

# Ratio of counts vs ratio of rates in Poisson processes

Giulio D'Agostini

Università “La Sapienza” and INFN, Roma, Italia

(giulio.dagostini@roma1.infn.it, <http://www.roma1.infn.it/~dagos>)

## Abstract

The often debated issue of ‘ratios of small numbers of events’ is approached from a probabilistic perspective, making a clear distinction between the predictive problem (forecasting numbers of events we might count under well stated assumptions, and therefore of their ratios) and inferential problem (learning about the relevant parameters of the related probability distribution, in the light of the observed number of events). The quantities of interests and their relations are visualized in a graphical model (‘Bayesian network’), very useful to understand how to approach the problem following the rules of probability theory. In this paper, written with didactic intent, we discuss in detail the basic ideas, however giving some hints of how real life complications, like (uncertain) efficiencies and possible background and systematics, can be included in the analysis, as well as the possibility that the ratio of rates might depend on some physical quantity. The simple models considered in this paper allow to obtain, under reasonable assumptions, closed expressions for the rates and their ratios. Monte Carlo methods are also used, both to cross check the exact results and to evaluate by sampling the ratios of counts in the cases in which large number approximation does not hold. In particular it is shown how to make approximate inferences using a Markov Chain Monte Carlo using JAGS/rjags. Some examples of R and JAGS code are provided.

## 1 Introduction

Many measurements in Physics are based on counting events belonging to a well defined ‘class’. They could be the number of electric pulses, registered within a given time interval, exceeding a properly set threshold, as in a Geiger counter; or the number of events observed, for a given integrated luminosity, in a region defined by properly

chosen ‘cuts’ in the multi-dimensional space defined on the basis of geometrical and kinematic variables of the final state particles, a typical problem in Particle Physics. However, the aim of physicists is not limited in counting how many events *will* occur in each ‘class’ satisfying some detector related criteria, but rather in *inferring* the *physical quantities* which are related to them, as the intensity of radioactivity or the production rate of a given physical final state resulting from the collision of two particles, to continue with our examples. This also implies that the ‘experimentally defined class’ (‘being inside cuts’) is only a proxy for the ‘physical class’ of interest, that might be a radioactive particle in a given energy range, or a particular final state resulting from a collision. This is in analogy with the case when we are interested in counting the number of individuals of a population infected by a specific agent using as a proxy the number of individuals tagged ‘positive’ by suitable tests, by their nature imperfect.<sup>1</sup>

If we change the conditions of the experiment, that is, going on with our examples, we place the Geiger counter in a different place, or we vary the initial energy of the colliding particles (or we tag somehow the final state), we usually register different numbers of events in our reference class. This could just be due to statistical fluctuations. But it could (also) be due to a variation of the related physical quantity. It is then crucial, as well understood, to associate an uncertainty to the ‘measured’ variation.

If the observed numbers are ‘large’, things get rather easy, thanks to the Gaussian approximation of the probability distributions of interest. When, instead, the numbers are ‘small’ the question can be quite troublesome (see, e.g., Refs. [2, 3, 4, 5, 6, 7]). For example, Ref. [3] focus on the “*errors on ratios of small numbers of events*”, leading the readers astray: we are *usually* not interested in the ratios of ‘counts’, but rather on the ratios of radioactivity levels or of production rates, and so on.

The aim of this paper is to review these questions following consistently the rules of probability theory. The initial, crucial point is to make a clear distinction between the empirical observations (the numbers of event of a given ‘experimentally defined class’) and the related physical quantities we are interested to infer, although in a probabilistic way. We start playing with the Poisson distribution in Sec. 2, referring to Appendix A for a reminder of how this distribution is related not only to the binomial (as well known), but also to other important distributions via the *Poisson process*, which has indeed its roots in the *Bernoulli process*. In Sec. 3 we show how to use the Bayes’ rule to infer Poisson  $\lambda$ ’s from the observed number of counts and then how to get the probability distribution of their ratio  $\rho$  making an exact propagation of uncertainties, that is  $f(\lambda_1/\lambda_2)$  from  $f(\lambda_1)$  and  $f(\lambda_2)$ . Then in Sec. 4 we move to

---

<sup>1</sup>This problem has been treated in much detail in Ref. [1], taking cue from questions related to the Covid-19 pandemic.

the inference of *intensities of the Poisson processes* (or ‘rates’  $r$ , in short), related to  $\lambda$  by  $\lambda=r\cdot T$ , with  $T$  being the ‘observation time’ – it can be replaced by ‘integrated luminosity’ or other quantities to which the Poisson parameter  $\lambda$  is proportional. In the same section the ‘anxiety-inducing’ [8] question of the priors, assumed ‘flat’ until Sec. 4.1, is finally tackled and the *conjugate priors* are introduced, showing, in particular, how to apply them in sequential measurements of the *same* rate. The technical question of getting the probability distribution of the ratio of rates is tackled in Sec. 5. Again, closed formulae are ‘luckily’ obtained, which can be extended to the more general problem of getting the probability density function (pdf), and its summaries, of a ratio of Gamma distributed variables.

When the game seems at the end, in Sec. 6 we modify the ‘graphical’ model (indeed a visualization of the underlying logical *causal model*) and restart the analysis, this time really *inferring directly*  $\rho$ , as it will be clear. The implications of the different models and of the priors appearing in each of them will be analyzed with some care. Finally, in Sec. 7 the same models are analyzed making use of Markov Chain Monte Carlo (MCMC) methods, exploiting JAGS. The purpose is twofold. First we want to cross-check the exact results obtained in the previous section, although the latter were limited to uniform priors of the ‘top parents’ of the causal model. Second this allows not only to take into account more realistic priors, but also to enlarge the models including efficiencies and background, for which examples of graphical model are provided. Another interesting question, that is how to fit the ratio of rates as a function of another physical question will be also addressed, showing how to modify the causal model, but without entering into the details. The related issue of ‘combining ratios’ is also discussed and it shows once more the importance of the underlying model.

## 2 Predicting numbers of counts, their difference and their ratio

The Poisson distribution hardly needs any introduction, beside, perhaps, that it can be framed within the *Poisson process*, which has indeed its roots in the *Bernoulli process*. This picture makes the Poissonian related to other important distributions, as reminded in Appendix A, which can be seen as a technical preface to the paper.

