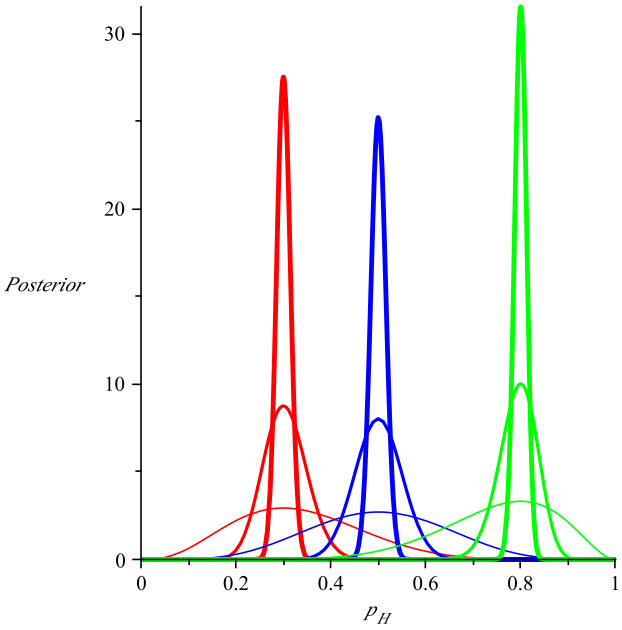

Figure 2: Posterior probability for the coin bias for $N = 10, 100, 1000$ (from thin to thick) and $H/N = 0.3, 0.5, 0.8$ (from left to right).

2 Single parameter inference

1. A coin is tossed N times and heads come up H times [...]

   Answer

   The likelihood function for p_H is given by the binomial

$$\mathcal{L}(p_H) = P(H|p_H, N) = \binom{N}{H} p_H^H (1-p_H)^{N-H}. \tag{15}$$

   If we choose a flat prior over p_H, *i.e.*, uniform over $0 \le p_H \le 1$, the posterior is numerically identical to the likelihood, apart from the normalizing constant (the evidence), and we obtain:

$$P(p_H|H, N) = \frac{\mathcal{L}(p_H)}{Z} \tag{16}$$

   where

$$Z = \binom{N}{H} \int_0^1 \mathrm{d}p_H \, p_H^H (1-p_H)^{N-H} = \binom{N}{H} \frac{H!(N-H)!}{(N+1)!}. \tag{17}$$

   For $N = 100$, the posterior on p_H is plotted in Fig. 2 for a few choices of H and $N = 10, 100, 1000$.

   A measure of the uncertainty of our estimate for p_H is the standard deviation of the posterior, which becomes very close to Gaussian for large N and not too small H, as apparent from Fig. 2. We can estimate the standard deviation by expanding the posterior to second order in p_H around the maximum, and the standard deviation is then given by minus the curvature of the log-posterior at the peak:

$$P(p_H|H, N) \approx P_0 \exp\left(-\frac{1}{2}\frac{(p_{\mathrm{ML}} - p)^2)}{\Sigma^2}\right), \tag{18}$$

where

$$\begin{aligned}\Sigma^{-2} &= -\frac{\partial^2 \ln P(p_H|H,N)}{\partial p^2}\Big|_{p=p_{\rm ML}} = -\frac{\partial}{\partial p}\left(\frac{H}{p}-\frac{N-H}{1-p}\right)\Big|_{p=p_{\rm ML}} \\ &= \frac{H-2Hp+p^2N}{p^2(1-p)^2}\Big|_{p=p_{\rm ML}} = \frac{N}{\frac{H}{N}(1-H/N)}.\end{aligned} \tag{19}$$

and $p_{\rm ML}$ is the maximum of the posterior, given by $p_{\rm ML} = H/N$ (as apparent from derivating Eq. (16) wrt p_H and setting it to 0). So the standard deviation of the posterior is approximately given by

$$\Sigma \approx \frac{(\frac{H}{N}(1-H/N))^{1/2}}{\sqrt{N}}. \tag{20}$$

The probability of the $(N+1)$-th flip giving heads, given that in the past N flips we obtained H heads is given by:

$$P((N+1)\text{-th} = H|N,H) = \int \mathrm{d}r\, P((N+1)\text{-th} = H|p_H)P(p_H|N,H) \tag{21}$$

i.e., the prior predictive distribution for the next flip is the average of the likelihood for that flip, $P((N+1)\text{-th} = H|p_H)$ over the current posterior, $P(p_H|N,H)$ integrated over the parameter. This gives

$$P((N+1)\text{-th} = H|N,H) = \int \mathrm{d}p_H\, \frac{p_H^H(1-p_H)^{N-H}}{Z} = \frac{H+1}{N+2}. \tag{22}$$

