1 Coin tossing

The solutions are:

(a) The likelihood function is given by

$$\mathcal{L}(p) = P(r = H|p, n) = \binom{n}{H} p^H (1-p)^{n-H}, \tag{1}$$

where the unknown parameter is p and the data are the number of heads, H (for a fixed number of trials, $n = 10$ here).

(b) The Maximum Likelihood Estimator (MLE) for the success probability p is found by maximising the log likelihood, see Example 20 in the handout:

$$\begin{aligned} \frac{\partial \ln \mathcal{L}(p)}{\partial p} &= \frac{\partial}{\partial p}\left(\ln \binom{n}{H} + H \ln p + (n-H)\ln(1-p)\right) = \frac{H}{p} - \frac{n-H}{1-p} \stackrel{!}{=} 0 \\ &\Leftrightarrow p_{\rm ML} = \frac{H}{n}. \end{aligned} \tag{2}$$

Therefore the ML value for p is $p_{\rm ML} = 0.8$.

(c) We approximate the likelihood function as a Gaussian, with standard deviation given by minus the curvature of the log-likelihood at the peak:

$$\mathcal{L}(p) \approx \mathcal{L}_{\rm max} \exp\left(-\frac{1}{2}\frac{(p_{\rm ML} - p)^2)}{\Sigma^2}\right), \tag{3}$$

where (see Eq. (65) in the handout)

$$\begin{aligned} \Sigma^{-2} &= -\frac{\partial^2 \ln \mathcal{L}(p)}{\partial p^2}\Big|_{p=p_{\rm ML}} = -\frac{\partial}{\partial p}\left(\frac{H}{p} - \frac{n-H}{1-p}\right)\Big|_{p=p_{\rm ML}} \\ &= \frac{H - 2Hp + p^2 n}{p^2(1-p)^2}\Big|_{p=p_{\rm ML}} = \frac{n}{\frac{H}{n}\left(1-\frac{H}{n}\right)}. \end{aligned} \tag{4}$$

The 1σ confidence interval for p is given by $\Sigma = 0.13$. Therefore the result would be reported as $p = 0.80 \pm 0.13$.

(d) Following the hint, the number of σ confidence with which the hypothesis that the coin is fair can be ruled out is given by

$$\frac{|p_{\rm ML} - \frac{1}{2}|}{\Sigma} = \frac{0.8 - 0.5}{0.13} = 2.31. \tag{5}$$

Therefore the fairness hypothesis can be ruled out at the $\sim 2.3\ \sigma$ level.

(e) Using above equations, the MLE for the success probability is still $p_{\rm ML} = 0.8$, as before. However, the uncertainty is now much reduced, because of the large number of trials. In fact, we get $\Sigma = 0.013$ (notice how the uncertainty has decreased by a factor of $\sqrt{n}$, as expected. I.e., 100 times more trials correspond to a reduction in the uncertainty by a factor of 10). The fairness hypothesis can now be excluded with much higher confidence:s of $p = 1/2$, expressed in number of sigmas:

$$\text{number of sigmas} = \frac{|p_{ML} - \frac{1}{2}|}{\Sigma} = \frac{0.8 - 0.5}{0.013} = 23.1 \approx 23. \tag{6}$$

This constitutes more than decisive evidence against the hypothesis that the coin is fair. Notice however that the Gaussian approximation to the likelihood we employed will most probably not be accurate so far into the tails of the likelihood function (i.e., the Taylor expansion on which it is based is a *local* expansion around the peak).

