes horizontal at large $|x|$, resulting in zero coverage for large θ values. The transformation $\rho \to \theta = 1/\rho$ is one-to-one but not continuous from the left, explaining why the nice properties of equal-tailed intervals for ρ do not transfer to θ.

Finally we examine intrinsic loss credible regions. The intrinsic discrepancy loss, equation (61), becomes

$$\delta\{\theta_0, \theta\} \;=\; \frac{1}{2}\left(\frac{1/\theta_0 - 1/\theta}{\sigma}\right)^2 . \tag{67}$$

The posterior expected intrinsic discrepancy loss is then

$$d\{\theta_0; x\} \;=\; \int \delta\{\theta_0, \theta\}\, p(\theta; x)\, \mathrm{d}\theta \;=\; \frac{1}{2}\left[1 + \left(\frac{1/\theta_0 - 1/x}{\sigma}\right)^2\right], \tag{68}$$

and is plotted as a dashed line in figure 6(a). The minimum-loss estimate of θ is x, and minimum-loss credible regions are given by equation (66) and shown in figure 6(d). In this case the intrinsic discrepancy loss formalism has automatically produced point and interval estimates that correspond to HPD in the curvature radius parameterisation. ■

4.1 Binomial efficiencies

The binomial likelihood for an efficiency ϵ, after having observed x successes out of n trials, is

$$L(n, x; \epsilon) \;=\; \binom{n}{x} \epsilon^{x} (1-\epsilon)^{n-x}\,, \tag{69}$$

from which Jeffreys' prior is found to be

$$\pi_J(\epsilon) \;\propto\; \sqrt{\frac{n}{\epsilon(1-\epsilon)}}\,. \tag{70}$$

The properly normalised posterior is a Beta$(x+\frac{1}{2}, n-x+\frac{1}{2})$ distribution:

$$p(\epsilon; x) \;=\; \frac{\epsilon^{x-\frac{1}{2}}(1-\epsilon)^{n-x-\frac{1}{2}}}{B(x+\frac{1}{2}, n-x+\frac{1}{2})}\,. \tag{71}$$

The endpoints of an equal-tailed, $(1-\alpha)$ CL Bayesian interval $[\epsilon_{low}, \epsilon_{up}]$ for ϵ are the $\frac{\alpha}{2}$- and $1-\frac{\alpha}{2}$-quantiles of this posterior:

$$\epsilon_{low} \;=\; B_{x+\frac{1}{2}, n-x+\frac{1}{2}, \frac{\alpha}{2}} \quad \text{and} \quad \epsilon_{up} \;=\; B_{x+\frac{1}{2}, n-x+\frac{1}{2}, 1-\frac{\alpha}{2}}\,. \tag{72}$$

These can be compared with the frequentist formulæ (29) and (30). In contrast with Clopper–Pearson intervals, Jeffreys intervals tend to be shorter but do not guarantee exact coverage. The coverage of both constructions oscillates as a function of the true value of ϵ. For Clopper–Pearson intervals these oscillations all remain above the nominal confidence level $1-\alpha$, whereas for Jeffreys intervals they straddle $1-\alpha$ [Cai05].

4.2 Poisson means

For a Poisson-distributed number of events n the likelihood is

$$L(n; \theta) \;=\; \frac{(\theta+\nu)^{n}\, e^{-\theta-\nu}}{n!}\,, \tag{73}$$

where, as before, θ is the signal strength of interest and ν is the level of a known background contamination. Jeffreys' rule (63) gives:

$$\pi_J(\theta) \;\propto\; \frac{1}{\sqrt{\theta+\nu}}\,, \tag{74}$$

and the corresponding posterior is a shifted Gamma distribution:

$$p(\theta; n) \;=\; \frac{(\theta+\nu)^{n-\frac{1}{2}}\, e^{-\theta-\nu}}{\Gamma(n+\frac{1}{2})\,[1-P(n+\frac{1}{2}, \nu)]}\,, \quad \text{with} \quad P(a, \nu) \;\equiv\; \int_0^{\nu} \frac{t^{a-1} e^{-t}}{\Gamma(a)}\, \mathrm{d}t\,. \tag{75}$$

An $(1-\alpha)$ CL upper limit $\theta_{1-\alpha}$ is given by the $(1-\alpha)$-quantile of this posterior:

$$1-\alpha \;=\; \int_0^{\theta_{1-\alpha}} p(\theta; n)\,\mathrm{d}\theta \;=\; \frac{P(n+\frac{1}{2}, \nu+\theta_{1-\alpha}) - P(n+\frac{1}{2}, \nu)}{1-P(n+\frac{1}{2}, \nu)}$$
$$= \frac{P[\chi^2_{2n+1} \leq 2(\nu+\theta_{1-\alpha})] - P(n+\frac{1}{2}, \nu)}{1-P(n+\frac{1}{2}, \nu)}, \qquad (76)$$

where, similarly to what was done in section 3.3.4, we converted an incomplete Gamma function into the tail probability of a χ^2 distribution. Solving for the latter yields

$$P[\chi^2_{2n+1} \leq 2(\nu+\theta_{1-\alpha})] \;=\; 1-\alpha', \quad \text{with} \quad 1-\alpha' \equiv 1-\alpha+\alpha P(n+\frac{1}{2}, \nu)\,. \qquad (77)$$

Hence we find

$$\theta_{1-\alpha} \;=\; \frac{1}{2}\chi^2_{2n+1,1-\alpha'} - \nu\,, \qquad (78)$$

which can be compared to equation (34). In contrast with the frequentist result, the Jeffreys upper limit never becomes negative, thanks to the dependence of α' on ν.