Using the notation introduced there, the Poisson probability function is given by<sup>2</sup>

$$f(x | \lambda) \equiv P(X=x | \lambda) = \frac{\lambda^x}{x!} \cdot e^{-\lambda} \quad \begin{cases} 0 < \lambda < \infty \\ x = 0, 1, \dots, \infty \end{cases} \quad (1)$$

and it quantifies how much we believe that  $x$  counts will occur, *if we assume* an exact value of the parameter  $\lambda$ .<sup>3</sup> As well known, expected value and standard deviation of  $X$  are  $\lambda$  and  $\sqrt{\lambda}$ . The most probable value of  $X$  ('mode') is equal to the integer just below  $\lambda$  ('`floor`( $\lambda$ )') in the case  $\lambda$  is not integer. Otherwise is equal to  $\lambda$  itself, and also to  $\lambda - 1$  (remember that  $\lambda$  cannot be null).

If we have two *independent* Poisson distributions characterized by  $\lambda_1$  and  $\lambda_2$ , i.e.

$$\begin{aligned} X_1 &\sim \mathcal{P}_{\lambda_1} \\ X_2 &\sim \mathcal{P}_{\lambda_2}, \end{aligned}$$

we 'expect' their difference  $D = X_1 - X_2$  'to be'  $(\lambda_1 - \lambda_2) \pm \sqrt{\lambda_1 + \lambda_2}$ , as it results from well known theorems of probability theory<sup>4</sup> (hereafter, unless indicated otherwise, the notation ' $xxx \pm yyy$ ' stands for 'expected value of the quantity  $\pm$  its *standard uncertainty*' [9], that is the standard deviation of the associated probability distribution).

The probability distribution of  $D$  can be obtained 'from the inventory of the values of  $X_1$  and  $X_2$  that result in each possible value of  $D$ ', that is

$$P(D=d | \lambda_1, \lambda_2) \equiv f(d | \lambda_1, \lambda_2) = \sum_{\substack{x_1, x_2 \\ x_1 - x_2 = d}} f(x_1 | \lambda_1) \cdot f(x_2 | \lambda_2). \quad (2)$$

For example, in the case of  $\lambda_1 = \lambda_2 = 1$ , the most probable contributions to  $D$  are shown in Tab. 1. For instance, the probability to get  $D = 0$  sums up to 30.9%. The probability decreases symmetrically for larger absolute values of the difference.

---

<sup>2</sup>I try, whenever it is possible, to stick to the convention of capital letters for the name of a variable and small letters for its possible values. Exceptions are Greek letters and quantities naturally defined by a small letter, like  $r$  for a 'rate'.

<sup>3</sup>If, instead, we are uncertain about  $\lambda$  and quantify its uncertainty by the probability density function  $f(\lambda | I)$ , where  $I$  stands for our status of information about that quantity, the distribution of the counts will be given by  $f(x | \lambda, I) = \int_0^\infty f(x | \lambda, I) \cdot f(\lambda | I) d\lambda$ .

<sup>4</sup>In brief: the expected value of a linear combination is the linear combination of the expected values; the variance of a linear combination is the linear combination of the variances, with squared coefficients.

		$X_2$					
		0	1	2	3	4	5
$X_1$	0	<b>0</b> [ <b>0.135335</b> ]	-1 [0.135335]	-2 [0.067668]	-3 [0.022556]	-4 [0.005639]	-5 [0.001128]
	1	1 [0.135335]	<b>0</b> [ <b>0.135335</b> ]	-1 [0.067668]	-2 [0.022556]	-3 [0.005639]	-4 [0.001128]
	2	2 [0.067668]	1 [0.067668]	<b>0</b> [ <b>0.033834</b> ]	-1 [0.011278]	-2 [0.002819]	-3 [0.000564]
	3	3 [0.022556]	2 [0.022556]	1 [0.011278]	<b>0</b> [ <b>0.003759</b> ]	-1 [0.000940]	-2 [0.000188]
	4	4 [0.005639]	3 [0.005639]	2 [0.002819]	1 [0.000940]	<b>0</b> [ <b>0.000235</b> ]	-1 [0.000047]
	5	5 [0.001128]	4 [0.001128]	3 [0.000564]	2 [0.000188]	1 [0.000047]	<b>0</b> [ <b>0.000009</b> ]

Table 1: Table of the most probable differences  $D = X_1 - X_2$  for  $\lambda_1 = \lambda_2 = 1$  (probability of each entry in the table within square brackets).

Without entering into the question of getting a closed form of  $f(d | \lambda_1, \lambda_2)$ ,<sup>5</sup> it can be instructive to implement Eq. (2), although in an approximate and rather inefficient way, in a few lines of R code [11]:<sup>6</sup>

```
dPoisDiff <- function(d, lambda1, lambda2) {
  xmax = round(max(lambda1, lambda2)) + 20*sqrt(max(lambda1, lambda2))
  sum( dpois((0+d):xmax, lambda1) * dpois(0:(xmax-d), lambda2) )
}
```

This function is part of the code provided in Appendix B.1, which produces the plot of Fig. 1, evaluating also expected value and standard deviation (indeed approximated values, being `xmax` not too large).

Moving to the ratio of counts, numerical problems might arise, as shown in Tab. 2, analogue of Tab. 1. In fact for rather small values of  $\lambda_2$  there is high chance (exactly

---

<sup>5</sup>Such a distribution is known in the literature as Skellam distribution [10] and it is available in R [11] installing the homonym package [12]. The distribution of the differences corresponding to the cases of Tab. 1 can be easily plotted by the following R commands, producing a bar plot similar to that of Fig. 1,

```
library(skellam)
d = -5:5
barplot(dskellam(d,1,1), names=d, col='cyan')
```

<sup>6</sup>This function, hopefully having a didactic value, is not optimized at all and it uses the fact that the R function `dpois()` returns zero for negative values of the variable.