2 Counting experiment

(a) The discrete PMF for the number of counts r of a Poisson process with average rate λ is (assuming a unit time, $t = 1$ throughout)

$$P(r) = \frac{\lambda^r}{r!} e^{-\lambda}$$

(b) In this case

$$P(\hat{r}_i|\,\lambda) = \frac{\lambda^{\hat{r}_i}}{\hat{r}_i!} e^{-\lambda}\,,$$

for each independent measurement $\hat{r}_i$. So the joint likelihood is given by (as measurements are independent)

$$\mathcal{L}(\lambda) = \prod_{i=1}^{M} P(\hat{r}_i|\,\lambda) = \prod_{i=1}^{M} \frac{\lambda^{\hat{r}_i}}{\hat{r}_i!} e^{-\lambda}\,. \tag{7}$$

(c) The Maximum Likelihood Principle states that the estimator for λ can be derived by finding the maximum of the likelihood function. The maximum is found more easily by considering the log of the likelihood

$$\ln \mathcal{L}(\lambda) = \sum_{i=1}^{M} \left[\hat{r}_i \ln(\lambda) - \ln(\hat{r}_i!) - \lambda\right]\,.$$

with the maximum given by the condition $d\ln\mathcal{L}/d\lambda = 0$.
We have

$$\begin{aligned} \frac{d\ln\mathcal{L}}{d\lambda} &= \sum_{i=1}^{M} \left[\frac{\hat{r}_i}{\lambda} - 1\right] \\ &= \frac{1}{\lambda}\sum_{i=1}^{M} \hat{r}_i - M\,. \end{aligned}$$

So the Maximum Likelihood (ML) estimator for λ is

$$\lambda_{\rm ML} = \frac{1}{M}\sum_{i=1}^{M} \hat{r}_i\,,$$

which is just the average of the observed counts.

(d) The Taylor expansion is (see Eq. (63) in the handout)

$$\ln\mathcal{L}(\lambda) = \ln\mathcal{L}(\lambda_{\rm ML}) + \left.\frac{d\ln\mathcal{L}}{d\lambda}\right|_{\lambda=\lambda_{\rm ML}} (\lambda - \lambda_{\rm ML}) + \frac{1}{2}\left.\frac{d^2\ln\mathcal{L}}{d\lambda^2}\right|_{\lambda=\lambda_{\rm ML}} (\lambda - \lambda_{\rm ML})^2 + \ldots\,.$$

By definition the linear term vanishes at the maximum so we just need the curvature around the ML point

$$\frac{d^2\ln\mathcal{L}}{d\lambda^2} = -\sum_{i=1}^{M} \frac{\hat{r}_i}{\lambda^2}\,,$$

such that

$$\left.\frac{d^2\ln\mathcal{L}}{d\lambda^2}\right|_{\lambda=\lambda_{\rm ML}} = -\frac{1}{\lambda_{\rm ML}^2}\sum_{i=1}^{M}\hat{r}_i = -\frac{M\lambda_{\rm ML}}{\lambda_{\rm ML}^2} = -\frac{M}{\lambda_{\rm ML}}\,.$$

Putting this into the Taylor expansion gives

$$\ln \mathcal{L}(\lambda) = \ln \mathcal{L}(\lambda_{\rm ML}) - \frac{1}{2}\frac{M}{\lambda_{\rm ML}}(\lambda - \lambda_{\rm ML})^2 \,,$$

which gives an approximation of the likelihood function around the ML point

$$\mathcal{L}(\lambda) \approx L_0 \exp\left(-\frac{1}{2}\frac{M}{\lambda_{\rm ML}}(\lambda - \lambda_{\rm ML})^2\right) \,,$$

(the normalisation constant L_0 is irrelevant).
So the likelihood is approximated by a Gaussian with variance

$$\Sigma^2 = \frac{\lambda_{\rm ML}}{M} \,.$$

(e) Comparing this with the standard result for the variance of the mean for the Gaussian case, i.e.