So for example, for $N=0$, $H=0$ you recover $P(\text{first} = H|0,0) = 1/2$, *i.e.*, if you haven't observed anything yet you have a 50% probability of heads in your first trial. for $N=10$, $H=9$ you predict $P(11\text{-th}|10,9) = 83.3\%$, *i.e.*, you are becoming confident that your coin is biased. For $N=100$, $H=90$ you get $P(101\text{-st}|100,90) = 89.2\%$ and so on. Notice that, as you would expext, for $H=N/2$ (*i.e.*, over N flips exactly half of them were heads) your prediction reduces to $(H+1)/(2H+2) = 1/2$, as it should.

2. An astronomer wishes to know the (mono-chromatic) flux of a particular source [...]

Answer

(a) The true flux of the source, $F_{\rm src}$. (Even though this is a definite physical number, it is reasonable to consider it's value in probabilistic terms, as it is not uniquely/logically determined by the data.)

(b) The datum is $N_{\rm src}$, the number of photons registered in the measurement of the source.

(c) The starting point for answering this question is to see that photons from the source hit the detector at a given rate ($F_{\rm src}/C$ per unit observation time) but that the photons propagate independently. This implies that the number of photons that hit the detector in a given period is Poisson distributed, and so

$$\Pr(N_{\rm src}|F_{\rm src}) = \frac{(F_{\rm src}/C)^{N_{\rm src}}e^{-F_{\rm src}/C}}{N_{\rm src}!}. \tag{23}$$

In the case of bright sources, for which $F_{\rm src}/C \gg 1$, the distribution of $N_{\rm src}$ is still Poisson, although mathematically extremely well approximated as a Gaussian of the form

$$\Pr(N_{\rm src}|F_{\rm src}) \propto \frac{1}{(F_{\rm src}/C)^{1/2}} e^{-1/2(N_{\rm src}-F_{\rm src}/C)^2/(F_{\rm src}/C)}, \tag{24}$$

where, in the large $N_{\rm src}$ limit, it is being treated as a continuous variable. This equation is no longer correctly normalised as an awkward sum over $N_{\rm src}$ must be done; however the relative

probabilities of the different possible $N_{\rm src}$ values for a given $F_{\rm src}$ are correct. More importantly, the likelihood is a smooth function of $F_{\rm src}$, and it is this interpretation that will be required for later inference. However, whilst $\Pr(N_{\rm src}|F_{\rm src})$ is a Gaussian in $N_{\rm src}$, it is not Gaussian in terms of $F_{\rm src}$, as $F_{\rm src}$ appears in the normalising constant and in the denominator of the exponential.

It is important not only to obtain the mathematical form of the likelihood but also to understand what it means. It is *not* the probability of $F_{\rm src}$, even though in some cases it might have a similar form (*e.g.*, peaked in the same place, or with a similar spread). It *is* only the probability that $N_{\rm src}$ photons would be received from the source *if* its flux was $F_{\rm src}$.

(d) You, as an astronomer, are very far from total ignorance about astronomical sources and their fluxes. If you know the type of the source (*e.g.*, a quasar or a Galactic star, *etc.*) then previous astronomical knowledge about all sorts of astronomical sources. Even without any particular knowledge about the type of source, there is the generic fact that, due to geometry, there are significantly more faint sources than bright sources. The immediate implication is that, in any situation where the data do not strongly constrain the source's flux, it will be important to include the preponderence of faint sources in the prior.

(e) The complicated nature of astronomical surveys – and particular their attendant selection effects – makes this a potentially difficult question to answer. However the underlying principle is that the observed flux distribution of the sources in question would serve as a good, if approximate, prior for the flux of the source of interest.

(f) The prior implied is (up to a normalisation constant)

$$\Pr(F_{\rm src})\Theta(F_{\rm src}) \propto F_{\rm src}^{-5/2}, \tag{25}$$

where $\Theta(x)$ is the Heavyside step function, to ensure that the prior is zero for negative fluxes. This might seem a little fussy, but in exploring an unfamiliar problem it is generally worth being more careful/explicit about the assumptions you're making.