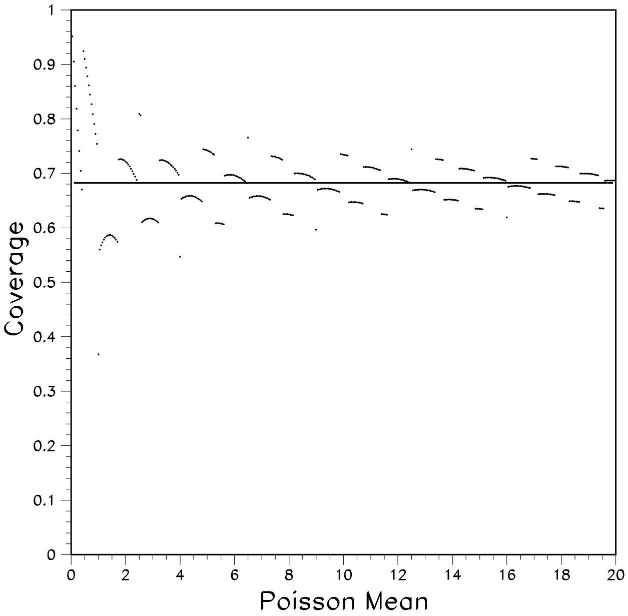

Figure 7: Frequentist coverage of 68% CL Bayesian central intervals for the mean of a Poisson distribution, using Jeffreys' prior. The coverage is evaluated in increments of 0.1 in the value of the true Poisson mean, and the Bayesian credibility of the construction is indicated by the solid horizontal line.

Figure 7 shows how the coverage of the Jeffreys limit oscillates as a function of the true value of θ, with downward oscillations dipping below the Bayesian credibility of the interval. For this reason a flat prior is sometimes preferred, as the resulting coverage oscillations remain above the credibility. For a flat prior the upper limit is given by:

$$\theta^{\text{flat}}_{1-\alpha} \;=\; \frac{1}{2}\chi^2_{2n+2,1-\alpha''} - \nu\,, \quad \text{where} \quad 1-\alpha'' \;\equiv\; 1-\alpha+\alpha P(n+1, \nu)\,. \qquad (79)$$

The good coverage properties of the flat prior are only true for upper limits, however; for lower limits and two-sided intervals Jeffreys' rule performs better.

Table 3: Bayesian interval constructions for the mean of a Poisson distribution when n events are observed. All results were obtained with Jeffreys' prior. The ordering rules shown are 95% CL upper limit (column 2), 68% CL equal-tailed interval (column 3), and 95% and 68% CL lowest posterior-expected intrinsic discrepancy loss (columns 4 and 5).

	Bayesian intervals with Jeffreys' prior			
	Upper limit	Equal-tailed	Lowest post. exp. intr. loss	
n	95% CL	68% CL	95% CL	68% CL
0	1.92	[0.02, 0.99]	[0.00, 1.92]	[0.02, 0.91]
1	3.90	[0.42, 2.59]	[0.01, 3.93]	[0.41, 2.57]
2	5.53	[1.02, 3.97]	[0.28, 5.82]	[1.03, 3.97]
3	7.03	[1.72, 5.28]	[0.69, 7.48]	[1.71, 5.27]
4	8.46	[2.45, 6.54]	[1.17, 9.03]	[2.45, 6.54]
5	9.83	[3.22, 7.77]	[1.72, 10.50]	[3.22, 7.77]
6	11.18	[4.01, 8.98]	[2.32, 11.93]	[4.01, 8.98]
7	12.49	[4.82, 10.17]	[2.94, 13.33]	[4.82, 10.17]
8	13.79	[5.64, 11.35]	[3.59, 14.69]	[5.64, 11.35]
9	15.07	[6.47, 12.52]	[4.25, 16.03]	[6.47, 12.52]
10	16.33	[7.31, 13.69]	[4.94, 17.35]	[7.31, 13.68]

For the Poisson model the intrinsic discrepancy loss is given by

$$\delta\{\theta_0, \theta\} \;=\; |\theta_0 - \theta| - \big[\nu + \min(\theta_0, \theta)\big] \left| \ln \frac{\nu + \theta_0}{\nu + \theta} \right| , \tag{80}$$

and its posterior expectation can be computed numerically. Bayesian intervals derived from this loss function are shown in table 3, together with regular upper limits and equal-tailed intervals. Note that the 95% CL intrinsic interval coincides with the upper limit when zero events are observed. This is due to the fact that in order to obtain a higher credible interval one has to tolerate a higher loss, which eventually becomes larger than the loss at $\theta = 0$. At 99% CL for example, the intrinsic interval coincides with the upper limit for $N = 0$, 1, and 2. This unification of two-sided intervals and upper limits is reminiscent of Feldman–Cousins intervals in the frequentist case, where it could be used to test a parameter value on the boundary. However, it does not have the same significance here, because the duality between confidence intervals and hypothesis tests only exists in the frequentist paradigm.