$$\Sigma^2 = \frac{\sigma^2}{M} \,,$$

where M is the number of measurements and σ is the standard deviation of each measurement, we can conclude that the variance of the Poisson distribution itself is indeed

$$\sigma^2 = \lambda \,.$$

3 **Gaussian measurements with variable variance**

(a) The photon counts follow a Poisson distribution. We know that the MLE for the Poisson distribution is the observed number of counts (n) and its standard deviation is $\sqrt{n}$. However, for large n ($\gg 20$) the Poisson distribution is well approximated by a Gaussian of mean n and standard deviation $\sqrt{n}$. In this case, n is of order 10^5, hence the standard deviation intrinsic to the Poisson process (the so-called "shot noise") is of order $\sqrt{10^5} \approx 3 \cdot 10^2$. The quoted experimental uncertainty is much larger than that (of order 10^4 for each datum), hence we can conclude that the statistical error is dominated by the noise in the detector rather than by the Poisson variance.
Therefore we can approximate the likelihood for each observation as a Gaussian with mean given by the observed counts $\hat{n}_i$ and standard deviation given by the quoted error, $\hat{\sigma}_i$:

$$\mathcal{L}_i(F) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_i} \exp\left(-\frac{1}{2}\frac{(F - \hat{n}_i)^2}{\hat{\sigma}_i^2}\right) \quad (i = 1, \ldots, 4). \tag{8}$$

(b) Since the measurements are independent, the total likelihood is the product of the 4 terms:

$$\mathcal{L}(F) = \prod_{i=1}^{4} \mathcal{L}_i(F). \tag{9}$$

(c) To estimate the mean of the distribution, we apply the MLE procedure for the mean (F), obtaining:

$$\begin{aligned} \frac{\partial \ln \mathcal{L}(F)}{\partial F} &= -\sum_i \frac{F - \hat{n}_i}{\hat{\sigma}_i^2} \stackrel{!}{=} 0 \\ \Leftrightarrow F_{\rm ML} &= \sum_i \frac{\hat{n}_i}{\hat{\sigma}_i^2/\bar{\sigma}^2}, \end{aligned} \tag{10}$$

where

$$\frac{1}{\bar{\sigma}^2} \equiv \sum_i \frac{1}{\hat{\sigma}_i^2}. \tag{11}$$

We thus see that the ML estimate for the mean is the mean of the observed counts weighted by the inverse error on each on them (verify that Eq. (**??**) reverts to the usual expression for the sample mean for $\hat{\sigma}_i = \hat{\sigma}$ for $(i = 1, \ldots, 4)$, i.e., if all observations have the same error). This automatically gives more weight to observations with a smaller error.

(d) From the given observations, one thus obtains $F_{\rm ML} = 29.2 \times 10^4$ photons/cm^2. By comparison the sample mean is $\bar{F} = 30.3 \times 10^4$ photons/cm^2.

(e) The inverse variance of the mean is given by the second derivative of the log-likelihood evaluated at the ML estimate, see Eq. (65) in the handout:

$$\Sigma^{-2} = -\frac{\partial^2 \ln \mathcal{L}(F)}{\partial F^2}\Big|_{F=F_{\rm ML}} = \sum_i \frac{1}{\hat{\sigma}_i^2}. \tag{12}$$

(again, it is simple to verify that the above formula reverts to the usual $N/\hat{\sigma}^2$ expression if all measurements have the same error).

Therefore the variance of the mean is given by $\Sigma^2 = 2.16 \times 10^8$ (photons/cm^2)^2, and the standard deviation is $\Sigma = 1.47 \times 10^4$ photons/cm^2. Our measurement can thus be summarized as $F = (29.2 \pm 1.5) \times 10^4$ photons/cm^2.

4 Photon counts and source strength

(a) The likelihood function is given by the Poisson distribution

$$\mathcal{L}(\hat{r}) = P(\hat{r}|\lambda) = \frac{(\lambda t)^{\hat{r}}}{\hat{r}!} \exp(-\lambda t), \tag{13}$$

where t is the time of observation in minutes. The unknown parameter is the source strength λ (in units of photons/min), while the data are the observed counts, $\hat{r}$.