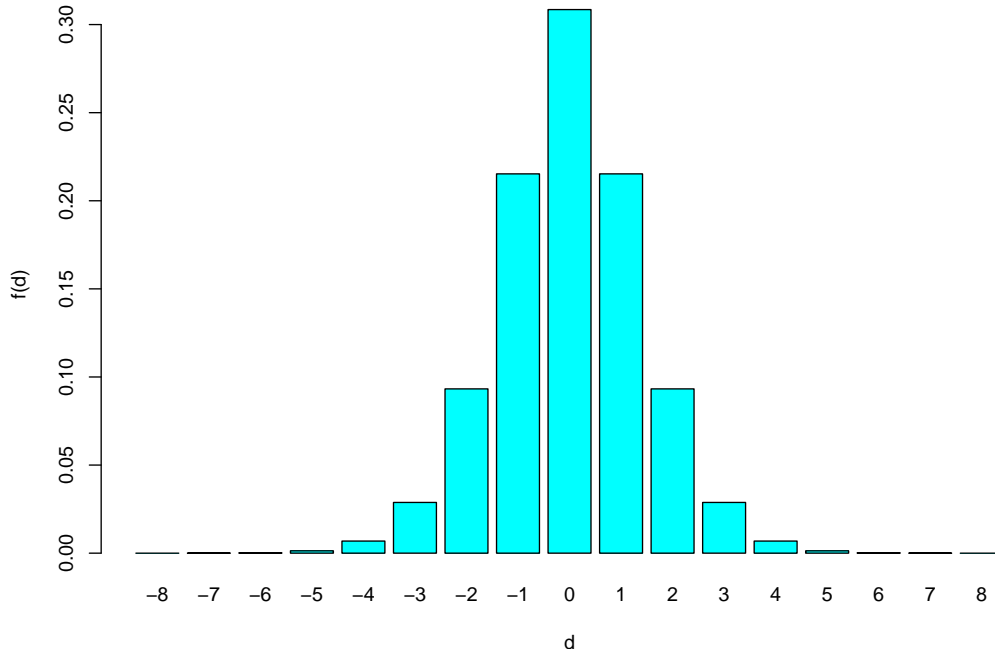


Figure 1: Distribution of the difference of counts resulting from two Poisson distributions with  $\lambda_1 = \lambda_2 = 1$ .

		$X_2$					
		0	1	2	3	4	5
$X_1$	0	NaN [0.135335]	0 [0.135335]	0 [0.067668]	0 [0.022556]	0 [0.005639]	0 [0.001128]
	1	Inf [0.135335]	1 [0.135335]	1/2 [0.067668]	1/3 [0.022556]	1/4 [0.005639]	1/5 [0.001128]
	2	Inf [0.067668]	2 [0.067668]	1 [0.033834]	2/3 [0.011278]	1/2 [0.002819]	2/5 [0.000564]
	3	Inf [0.022556]	3 [0.022556]	3/2 [0.011278]	1 [0.003759]	3/4 [0.000940]	3/5 [0.000188]
	4	Inf [0.005639]	4 [0.005639]	2 [0.002819]	4/3 [0.000940]	1 [0.000235]	4/5 [0.000047]
	5	Inf [0.001128]	5 [0.001128]	5/2 [0.000564]	5/3 [0.000188]	5/4 [0.000047]	1 [0.000009]

Table 2: Table of the most probable ratios  $X_1/X_2$  for  $\lambda_1 = \lambda_2 = 1$ . 'NaN' and 'Inf' are the R symbols for undefined ('not a number') and infinity, resulting from a vanishing denominator.

the probability of getting  $X_2 = 0$ ) that the ratio results in an undefined form or an infinite, reported in the table using the R symbols `NaN` ('not a number') and `Inf`, respectively. As we can see, we have now quite a variety of possibilities and the probability distribution of the ratios is rather irregular. For this reason, in this case we evaluate it by Monte Carlo methods using R.<sup>7</sup> Figure 2 shows the distributions of the ratio for  $\lambda_1 = \lambda_2 = 1, 2, 3$ . The figure also reports the probability to get an infinite or an undefined expression, equal to  $P(X_2 = 0 | \lambda_i)$ . When  $\lambda_2$  is very large the probability to get  $X_2 = 0$ , and therefore of  $X_1/X_2$  being equal to `Inf` or `NaN`, vanishes. But the distribution of the ratio remains quite 'irregular', if looked into detail, even for  $\lambda_1$  'not so small', as shown in Fig. 3 for the cases of  $\lambda_1 = 5, 10, 20, 50$  and  $\lambda_2 = 1000$ .

However we should not be worried about this kind of distributions, which are not more than entertaining curiosities, as long as physics questions are concerned. Why should we be interested in the ratio of counts that we might observe for different  $\lambda$ 's? If we want to get an idea of how much the counts could differ, we can just use the probability distribution of their possible differences, which has a regular behavior, without divergences or undefined forms.

The deep reason for speculating about "*ratios of small numbers of events*" and their "*errors*"[3] is due to a curious ideology at the basis of a school of Statistics which limits the applications of Probability Theory. Indeed, we, as physicists, are often interested in the *ratio of the rates of Poisson processes*, that is in  $\rho = r_1/r_2$ , being this quantity related to some physical quantities having a theoretical relevance. Therefore we aim to learn which values of  $\rho$  are more or less probable in the light of the experimental observations. Stated in this terms, we are interested in evaluating 'somehow' (not always in closed form) the probability density function (pdf)  $f(\rho | x_1, T_1, x_2, T_2, I)$ , given the observations of  $x_1$  counts during  $T_1$  and of  $x_2$  counts during  $T_2$  (and also conditioned on the background state of knowledge, generically indicated by  $I$ ). But there are statisticians who maintain that we can only talk about the probability of  $X$  counts, assuming  $\lambda$ , and not of the probability distribution of  $\lambda$  having observed  $X = x$ , and even less of  $\lambda_1/\lambda_2$  (same as  $r_1/r_2$ , if  $T_1 = T_2$ ) having observed  $x_1$  and  $x_2$  counts.<sup>8</sup>

---

<sup>7</sup>The core of the R code is given, for the case of  $\lambda_1 = \lambda_2 = 1$ , by

```
lambda1 = lambda2 = 1; n = 10^6
x1 = rpois(n,lambda1)
x2 = rpois(n,lambda2)
rx = x1/x2
rx = rx[!is.nan(rx) & (rx != Inf)]
barplot(table(rx)/n, col='cyan', xlab='x1/x2', ylab='f(x1/x2)')
```