5 Graphical comparison of interval constructions

The effect of a physical boundary on frequentist and Bayesian interval constructions is illustrated in figures 8 and 9 for the measurement of the mean θ of a Gaussian with unit standard deviation. The true mean θ is constrained to be positive. All intervals are based on a single observation x, which can be positive or negative due to resolution effects. This is a simplified model corresponding for example to the measurement of the square of a neutrino mass discussed in [FC98]. As pointed out in section 2, intervals have many properties that are worth studying: here we only examine the Bayesian credibility of frequentist constructions and the frequentist coverage of Bayesian constructions.

Figure 8 shows only frequentist constructions. Due to the positivity constraint on θ, the 68% CL equal-tailed (or central) interval (figure 8(a)) is empty whenever the observation x is below -1. For x between -1 and $+1$ the interval is an upper limit, and for x higher than $+1$ it is two-sided. Since this is an exact frequentist construction, its coverage is 68% for all physical values of θ. From a frequentist point of view empty intervals are not meaningless: they simply indicate that no physical value of θ can account for the observation *at the stated confidence level*. However, empty intervals have a drastic effect on Bayesian credibility. We can investigate this with the help of Jeffreys' prior, which for this problem is zero for $\theta < 0$ and a positive constant for $\theta \geq 0$. For each value of x the integral of the posterior density over the corresponding θ interval yields the latter's credibility. The result is shown in figure 8(b): the credibility vanishes for $x < -1$, then rises sharply up to a maximum at $x = 1$, and finally for $x > 2$ it settles down to a value very close to the frequentist coverage.

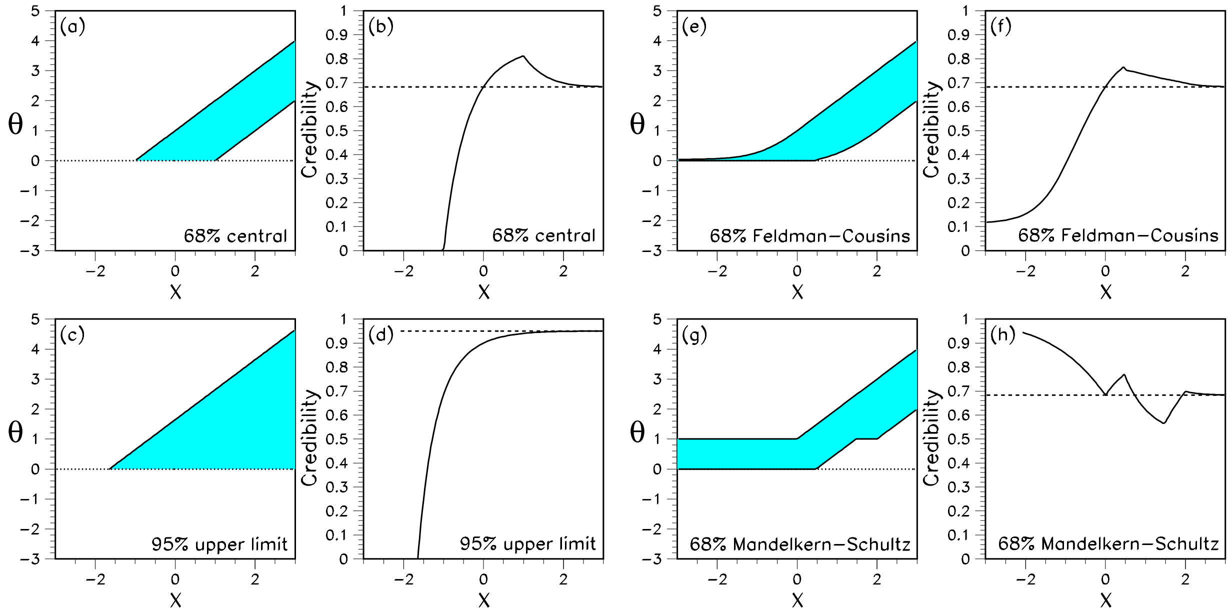

Figure 8: Frequentist interval constructions. Panels (a), (c), (e) and (g) show graphs of θ versus x, with dotted lines indicating the lower boundary of the physical region. Panels (b), (d), (f) and (h) show the corresponding Bayesian credibility levels based on Jeffreys' prior, with dashed lines indicating the frequentist coverage.

The remaining pairs of panels in figure 8 are similarly organised, showing a frequentist construction on the left and the corresponding Bayesian credibility on the right. It can be seen that upper limits have the same credibility problem as central intervals. The remaining two frequentist constructions mitigate the credibility problem by avoiding empty intervals. Feldman-Cousins intervals, shown in figure 8(e) and (f), use x as estimator of θ and are based on a likelihood ratio ordering rule [FC98]. They still have low credibility for negative

X values. Mandelkern–Schultz intervals, presented in figure 8(g) and (h), use $\max\{0, X\}$ as estimator of θ and are based on an equal-tails ordering rule [MS00]. These intervals are the same for any negative X as for zero X, resulting in excess credibility at negative X.