The posterior distribution of the source's true flux would then be (up to a normalisation constant)

$$\Pr(F_{\rm src}|N_{\rm src}) \propto \Theta(F_{\rm src})(F_{\rm src}/C)^{N_{\rm src}-5/2}e^{-F_{\rm src}/C}. \tag{26}$$

In the limit of a large number of photons, the Gaussian approximatin invoked above leads to the posterior

$$\Pr(F_{\rm src}|N_{\rm src}) \propto \Theta(F_{\rm src})F_{\rm src}^{-3}e^{-1/2(N_{\rm src}-F_{\rm src}/C)^2/(F_{\rm src}/C)}. \tag{27}$$

The prior is not normaliseable unless a minimum flux, $F_{\rm min}$ is assumed (or justified somehow), and so care must be taken with these posteriors to check that they are normaliseable. The obvious potential problem is as $F_{\rm src} \to 0$, as it is here that the improper prior becomes infinite. The prior diverges as a power-law, as does the likelihood, when expressed as a function of $F_{\rm src}$, although the latter is dominant provided $N_{\rm src} > 5/2$, so the posterior is bounded and integrable. The Gaussian approximation does not have this property, however, and the likelihood is finite, if very small, at $F_{\rm src} = 0$, leading to a sharp "spike" in the posterior at $F_{\rm src} = 0$ that contains infinite probability. This is an artefact of the Gaussian approximation to the Poisson likelihood and is not a serious problem in practice.

(g) The likelihoods and unnormalised posterior distributions are shown in Fig. 3. In the $N_{\rm src} = 5$ case the full Poisson formula is used; in the $N_{\rm src} = 10^4$ case the Gaussian approximation is adopted. In the latter case the posterior and likelihood are almost indistinguishable and also both very close to Gaussian. The prior does not play a strong role as the high-precision measurement is much more informative. In the $N_{\rm src} = 5$ case, however, the measurement contains far less information and the source is probably fainter than the data might naively be taken to indicate.

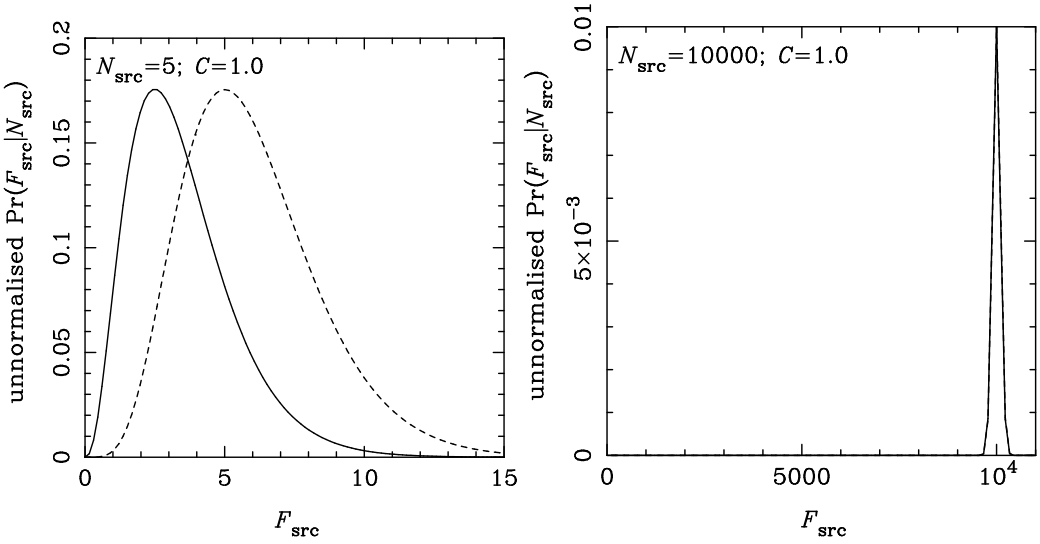

Figure 3: Unnormalised posterior in the source flux, $F_{\rm src}$ in the cases where $N_{\rm src}$ = 5 (left) and $N_{\rm src}$ = 10^4 (right). In both cases the dashed lines show the likelihood as a function of $F_{\rm src}$.

(h) The full answer to any Bayesian parameter estimation problem is the posterior distribution in the parameter(s) of interest. However in many practical situations (*e.g.*, reporting flux estimates of millions of sources) there is no way of assimilating or visualising the full distribution. Hence it is useful to try and condense it into, *e.g.*, an estimated value and an error. That said, there can be no definitive algorithm for doing this. In some cases a few parameters can completely encapsulate the posterior (*e.g.*, the mean/mode/median and standard deviation if it's Gaussian), but in most cases this is not strictly possible.

