DRAFT 0.1

Glen Cowan
1 July, 2013

Comment on use of a different distributions for constraining nuisance parameters

Suppose an analysis contains a parameter of interest μ and a nuisance parameter θ. Sometimes the uncertainty on θ is characterized by the statement that its value could lie anywhere in an interval $[a, b]$. It can be tempting to interpret this as meaning that one's degree of belief about the true value of the parameter is uniformly distributed between a and b and is zero outside these limits. In a Bayesian analysis this is equivalent to taking a uniform (box) distribution for the prior pdf

$$\pi_\theta(\theta) = \begin{cases} \frac{1}{b-a} & a \leq \theta \leq b \, , \\ \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Bayes' theorem is used to find the posterior probability for θ given the data, that we denote here as x,

$$p(\theta|x) \propto L(x|\mu,\theta)\pi_\mu(\mu)\pi_\theta(\theta) \; . \tag{2}$$

Here we have assumed that the prior pdfs for μ and θ factorize, which is to say the two parameters are independent. Inference about the parameter of interest μ is then obtained by integrating (marginalizing) over θ,

$$p(\mu|x) = \int p(\mu, \theta|x) \, d\theta \; . \tag{3}$$

It is difficult to imagine, however, that one's degree of belief in θ changes from a constant finite value just inside the interval to zero just outside. The usual case is that if one only has a best estimate $\tilde{\theta}$ for θ, then there are many possible ways how or reasons why this could depart from the true value θ. The deviation $\theta - \tilde{\theta}$ is the sum of all of these contributions. One can then argue on the basis of the central limit theorem that a more realistic model for one's uncertainty is a Gaussian distribution.

On the other hand, the tails of a Gaussian distribution fall off extremely quickly and this is also often an unrealistic model for one's genuine uncertainty. To have a bell-shaped curve that has longer tails one can use a prior based on a Student's t distribution,

$$\pi(\theta) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} \; . \tag{4}$$

Here

$$t = \frac{\theta - \tilde{\theta}}{\lambda} \; , \tag{5}$$

$\tilde{\theta}$ is the best estimate of θ, λ is a scale parameter that determines the width of the distribution. The parameter ν is the number of degrees of freedom, which controls the tails. This is a continuous parameter defined for $\nu \geq 1$. For $\nu \to \infty$ the Student's t becomes a Gaussian distribution; for $\nu = 1$ it is the Cauchy distribution, which has infinite variance. For $\nu > 2$ the variance is

$$\sigma_t^2 = \frac{\nu}{\nu - 2} \tag{6}$$

so that the standard deviation of the prior $\pi(\theta)$ is

$$\sigma_\theta = \lambda \sqrt{\frac{\nu}{\nu - 2}} \; . \tag{7}$$

For any of the priors mentioned above (box, Gaussian, Student) one can ask what the equivalent (or closest matching) frequentist procedure would be. To answer this one can ask what information led to the degree of belief encapsulated by $\pi(\theta)$, which is centred about our best estimate $\tilde{\theta}$. Suppose, for example, that before we obtained the estimate $\tilde{\theta}$ we had "no information" about θ and on this basis we take the prior to be a constant.[1] Suppose further that we treat the best estimate $\tilde{\theta}$ as a measured quantity with a likelihood $L(\tilde{\theta}|\theta)$. That is, if we were to repeat the measurement many times, the sampling distribution of $\tilde{\theta}$ would follow the distribution given by this likelihood.

We can then ask what form $L(\tilde{\theta}|\theta)$ would have in order to obtain, starting from a constant prior for θ, one of the priors mentioned above. This "intermediate" prior will be obtained from the original constant prior, $\pi_0(\theta) =$ const. and the likelihood $L(\tilde{\theta}|\theta)$ using Bayes' theorem:

$$\pi(\theta|\tilde{\theta}) \propto L(\tilde{\theta}|\theta)\pi_0(\theta) \; . \tag{8}$$

Because $\pi_0(\theta)$ is constant, we simply need a likelihood that satisfies

$$L(\tilde{\theta}|\theta) \propto \pi(\theta|\tilde{\theta}) \; . \tag{9}$$

In all three cases above (box, Gaussian, Student) this is a sampling distribution with the same functional form as the prior. For example, for the box one imagines that $\tilde{\theta}$ follows a sampling distribution given by the likelihood

$$L(\tilde{\theta}|\theta) = \begin{cases} \frac{1}{\Delta\theta} & |\tilde{\theta} - \theta| < \Delta\theta/2 \; , \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

This follows directly from the prior (1) centred about $\tilde{\theta} = (a+b)/2$ and having a width $\Delta\theta = b - a$.

One can also imagine examples of priors for which it is not possible to write down a meaningful frequentist equivalent. Suppose, for example, one would like to represent uncertainty in a parameter θ by an exponential distribution

$$\pi(\theta) = \frac{1}{\tilde{\theta}} e^{-\theta/\tilde{\theta}} \; . \tag{11}$$

[1] Having "no information" about a parameter is not really well defined. For example, a uniform distribution about θ implies a nonuniform pdf for a nonlinear function of θ.

If we suppose that this prior emerged from an earlier constant prior, then it is not possible to write down a normalizable sampling distribution for $\tilde{\theta}$ that would lead to Eq. (11).

Let us suppose, however, that we are able to regard $\tilde{\theta}$ as a measured quantity with a likelihood $L(\tilde{\theta}|\theta)$. Then the full likelihood for the rest of the measured quantities, which we are writing as x, and $\tilde{\theta}$ is

$$L(x, \tilde{\theta}|\mu, \theta) = L_x(x|\mu, \theta) L_{\tilde{\theta}}(\tilde{\theta}|\theta) \; . \tag{12}$$

where subscripts x and $\tilde{\theta}$ have been introduced to indicate the parts of the likelihood that describe the corresponding measured quantities.

One can now use the likelihood (12) in a frequentist analysis where the nuisance parameter θ is eliminated by forming the profile likelihood

$$L_{\rm p}(\mu) = L(x, \tilde{\theta}|\mu, \hat{\hat{\theta}}(\mu)) \; . \tag{13}$$

Using this in a test of μ based, e.g., on the profile likelihood ratio

$$\lambda(\mu) = \frac{L_{\rm p}(\mu)}{L(\hat{\mu}, \hat{\theta})} \tag{14}$$

will lead to a result that is similar but not identical to the Bayesian result obtained from the marginalization integral (3).