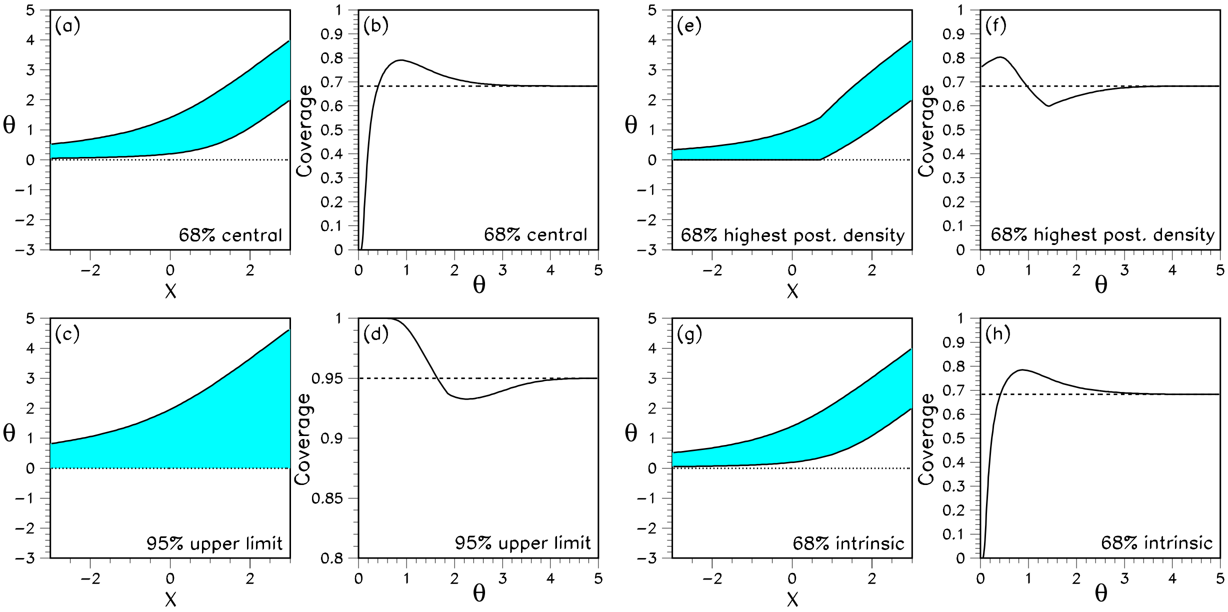

Figure 9: Bayesian interval constructions. Panels (a), (c), (e) and (g) show graphs of θ versus x, with dotted lines indicating the lower boundary of the physical region. Panels (b), (d), (f) and (h) show the corresponding frequentist coverage levels, with dashed lines indicating the Bayesian credibility.

Figure 9 shows four Bayesian constructions in paired panels. For each pair, the left panel shows the credibility belt and the right one the corresponding frequentist coverage. All constructions use Jeffreys' prior for θ and differ only by the ordering rule used. Panel pairs (a) and (b), (c) and (d), and (e) and (f) use equal-tailed, upper limit and highest posterior density ordering, respectively. On panel (g) and (h) the ordering is according to the intrinsic discrepancy loss, which for this problems equals $\delta\{\theta_0, \theta\} = (\theta - \theta_0)^2/2$ and coincides with quadratic loss. All four constructions have reasonable frequentist coverage, except near $\theta = 0$, where the curves for equal-tailed and intrinsic intervals dip to zero.

A noteworthy feature of both figures 8 and 9 is that frequentist coverage and Bayesian credibility always agree with each other when one is far enough from the physical boundary.

6 The role of intervals in search procedures

Suppose that we are using a collider experiment to search for a new particle with unknown production rate θ. From a statistical point of view this problem can be formulated as a hypothesis test of

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta > 0\,, \tag{81}$$

since H_0 corresponds to non-existence of the particle and H_1 to its existence. Here we are following the formalism in the test inversion section 3.2. In a frequentist approach we calculate a p-value p_0 to test H_0, and claim discovery if $p_0 \leq \epsilon$, where ϵ is a pre-specified type I error rate (typically $2.87 \cdot 10^{-7}$, corresponding to 5σ in the tail of a Gaussian distribution). It is then customary to accompany the discovery claim with point and interval estimates of θ, with the interval being two-sided and providing 68% confidence. If on the other hand $p_0 > \epsilon$, we fail to reject H_0. However, this decision does not imply that all values of θ under H_1 are now rejected. In particular, there are values of θ that the experiment is simply not sensitive to, and values that the data won't allow us to exclude. Hence we need to investigate this more closely: For which values θ_0 can the hypothesis $H_1' : \theta = \theta_0$ be excluded? More precisely, we need to test:

$$H_1'[\theta_0] : \theta = \theta_0 \qquad \text{versus} \qquad H_0'[\theta_0] : \theta < \theta_0\,, \tag{82}$$

where for later convenience we specified the tested value of θ as an argument to the hypotheses. Comparing with equation (3) in section 3.2 shows that the set of θ_0 values for which $H_1'[\theta_0]$ cannot be excluded, that is, the set of particle production rates that our data cannot exclude, is of the form $[0, \theta_{up}]$ for some upper limit θ_{up}. The value of θ_{up} depends on the size of test (82). A common choice in high energy physics is 5%, so that the upper limit θ_{up} will have 95% CL.