For singly-peaked distributions it is reasonable to use the peak of the posterior, or the median or the mean. Whichever of these characterising numbers is chosen will be less than the "natural" estimator, $\hat{F}_{\rm src} = CN_{\rm src}$. This result is potentially counter-intuitive, especially if you've gotten used to using sampling statistis. One of the first tests many people would run to test an algorithm being used to estimate some quantity of interest would be to generate lots of fake data with the flux equal to some known $F_{\rm src}$ and then see if the resultant estimates (from the peak or mean or whatever) are centred around the true value. Bayesian estimates do *not* satisfy this test (unless the prior happens to be symmetric about $F_{\rm src}$). The reason is that the prior distribution reflects the distribution of source fluxes in the Universe, which is explicitly contradicted if one simulates data with a single flux value.

Put another way, in any real astronomical measurement most of the sources with photon counts $N_{\rm src}$ will have true fluxes which are less than $F_{\rm src} = CN_{\rm src}$ as there are more faint sources which are randomly scattered bright than there are brighter sources scattered faint. This phenomenon has long been known as Eddington bias, where the term "bias" is used because of the fact that conventional flux estimates are biased high. In terms of Bayesian statistics it would simply be the result of having made a poor choice of prior (that didn't reflect the prevalence of faint sources).

3. This problem takes you through the steps to derive the posterior distribution for a quantity of interest [...]

Answer

(a) The likelihood is given by

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\theta - \hat{x}_i)^2}{\sigma^2}\right). \tag{28}$$

Consider now the exponential term:

$$\begin{aligned} &\frac{1}{2}\sum_i \frac{(\theta - \hat{x}_i)^2}{\sigma^2} = \frac{1}{2\sigma^2}\left(N\theta^2 - 2\sum_i \hat{x}_i\theta + \sum_i \hat{x}_i^2\right) \\ &= \frac{N}{2\sigma^2}\left(\theta^2 - 2\theta\bar{x} + \bar{x}^2 - \bar{x}^2 + \frac{1}{N}\sum_i \hat{x}_i^2\right) = \frac{N}{2\sigma^2}(\theta - \bar{x})^2 + \frac{N}{2\sigma^2}\left(\frac{1}{N}\sum_i \hat{x}_i^2 - \bar{x}^2\right) \end{aligned} \tag{29}$$

So the likelihood can be written as

$$L(\theta) = L_0 \exp\left(-\frac{1}{2}\frac{(\theta - \bar{x})^2}{\sigma^2/N}\right), \tag{30}$$

where L_0 is a constant that does not depend on θ.

(b) The posterior pdf for θ is proportional to the likelihood times the prior (dropping the normalization constant in Bayes' Theorem):

$$p(\theta|\hat{x}) \propto \mathcal{L}(\theta)p(\theta) \propto \exp\left(-\frac{1}{2}\frac{(\theta - \bar{x})^2}{\sigma^2/N}\right)\exp\left(-\frac{1}{2}\frac{\theta^2}{\Sigma^2}\right), \tag{31}$$

where we have dropped normalization constants which do not depend on θ and we have used the Gaussian form of the prior. Collecting terms that depend on θ in the exponent and completing the square we get

$$p(\theta|\hat{x}) \propto \exp\left(-\frac{1}{2}\frac{(\theta - \bar{x}\frac{\Sigma^2}{\Sigma^2 + \frac{\sigma^2}{N}})^2}{\left[\frac{1}{\Sigma^2} + \frac{N}{\sigma^2}\right]^{-1}}\right), \tag{32}$$

which shows that the posterior for θ is a Gaussian with the mean and variance as given in the question.

(c) When $N \to \infty$, we have that the variance $\left[\frac{1}{\Sigma^2} + \frac{N}{\sigma^2}\right]^{-1} \to \sigma^2/N$ (as $\frac{N}{\sigma^2} \gg \frac{1}{\Sigma^2}$) and the mean $\bar{x}\frac{\Sigma^2}{\Sigma^2 + \frac{\sigma^2}{N}} \to \bar{x}$ (as $\Sigma^2 \gg \frac{\sigma^2}{N}$ and the fraction goes to unity). Thus the posterior pdf becomes

$$p(\theta|\hat{x}) \to \exp\left(-\frac{1}{2}\frac{(\theta - \bar{x})^2}{\sigma^2/N}\right), \tag{33}$$

which shows that the posterior converges to the likelihood and the prior dependence disappears.

(d) From the above result, we can use the posterior pdf to compute the posterior mean of θ:

$$\langle\theta\rangle = \int \theta\, p(\theta|\hat{x})\, d\theta = \bar{x}. \tag{34}$$

Therefore the posterior mean tends to the sample mean, $\bar{x}$, which as we know is also the MLE for the mean.