(b) We can compute the requested probability by substituting in the Poisson distribution above the values for $\hat{r}$ and λ, obtaining:

$$P(\hat{r} = 15|\lambda = 10, t = 1 \text{ min}) = 0.0347. \tag{14}$$

(c) The maximum likelihood estimate is obtained by finding the maximum of the log likelihood as a function of the parameter (here, the rate λ). Hence we need to find the value of λ such that:

$$\frac{\partial \ln \mathcal{L}(\hat{r})}{\partial \lambda} = 0. \tag{15}$$

The derivative gives (see Example 21 in the handout)

$$\frac{\partial \ln \mathcal{L}(\hat{r})}{\partial \lambda} = \frac{\partial}{\partial \lambda}\left(\hat{r} \ln(\lambda t) - \ln \hat{r}! - \lambda t\right) = \hat{r}\frac{t}{\lambda t} - t = 0 \Leftrightarrow \lambda_{MLE} = \frac{\hat{r}}{t}. \tag{16}$$

So the maximum likelihood estimator for the rate is the observed number of counts divided by the time, in agreement with Eq. (62) in the handout. In this case, $t = 1$ min so the MLE for λ is 10 photons per minute.

(d) The likelihood function now needs to be modified to account for the fact that the observed counts are the superposition of the background rate and the source rate (the star). According

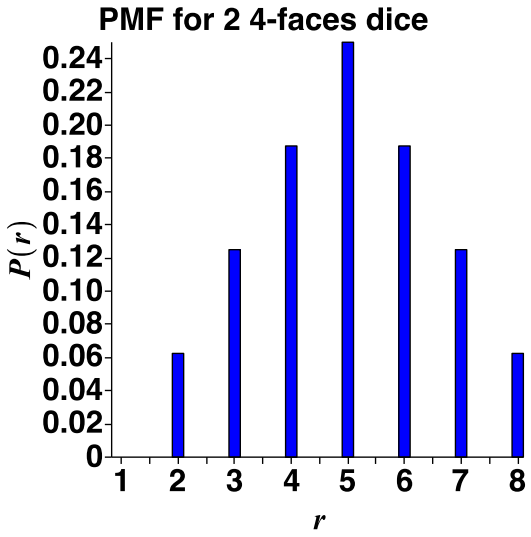

Figure 1: Distribution of the outcomes for the sum of the values of the two dice.

to the hint, the likelihood for the total number counts, $\hat{r}_t$, is Poisson with rate $\lambda = \lambda_s + \lambda_b$, and thus

$$P(\hat{r}_t|\lambda = \lambda_s + \lambda_b) = \frac{(\lambda t_t)^{\hat{r}_t}}{\hat{r}_t!}\exp(-\lambda t_t). \tag{17}$$

Similarly to what we have done above, the MLE estimate for λ_s is found by setting to 0 the derivative of the log likelihood wrt λ_s:

$$\frac{\partial \ln P(\hat{r}_t|\lambda = \lambda_s + \lambda_b)}{\partial \lambda_s} = \hat{r}_t \frac{t_t}{(\lambda_s + \lambda_b)t_t} - t_t = 0 \Leftrightarrow \lambda_s = \frac{\hat{r}_t}{t_t} - \lambda_b. \tag{18}$$

So the MLE for the source is given by the observed average total rate ($\frac{\hat{r}_t}{t_t}$) minus the background rate.

(e) Inserting the numerical results, we have that $\lambda_s = 3$. The MLE estimate for λ_s gives a negative rate if $\hat{r}_t/t_t < \lambda_b$, which is clearly non-physical. However, this can definitely happen because of downwards fluctuations in the number counts due to the Poisson nature of the signal (even if the background is assumed to be known perfectly). So this is an artefact of the MLE estimator (nothing to do with physics! We *know* that the actual physical source rate has to be a non-negative quantity!). The solution is to use Bayes theorem instead.

5 Dice throwing

The possible outcomes for the sum of the values are $\{2, 3, 4, 5, 6, 7, 8\}$, with probabilities given respectively by $\{\frac{1}{16}, \frac{2}{16}, \frac{3}{16}, \frac{4}{16}, \frac{3}{16}, \frac{2}{16}, \frac{1}{16}\}$. Those are plotted in Fig. 1.