From a frequentist point of view there are two problems with the scenario of discovery versus non-discovery just outlined: one concerns coverage, and the other measurement sensitivity. We discuss these issues in the next two subsections.

6.1 Coverage

When we claim discovery, we typically quote a 68% CL, two-sided interval on the new particle production rate θ. When we fail to claim discovery, we quote a 95% CL upper limit. What is the reference ensemble (see section 3.1.2) of these confidence level statements? It might seem sensible to refer the 68% CL intervals to the ensemble $\mathcal{E}_1$ of all searches that claim discovery, and the 95% CL limits to the ensemble $\mathcal{E}_2$ of all searches that don't. Unfortunately this doesn't work, because in $\mathcal{E}_2$ the upper limits undercover at large values of θ and in $\mathcal{E}_1$ the two-sided intervals undercover at low values of θ. One might perhaps think that this problem would disappear if we had just one common reference ensemble instead of two; and that this could be achieved if we decided to quote the same confidence level for both upper limits and two-sided intervals, say 90%. However, as shown in Ref. [FC98] this doesn't work either, because there is then a set of intermediate values of θ where the coverage is only 85%.

The real source of the problem lies in the fact that the decision regarding the type of interval to quote is based on the observation itself; reference [FC98] calls this policy "flip-flopping based on the data". There would be no undercoverage if somehow the decision could be made *before* looking at the data. Since this is not possible in the context of a search for new physics, [FC98] advocates the use of the likelihood ratio ordering rule described in section 3.1.3. With this rule, intervals of a given confidence level are two-sided (see figure 8(e)) when the observation is above a certain threshold, and one-sided when it is below. However, this does not yet completely solve the problem, because as mentioned previously, the search procedure used in practice involves three confidence levels: 5σ to decide on a discovery claim, 95% for upper limits, and 68% for two-sided intervals. Thus, for example, there would still be undercoverage if one were to report 68% CL likelihood ratio intervals *only* when claiming discovery, etc. The solution here is to *always* report both the 68% and 95% CL intervals.

6.2 Sensitivity

The sensitivity issue arises when there is no convincing evidence favouring the existence of a new particle, and we cannot reject the background-only hypothesis $H_0 : \theta = 0$. We then proceed to set an upper limit θ_{up} on the new particle production rate θ, typically at the 95% confidence level. The *desired* interpretation of θ_{up} is that θ values above it are both within the sensitivity range of the experimental apparatus and excluded by the observations; θ values below θ_{up} are either outside the sensitivity range or not excluded by the observations. Unfortunately, when θ_{up} is determined by a frequentist method, there is a finite probability that it will exclude θ values to which the experiment is not sensitive. As a simple illustration, consider the case where the observation is a Poisson distributed number of events x with mean $\theta+\nu$, where ν is a known background contamination. We first encountered this example in section 3.3.4, where the frequentist upper limit is given by equation (34). Remarkably, that upper limit decreases as the background contamination increases, and could even be negative. However, a negative upper limit means that all physical values of θ are excluded by the experiment, which is clearly implausible. In the case where no events are observed, the formula gives $\theta_{up} = -\ln\alpha - \nu$, which is negative whenever $\nu \geq -\ln\alpha$. In the absence of signal, the probability of no events is $e^{-\nu}$ and therefore the probability of a negative upper limit could be as high as $e^{\ln\alpha} = \alpha$. For a 95% CL limit this is 5%, which is considered quite substantial by many physicists.

Several attempts have been made to handle this problem [Hig87], none entirely satisfactory. All have to deal with the ambiguity of deciding which θ values are outside the sensitivity reach of the experiment, and whether this set of values depends on the confidence level of the upper limit or even on the strength of evidence provided by the data [Cou11]. To compare approaches it is convenient to introduce some notation: let p_0 be the p-value used to test H_0 in test (81) and $p_1(\theta_0)$ the p-value used to test $H_1'[\theta_0]$ in test (82). The standard frequentist $(1-\alpha)$ CL upper limit construction rejects θ values for which $p_1(\theta) < \alpha$. A simple modification of this construction that addresses the sensitivity problem is to reject θ values for which *both* $p_1(\theta) < \alpha$ *and* $\theta \in \mathcal{S}$, where $\mathcal{S}$ is the subset of parameter space that contains all the θ values to which the experiment is deemed to be sensitive. There is no unique way of defining the sensitivity set $\mathcal{S}$. One approach [Kas+10] defines it as containing all θ values that have probability at least β of being detected at the γ significance level if H_1 is true:

$$\mathcal{S} \;=\; \Big\{\theta :\; P\Big[p_0 \leq \gamma \mid H_1[\theta]\Big] \geq \beta\Big\}. \tag{83}$$

Thus, in addition to the confidence level $1-\alpha$, this method requires the choice of two probabilities, β and γ.

An alternative definition of $\mathcal{S}$ is as the set of θ values for which the inequality $p_1(\theta) < \alpha$ is *expected* to occur with probability at least β if H_0 is true:

$$\mathcal{S} \;=\; \Big\{\theta :\; P\Big[p_1(\theta) \leq \alpha \mid H_0\Big] \geq \beta\Big\}. \tag{84}$$

The advantage here is that one needs to choose only one additional probability, namely β. The $\theta \in \mathcal{S}$ requirement is sometimes called a power constraint, due to the fact that the probabilities calculated in definitions (83) and (84) are power functions, i.e. they are probabilities for rejecting one hypothesis when the other is true.