<sup>8</sup>Ref. [3] is a kind of 'masterpiece' of the kind of convoluted reasoning involved. For example, the paper starts with the following *incipit* (quote marks original):

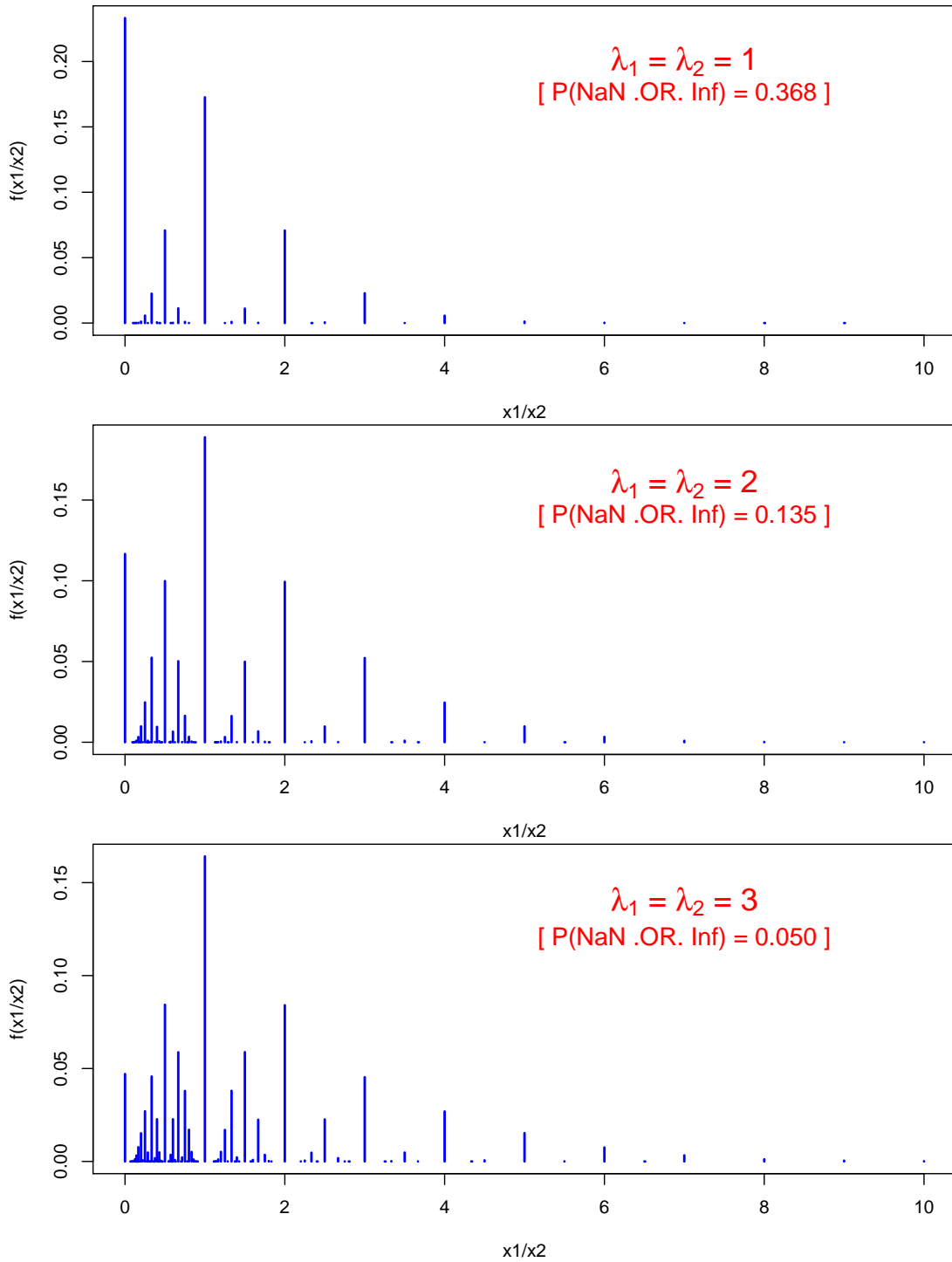


Figure 2: Monte Carlo distribution of the ratio of counts resulting from two Poisson distributions with  $\lambda_1 = \lambda_2$ .