For the case of 1000 dice, we appeal to the central limit theorem (CLT): the mean value of one dice is $E(X) = \sum_{i=1}^{4} p_i x_i = 2.5$ and its variance is $Var(X) = \sum_{i=1}^{4}(x_i - E(x))^2 p_i = 5/4$, where $p_i = 1/4$ $(i = 1, \ldots, 4)$. Thus the sum of 1000 such variables will be approximately Gaussian distributed in virtue of the CLT, with mean 2500 and variance $1000 \times 5/4 = 1250$. The standard deviation is therefore $\sqrt{1250} \approx 35$.

6 Bayesian estimation of the flux

(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

$$\mathrm{P}(N_{\rm src}|F_{\rm src}) = \frac{(F_{\rm src}/C)^{N_{\rm src}} e^{-F_{\rm src}/C}}{N_{\rm src}!}. \tag{19}$$

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

$$\mathrm{P}(N_{\rm src}|F_{\rm src}) \propto \frac{1}{(F_{\rm src}/C)^{1/2}} e^{-1/2(N_{\rm src}-F_{\rm src}/C)^2/(F_{\rm src}/C)}, \tag{20}$$

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 $\mathrm{P}(N_{\rm src}|F_{\rm src})$ 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)

$$\mathrm{P}(F_{\rm src})\Theta(F_{\rm src}) \propto F_{\rm src}^{-5/2}, \tag{21}$$

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)

$$\mathrm{P}(F_{\rm src}|N_{\rm src}) \propto \Theta(F_{\rm src})(F_{\rm src}/C)^{N_{\rm src}-5/2} e^{-F_{\rm src}/C}. \tag{22}$$

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

$$\mathrm{P}(F_{\rm src}|N_{\rm src}) \propto \Theta(F_{\rm src}) F_{\rm src}^{-3} e^{-1/2(N_{\rm src}-F_{\rm src}/C)^2/(F_{\rm src}/C)}. \tag{23}$$

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

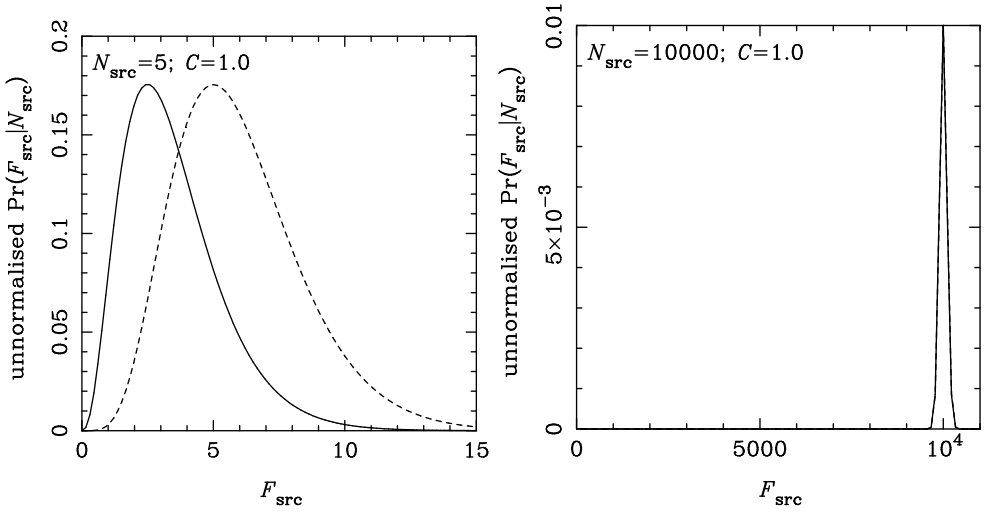

Figure 2: Unnormalised posterior in the source flux, $F_{\rm src}$ in the cases where $N_{\rm src} = 5$ (left) and $N_{\rm src} = 10^4$ (right). In both cases the dashed lines show the likelihood as a function of $F_{\rm src}$.

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. 2. 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.

(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).