Although power constraint methods provide some protection against excluding parameter values to which the experiment is not sensitive, they fail to address another problem of the frequentist limit (34), which is that if two experiments have different background contaminations but observe the same number of events, the experiment with the larger contamination will be able to exclude more signal [Rea02]. An approach which is arguably more successful at dealing with all manner of sensitivity problems is the so-called $\mathrm{CL_s}$ prescription [Jun99; Rea02]. A $(1-\alpha)$ CL $\mathrm{CL_s}$ upper limit construction rejects θ values for which

$$\mathrm{CL_s} \;\equiv\; \frac{p_1(\theta)}{1-p_0} \;<\; \alpha\,. \tag{85}$$

Note that this is a *stronger* requirement than the standard frequentist rejection criterion $p_1(\theta) < \alpha$. As a result, $\mathrm{CL_s}$ upper limits overcover from a frequentist point of view. On the other hand, in simple problems such as setting an upper limit on a Gaussian or Poisson mean, the $\mathrm{CL_s}$ result agrees with the Bayesian one for a constant prior.

It should be kept in mind that the $\mathrm{CL_s}$ prescription is nothing more than the rejection criterion (85). Just as with standard p-values, there is complete freedom in the choice of test statistic and method for handling nuisance parameters. Experiments at LEP, the Tevatron, and the LHC have all adopted different conventions and strategies in this regard, and one should be careful when attempting comparisons. In contrast with p-values however, the $\mathrm{CL_s}$ prescription is only used to compute upper limits.

Finally, we emphasise that Bayesian methods do not suffer from sensitivity problems due to the fact that they fully condition on the observations.

7 Final remarks and recommendations

One way to view the great assortment of interval constructions discussed in this chapter is as a set of answers to slightly different questions. Frequentist and Bayesian intervals with various ordering rules can all produce different inferences from the same data set. Whether these differences matter depends on the biases and expectations of the analyst, but also on objective factors such as the evidence available prior to the measurement, the sample size, systematic effects, and instrumental sensitivity. Thus if the consumer of the measurement result is provided with more than one interval estimate, for example a frequentist, a Bayesian, and an asymptotic construction, then he or she will better be able to judge the robustness and significance of the final result.

A recurring problem in high energy physics is the handling of nuisance parameters. When the sample size is large enough, asymptotic approximations based on the likelihood function can be trusted. Care is required in small samples however. An approximate frequentist approach is to first eliminate the nuisance parameter(s) by profiling or Bayesian integration, and then apply a test-inversion method on the parameter of interest. Although past experience with this approach has shown it to be reliable, one is always well advised to perform a few spot checks of the coverage. When using a Bayesian interval construction on small samples, one should of course evaluate the sensitivity of the final result to reasonable changes in the prior.

Another issue arises when the parameter space has physical boundaries, especially when the experiment has only weak sensitivity in the vicinity of such a boundary. The main concern is to avoid reporting intervals that exclude parameter values to which the apparatus is not sensitive. Bayesian methods appear to behave properly in this situation, but no single frequentist method is entirely satisfactory. This is another argument for reporting more than one type of interval.

8 Exercises

Exercise 1: Eliminating nuisance parameters by conditioning

In the frequentist paradigm, handling nuisance parameters can be a thorny problem. A method that sometimes works is based on the idea of *conditioning*. To illustrate this approach, suppose we measure an event count N that is Poisson-distributed with mean $\mu\nu$, where μ is the parameter of interest and ν a nuisance parameter. Assume that ν is constrained by the auxiliary measurement of a Poisson variate K with mean $\tau\nu$, where τ is a known constant:

$$N \sim \text{Poisson}(\mu\nu)\,, \tag{86}$$

$$K \sim \text{Poisson}(\tau\nu)\,. \tag{87}$$

In high energy physics one could think of μ as the production cross section for some process of interest and ν as a product of efficiencies, acceptances, and integrated luminosity. One can argue that the *sum* $M \equiv N + K$ provides no information about the *ratio* μ/τ of the above two Poisson means, or about μ itself. It is therefore interesting to seek inferences that condition on M. First, show that the conditional distribution of N given M is given by:

$$P[N = n \mid M = m] \;=\; \binom{m}{n}\left(\frac{\mu}{\tau+\mu}\right)^{n}\left(1-\frac{\mu}{\tau+\mu}\right)^{m-n}. \tag{88}$$

This is a binomial distribution that does not involve the nuisance parameter ν; it can therefore be used for inference about μ. Using the results from section 3.3.3 on binomial efficiencies for example, one can compute a confidence interval for the binomial parameter $\mu/(\tau + \mu)$. Assuming you have such an interval, transform it into an interval for μ and examine what happens when $n = 0$. What about when $k = 0$? Or when $n = k = 0$?

Next, suppose that the mean of N is the sum of μ and ν instead of their product, so we have:

$$N \sim \text{Poisson}(\mu + \nu)\,, \tag{89}$$

$$K \sim \text{Poisson}(\tau\nu)\,. \tag{90}$$

Can we still apply the conditioning method to eliminate the nuisance parameter ν here?