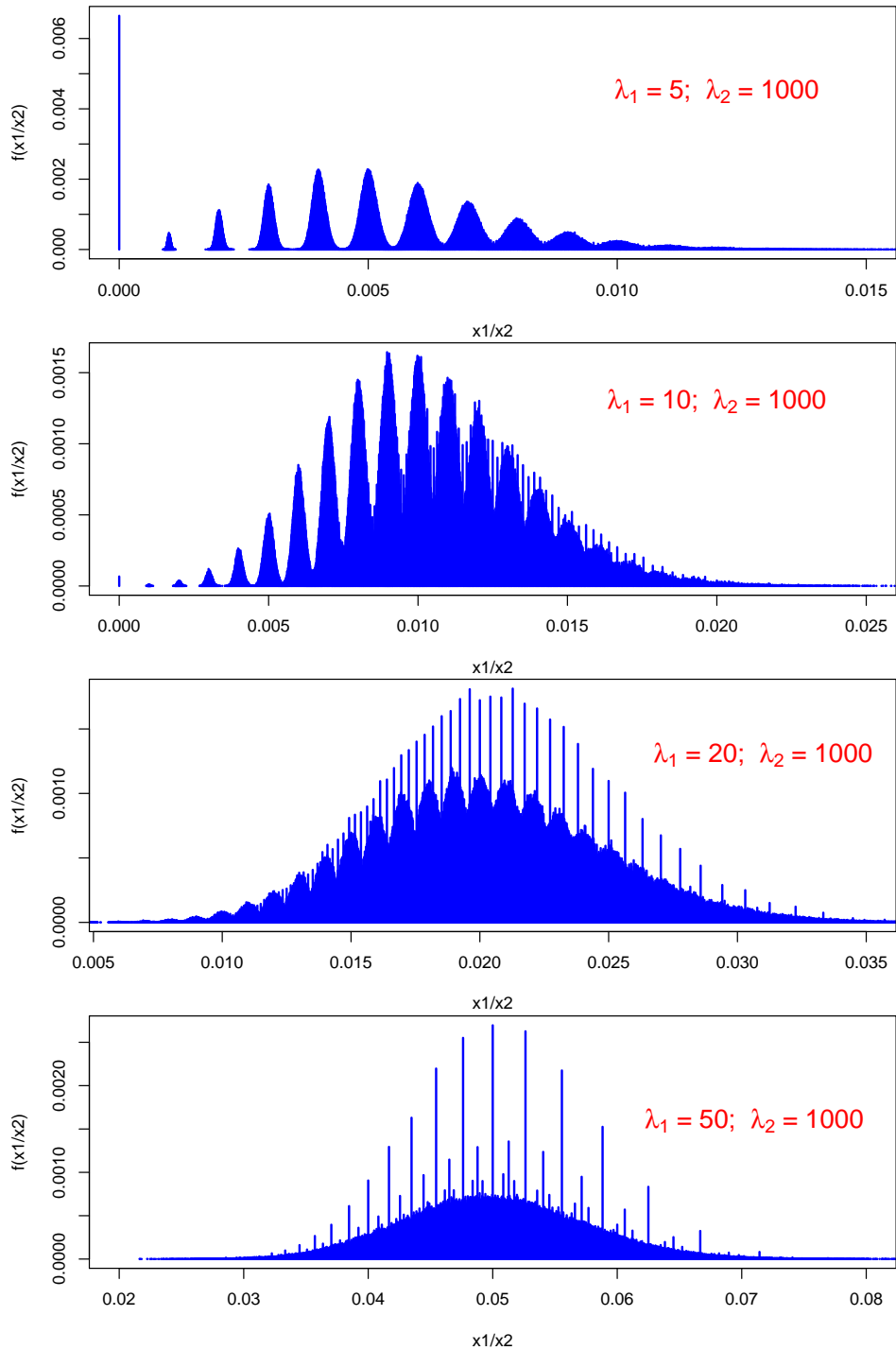


Figure 3: Monte Carlo distribution of the ratio of counts resulting from two Poisson distributions with  $\lambda_1 = 5, 10, 20, 50$  and  $\lambda_2 = 1000$ .

If, instead, we follow an approach closer to that innate in the human mind, which naturally attaches the concept of probability to whatever is considered uncertain [16], there is no need to go through contorted reasoning. For example, if we observe  $X = 1$ , then we tend to believe that this *effect* has been *caused* more likely by a value of  $\lambda$  around 1 rather than around 10 or 20, or larger, although we cannot rule out with *certainty* such values. Similarly, sticking to the observation of  $X = 1$ , we tend to believe to  $\lambda \approx 1$  much more than to  $\lambda \approx 10^{-2}$ , or smaller. In particular,  $\lambda = 0$  is *definitely ruled out*, because it could only yield 0 counts (this is just a limit case, since the Poisson  $\lambda$  is defined positive).

It is then clear that, as far as the ratio of  $\lambda_1/\lambda_2$  is concerned, there are no divergences, no matter how small the numbers of counts might be, obviously with two exceptions. The first is when  $X_2$  is observed to be exactly 0 (but in this case we could turn our interest in  $\lambda_2/\lambda_1$ , assuming  $X_1 > 0$ ). The second is when  $X_2$  is not zero, but there could be some background, such that  $r_2 = 0$  is not excluded with certainty [13, 17]. The effect of possible background is not going to be treated in detail in this paper, and only some hints on how to include it into the model will be given.

### 3 Inferring Poisson $\lambda$ 's and then *deducing* their ratio

We are now faced to the inference of  $r_1$  and  $r_2$  from the observed numbers of counts and the observation times  $T_1$  and  $T_2$ . For simplicity we start assuming  $T_1 = T_2$ , so that we can focus on  $\lambda_1$ ,  $\lambda_2$  and their ratio. The extension to the general case will be straightforward, as we shall see from Sec. 4 on.

---

*When the result of the measurement of a physical quantity is published as  $R = R_0 \pm \sigma_0$  without further explanation, it is implied that  $R$  is a gaussian-distributed measurement with mean  $R_0$  and variance  $\sigma_0^2$ . This allows one to calculate various confidence intervals of given “probability”, i.e., the “probability”  $P$  that the true value of  $R$  is within a given interval.*

However, nowhere in the paper is explained why *probability* is within quote marks. The reason is simply because the authors are fully aware that frequentist ideology, to which they overtly adhere, refuses to attach the concept of probability to *true values*, as well as to *model parameters*, and so on (see e.g. Ref. [13]). But authoritative statements of this kind might contribute to increase the confusion of practitioners [14], who then tend to take frequentist ‘confidence levels’ as if they were probability values [15].

### 3.1 Inference of $\lambda$ given $x$ , assuming $X \sim \mathcal{P}_\lambda$

The probability density function of  $\lambda$  is evaluated from the so called *Bayes' rule*:

$$f(\lambda|x) \propto f(x|\lambda) \cdot f_0(\lambda) \quad (3)$$

$$\propto \frac{\lambda^x \cdot e^{-\lambda}}{x!} \cdot f_0(\lambda), \quad (4)$$

where  $f_0(\lambda)$  is the so called ‘prior’.<sup>9</sup> Assuming for the moment a ‘flat’ prior, that is  $f_0(\lambda) = k$ , and neglecting all factors non depending on  $\lambda$ , we get

$$f(\lambda|x) \propto \lambda^x \cdot e^{-\lambda} \quad (5)$$

$$\propto \lambda^{(x+1)-1} \cdot e^{-\lambda}, \quad (6)$$

in which we recognize a Gamma pdf with  $\alpha = x + 1$  and  $\beta = 1$  (see Appendix A – for a detailed derivation see e.g. Ref. [13]), and therefore

$$f(\lambda|x) = \frac{1}{\Gamma(x+1)} \cdot \lambda^{(x+1)-1} \cdot e^{-\lambda} \quad (7)$$

$$= \frac{\lambda^x \cdot e^{-\lambda}}{x!}. \quad (8)$$

Expected value, standard deviation and mode are  $x + 1$ ,  $\sqrt{x + 1}$  and  $x$ , respectively. The advantage of having expressed the distribution of  $\lambda$  in terms of a Gamma is that we can use the probability distributions made available from programming languages, e.g. in R, which usually include also useful random generators (e.g. `rgamma()` in R). For example, making use of the R function `dgamma()` we can draw Fig. 4, which shows  $f(\lambda|x)$ , for  $x = 1, 2, \dots, 10$ , with the following few lines of code:

```
for (x.o in 0:10) {  
  curve(dgamma(x,x.o+1,1),xlim=c(0,20),ylim=c(0,1),col='blue',add=x.o>0,  
        xlab=expression(lambda),ylab=expression(paste('f(',lambda,')')))  
}
```