Exercise 2: Bayesian intervals for an exponential lifetime

Consider the exponential decay example of section 3.3.2, where the probability density of the data t is $f(t;\tau) = e^{-t/\tau}/\tau$. Derive Jeffreys' prior for this problem and compute the corresponding posterior. Construct equal-tailed intervals from this posterior and compare them to the corresponding frequentist intervals in table 1. What can you conclude about the relationship between Bayesian credibility and frequentist coverage for this problem?

Show that the intrinsic discrepancy loss for this problem is given by

$$\delta\{\tau_0, \tau\} \;=\; \min\left\{\frac{\tau_0}{\tau}, \frac{\tau}{\tau_0}\right\} - 1 + \left|\ln\left(\frac{\tau_0}{\tau}\right)\right|, \tag{91}$$

and the posterior expectation of this loss by

$$d\{\tau_0; t\} \;=\; -\left(1 + \frac{\tau_0}{t}\right) e^{-t/\tau_0} + \frac{\tau_0}{t} - 1 + \gamma + \ln\left(\frac{t}{\tau_0}\right) + \left(2 + \frac{t}{\tau_0}\right) \mathrm{E}_1\left(\frac{t}{\tau_0}\right), \tag{92}$$

where $\gamma = 0.57721\,56649\,01532\,86060\ldots$ is the Euler–Mascheroni constant and $\mathrm{E}_1(x) = \int_x^\infty e^{-t}/t\,dt$ is the exponential integral. Plot $d\{\tau \mid t\}$ as a function of τ, for $t = 1$, and compare the minimum-loss estimate of τ with its maximum-likelihood estimate. Intrinsic loss intervals can only be computed numerically for this problem. How do they compare with likelihood intervals? With frequentist intervals?

Exercise 3: Graphical representation of search procedures

Suppose we make a measurement X that has a Gaussian distribution with unknown mean θ and unit width. Suppose also that the value $\theta = 0$ has the special physical significance of "no signal", whereas $\theta > 0$ represents "signal". In this simple model, the measurement sensitivity can be quantified by the difference $\Delta\theta$ between the θ values under a given signal hypothesis and under the no-signal hypothesis. Following the discussion in section 6.2 about sensitivity, make a plot with p_1 along the y axis and p_0 along the x axis, and draw contours of constant $\Delta\theta$ (i.e. for a fixed value of $\Delta\theta$, how does p_1 vary with p_0 as the data X run through its range?) Note that the line of no sensitivity, $\Delta\theta = 0$, coincides with the second diagonal. Draw the line $p_0 = \epsilon$, corresponding to the threshold for rejecting the no-signal hypothesis. A line at $p_1 = \alpha$ corresponds to the standard frequentist exclusion limit. Draw the sensitivity sets $\mathcal{S}$ defined in equations ((83) and (84)) and draw the $\mathrm{CL_s}$ threshold of equation (85). Note that under the no-signal hypothesis p_0 values are uniformly distributed between 0 and 1. Therefore the standard frequentist probability of excluding a signal hypothesis is given by 1 minus the abscissa of the intersection of the corresponding $\Delta\theta$ contour with the line $p_1 = \alpha$. Show how this probability of exclusion is non-zero even when there is no sensitivity. Show how the $\mathrm{CL_s}$ criterion and the other two methods avoid this problem.

References

[Ber05] J. M. Bernardo. “Intrinsic credible regions: an objective Bayesian approach to interval estimation”. In: *Test* 14 (2005), p. 317 (cit. on p. 26).

[BS94] J. M. Bernardo and A. F. M. Smith. *Bayesian theory*. John Wiley & Sons, 1994 (cit. on pp. 26, 27).

[Bon88] J. V. Bondar. “Discussion of ”Conditionally acceptable frequentist solutions” by G. Casella”. In: *Statistical decision theory and related topics IV, Vol. 1*. Ed. by S. S. Gupta and J. O. Berger. Springer, 1988, p. 91 (cit. on p. 7).

[Cai05] T. Tony Cai. “One-sided confidence intervals in discrete distributions”. In: *J. Statist. Plan. Inf.* 131 (2005), p. 63 (cit. on p. 30).

[CB00] James Carpenter and John Bithell. “Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians”. In: *Stat. Med.* 19 (2000), p. 1141 (cit. on p. 18).

[CL00] Chin-Shan Chuang and Tze Leung Lai. “Hybrid resampling methods for confidence intervals”. In: *Statist. Sinica* 10 (2000), p. 1 (cit. on p. 22).

[CP34] C. J. Clopper and E. S. Pearson. “The use of confidence or fiducial limits illustrated in the case of the binomial”. In: *Biometrika* 26 (1934), p. 404 (cit. on p. 14).

[Cou11] R. D. Cousins. “Negatively biased relevant subsets induced by the most-powerful one-sided upper confidence limits for a bounded physical parameter”. arXiv:1109.2023v1. Sept. 2011 (cit. on p. 36).

[CH92] Robert D. Cousins and Virgil L. Highland. “Incorporating systematic uncertainties into an upper limit”. In: *Nucl. Instr. and Meth. A* 320 (1992), p. 331 (cit. on p. 22).