### 3.2 Distribution of the ratio of Poisson $\lambda$ 's by sampling

Once we have learned that the pdf of  $\lambda$ , in the light of the observation of  $x$  count and assuming a flat prior, is a Gamma distribution, the easiest way to evaluate the distribution of  $\lambda_1/\lambda_2$ , for  $x_2 > 0$ , is by sampling. For example, using the following lines of R code,

---

<sup>9</sup>This name is somehow unfortunate, because it might induce people think to time order, as discussed in Ref. [1], in which it is shown how, instead, the ‘prior’ can be applied in a second step, in particular by someone else, if a ‘flat prior’ used.

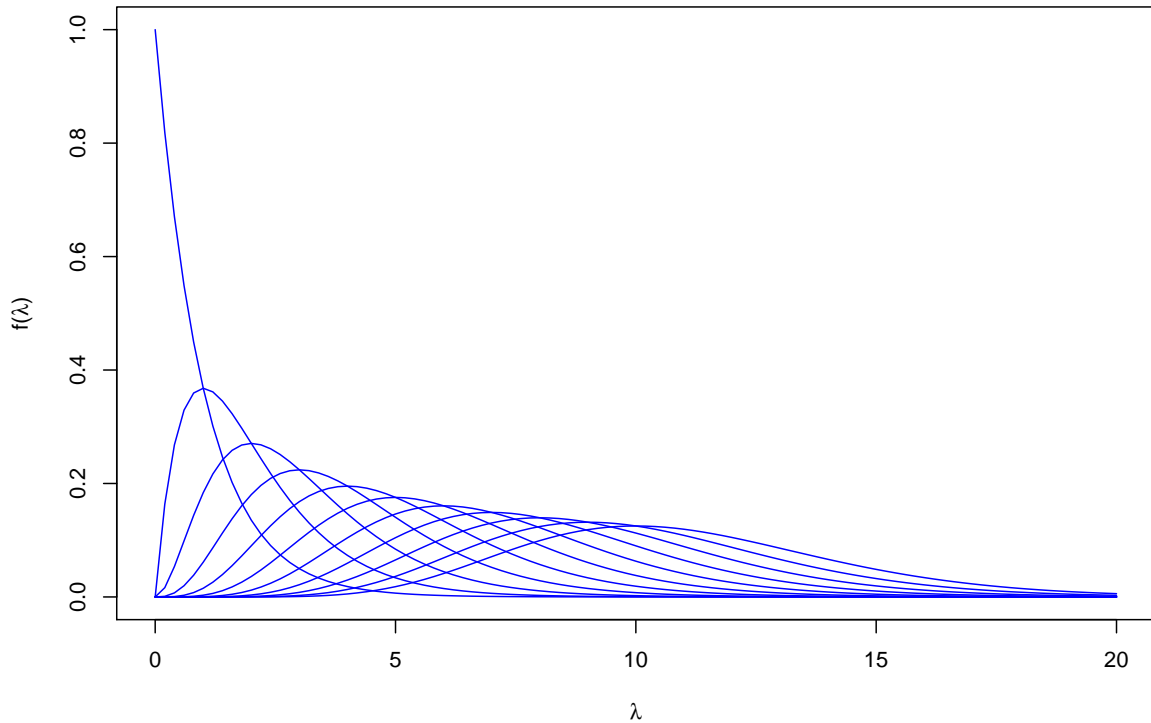


Figure 4: Inferred  $f(\lambda|x)$ , using a flat prior, for  $x = 0, 1, \dots, 10$ .

```

x1 = 1
x2 = 1
lambda1 = rgamma(n, x1+1, 1)
lambda2 = rgamma(n, x2+1, 1)
rho = lambda1/lambda2

```

Then, varying `x1` and `x2` we can get the plots of Figs. 5 and 6.<sup>10</sup> We immediately observe that the histograms are very regular, without divergences, although for small values of  $x_2$  there is quite a long tail up to infinity, which is however reached with vanishing probability (the figures report also the proportions of overflows, having chosen the horizontal scale of the plots in order to show the most interesting part of each probability distribution). The mean and standard deviation ('std') shown on each plot are calculated from the Monte Carlo samples.

The effect of the long tails is that there is quite a big difference between mean value and the most probable one, located around the highest bar of the histogram. This is not a surprise (the famous exponential distribution has modal value equal to zero independently of its parameter!), but it should sound as a *warning for those who*

---

<sup>10</sup>The complete script is provided in Appendix B.2.

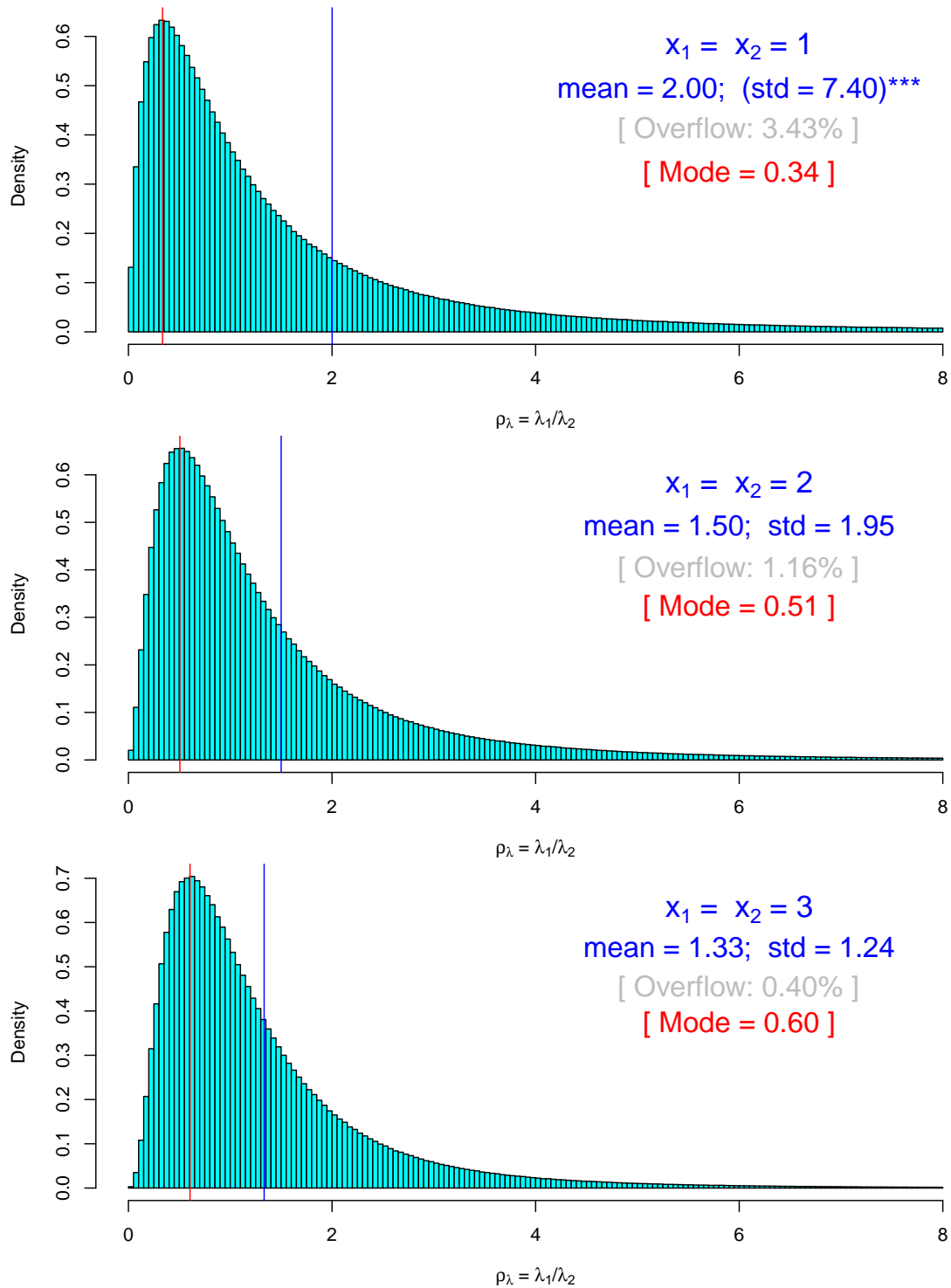


Figure 5: Estimate by sampling ( $n = 10^7$ ) of  $f(\rho_\lambda = \lambda_1/\lambda_2)$  for some 'observed' counts. (For the (non) meaning of standard deviation for  $x_1 = x_2 = 1$  see Sec. 3.3.)