[CHT10] Robert D. Cousins, Kathryn E. Hymes, and Jordan Tucker. “Frequentist evaluation of intervals estimated for a binomial parameter and for the ratio of Poisson means”. In: *Nucl. Instr. and Meth. A* 612 (2010), p. 388 (cit. on p. 15).

[Cow+11] Glen Cowan et al. “Asymptotic formulae for likelihood-based tests of new physics”. In: *Eur. Phys. J. C* 71 (2011), p. 1554 (cit. on p. 17).

[Cox58] D. R. Cox. “Some problems connected with statistical inference”. In: *Ann. Math. Statist.* 29 (1958), p. 357 (cit. on p. 7).

[DHY03] A. C. Davison, D. V. Hinkley, and G. A. Young. “Recent developments in bootstrap methodology”. In: *Statist. Sci.* 18 (2003), p. 141 (cit. on p. 18).

[DJP10] Luc Demortier, Supriya Jain, and Harrison Bertrand Prosper. “Reference priors for high energy physics”. In: *Phys. Rev. D* 82 (2010), p. 034002 (cit. on p. 27).

[DR95] Thomas J. DiCiccio and Joseph P. Romano. “On bootstrap procedures for second-order accurate confidence limits in parametric models”. In: *Statist. Sinica* 5 (1995), p. 141 (cit. on p. 20).

[ET93] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap.* Chapman & Hall, 1993 (cit. on pp. 8, 20).

[FC98] Gary J. Feldman and Robert D. Cousins. "Unified approach to the classical statistical analysis of small signals". In: *Phys. Rev. D* 57 (1998), p. 3873. arXiv: 9711021v2 [physics.data-an] (cit. on pp. 8, 15, 16, 33, 35).

[Gar36] F. Garwood. "Fiducial limits for the Poisson distribution". In: *Biometrika* 28 (1936), p. 437 (cit. on p. 15).

[Hig87] V. Highland. "Estimation of upper limits from experimental data". Temple University preprint C00-3539-38. 1987 (cit. on p. 36).

[HL63] J. L. Hodges Jr. and E. L. Lehmann. "Estimates of location based on rank tests". In: *Ann. Math. Statist.* 34 (1963), p. 598 (cit. on p. 3).

[JR75] F James and M. Roos. "Minuit – A System for Function Minimization and Analysis of the Parameter Errors and Correlations". In: *Comp. Phys. Comm.* 10 (1975), p. 343 (cit. on p. 17).

[Jun99] Thomas Junk. "Confidence level computation for combining searches with small statistics". In: *Nucl. Instr. and Meth. A* 434 (1999), p. 435 (cit. on p. 37).

[Kas+10] Vinay L. Kashyap et al. "On computing upper limits to source intensities". In: *Astrophys. J.* 719 (Aug. 2010), p. 900 (cit. on p. 36).

[KW96] Robert E. Kass and Larry Wasserman. "The selection of prior distributions by formal rules". In: *J. Amer. Statist. Assoc.* 91 (1996), p. 1343 (cit. on p. 2).

[MS00] M. Mandelkern and J. Schultz. "The statistical analysis of Gaussian and Poisson signals near physical boundaries". In: *J. Math. Phys.* 41 (2000), p. 5701. arXiv: 9910041v3 [hep-ex] (cit. on p. 34).

[Ney37] J. Neyman. "Outline of a theory of statistical estimation based on the classical theory of probability". In: *Phil. Trans. Roy. Soc. London A* 236 (1937), p. 333 (cit. on p. 4).

[Pra61] John W. Pratt. "Length of confidence intervals". In: *J. Amer. Statist. Assoc.* 56 (1961), p. 549 (cit. on p. 2).

[Pre+07] William H. Press et al. *Numerical recipes, the art of scientific computing.* 3rd ed. Cambridge University Press, 2007 (cit. on p. 14).

[Pun06] Giovanni Punzi. "Ordering algorithms and confidence intervals in the presence of nuisance parameters". In: *Statistical problems in particle physics, astrophysics and cosmology. Proceedings of PHYSTAT05.* Ed. by Louis Lyons and Müge Karagöz Ünel. Imperial College Press, 2006, p. 88 (cit. on p. 21).

[Rea02] A. L. Read. "Presentation of search results: the CLs technique". In: *J. Phys. G: Nucl. Part. Phys.* 28 (2002), p. 2693 (cit. on p. 37).

[SWW09] Bodhisattva Sen, Matthew Walker, and Michael Woodroofe. "On the unified method with nuisance parameters". In: *Statist. Sinica* 19 (2009), p. 301 (cit. on p. 22).

[TC05] Fredrik Tegenfeldt and Jan Conrad. "On Bayesian treatment of systematic uncertainties in confidence interval calculation". In: *Nucl. Instr. and Meth. A* 539 (2005), p. 407 (cit. on p. 22).

[Was89] L. A. Wasserman. "A robust Bayesian interpretation of likelihood regions". In: *Ann. Statist.* 17 (1989), p. 1387 (cit. on p. 26).

[Wil38] S. S. Wilks. "The large-sample distribution of the likelihood ratio for testing composite hypotheses". In: *Ann. Math. Statist.* 9 (1938), p. 60 (cit. on p. 17).