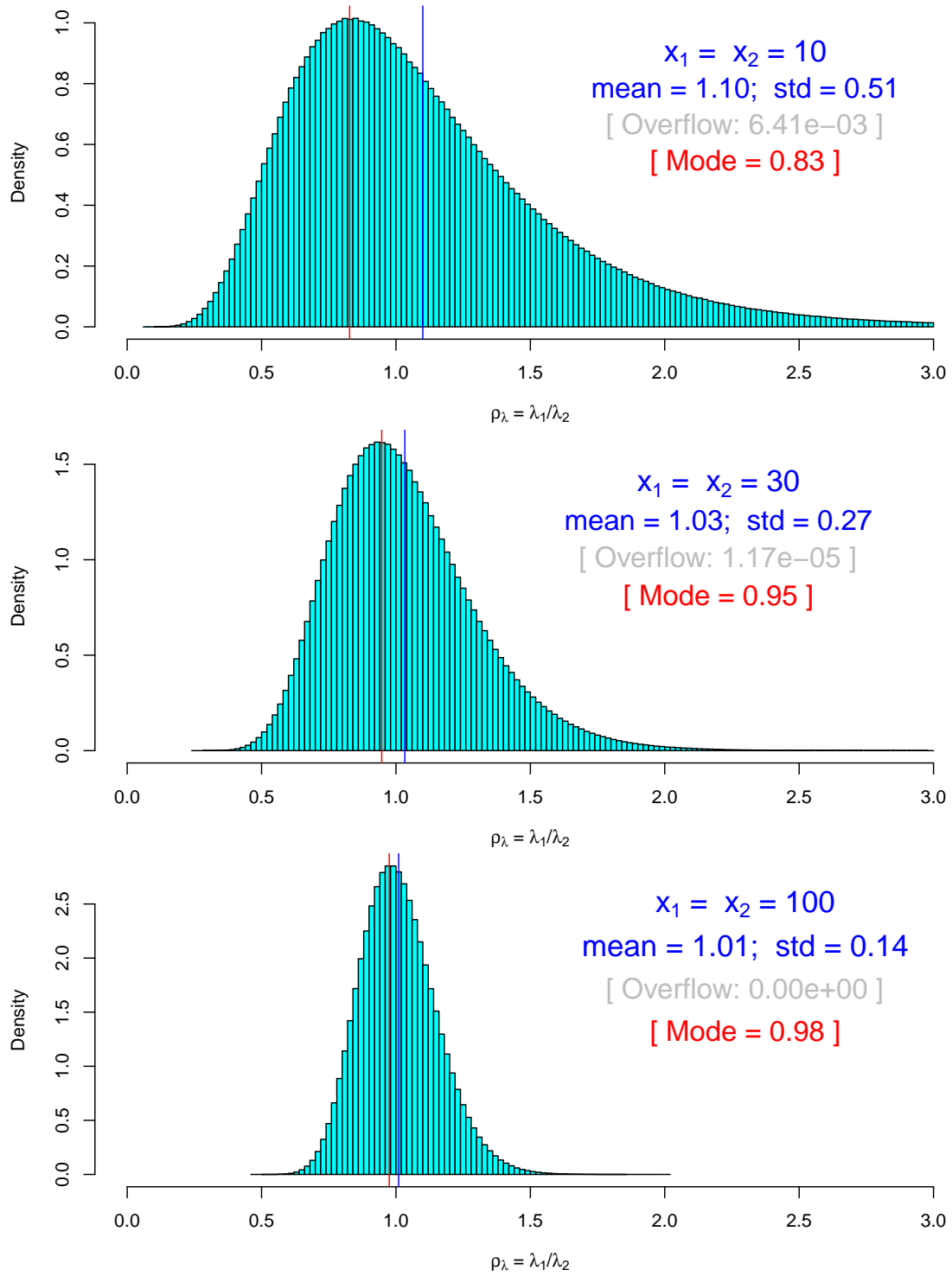


Figure 6: As Fig. 5 for larger values of the 'observed' counts.

use analysis methods which provide, as ‘estimator’, “the most probable value”<sup>11</sup> [13]. Moreover, for very small  $x_2$  the tails do not seem to go very fast to zero (in comparison e.g. to the exponential), leading to not-defined moments of the distribution (of the theoretical one, obviously, since in most cases mean and standard deviation of the Monte Carlo distribution have finite values). This question will be investigated in the next subsection, after the derivation of closed expressions.

When  $x_1$  and  $x_2$  become ‘quite large’ the Gamma distribution tends (*slowly* – think to the  $\chi^2$ , that is indeed a particular Gamma, as reminded in Appendix A) to a Gaussian, and likewise does (*a bit slower*) the ratio of two Gamma variables (as  $\lambda_1$  and  $\lambda_2$  are), as we can see for the case  $x_1 = x_2 = 100$  of Fig. 6,<sup>12</sup> in which some skewness is still visible and the mode is about 3% smaller than the mean value.

### 3.3 Distribution of the ratio of Poisson $\lambda$ ’s in closed form

Being the evaluation of ratio of rates (to which the ratio of  $\lambda$ ’s is related) an important issue in Physics, it is worth trying to get an analytic expression for its pdf. This can be done extending to the continuum Eq. (2),<sup>13</sup> that is replacing the sums by integrals, and applying the constraint between the two variables by a Dirac delta [13]:

$$f(\rho_\lambda | x_1, x_2) = \int_0^\infty \int_0^\infty \delta\left(\rho_\lambda - \frac{\lambda_1}{\lambda_2}\right) \cdot f(\lambda_1 | x_1) \cdot f(\lambda_2 | x_2) d\lambda_1 d\lambda_2. \quad (9)$$

Making use of the properties of the  $\delta()$ , we can rewrite it as

$$\delta\left(\rho_\lambda - \frac{\lambda_1}{\lambda_2}\right) = \frac{\delta(\lambda_1 - \lambda_1^*)}{\left|\frac{d}{d\lambda_1}\left(\rho_\lambda - \frac{\lambda_1}{\lambda_2}\right)\right|_{\lambda_1=\lambda_1^*}} \quad (10)$$

$$= \lambda_2 \cdot \delta(\lambda_1 - \lambda_1^*), \quad (11)$$

with  $\lambda_1^*$  root of the equation  $\rho_\lambda - \lambda_1/\lambda_2 = 0$ , and therefore equal to  $\rho_\lambda \cdot \lambda_2$ . Equation (9) becomes then

$$f(\rho_\lambda | x_1, x_2) = \int_0^\infty \int_0^\infty \lambda_2 \cdot \delta(\lambda_1 - \rho_\lambda \cdot \lambda_2) \cdot f(\lambda_1 | x_1) \cdot f(\lambda_2 | x_2) d\lambda_1 d\lambda_2 \quad (12)$$

$$= \int_0^\infty \lambda_2 \cdot \frac{(\rho_\lambda \cdot \lambda_2)^{x_1} \cdot e^{-\rho_\lambda \cdot \lambda_2}}{x_1!} \cdot \frac{\lambda_2^{x_2} \cdot e^{-\lambda_2}}{x_2!} d\lambda_2 \quad (13)$$

$$= \frac{\rho_\lambda^{x_1}}{x_1! x_2!} \cdot \int_0^\infty \lambda_2^{x_1+x_2+1} \cdot e^{-(\rho_\lambda+1) \cdot \lambda_2} d\lambda_2. \quad (14)$$

<sup>11</sup>For the reason of the quote marks see footnote 8.

<sup>12</sup>Zero overflow in that plot is only due to the ‘limited’ number of sampled, chosen to be  $10^7$ , and to the fact that the same script of the other plot has been used.

<sup>13</sup>for a different approach to get Eq. (9) see footnote 9, in which the integrand of Eq. (9) is interpreted as the joint pdf of  $\rho_\lambda$ ,  $\lambda_1$  and  $\lambda_2$ .

Once more we recognize in the integrand something related to the Gamma distribution. In fact, identifying the power of  $\lambda_2$  with ‘ $\alpha - 1$ ’ of a Gamma pdf, and ‘ $(1 + \rho_\lambda)$ ’ at the exponent with the ‘rate parameter’  $\beta$ , that is

$$\alpha - 1 = x_1 + x_2 + 1 \quad (15)$$

$$\beta = \rho_\lambda + 1, \quad (16)$$

the integrand in Eq. (14) can be rewritten as

$$\lambda_2^{\alpha-1} \cdot e^{-\beta \lambda_2} = \frac{\Gamma(\alpha)}{\beta^\alpha} \cdot \left( \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \lambda_2^{\alpha-1} \cdot e^{-\beta \lambda_2} \right) \quad (17)$$

in order to recognize within parentheses a Gamma pdf in the variable  $\lambda_2$ , whose integral over  $\lambda_2$  is then equal to one because of normalization. We get then

$$f(\rho_\lambda | x_1, x_2) = \frac{\rho_\lambda^{x_1}}{x_1! x_2!} \cdot \frac{\Gamma(\alpha)}{\beta^\alpha} \cdot \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \lambda_2^{\alpha-1} \cdot e^{-\beta \lambda_2} d\lambda_2 \quad (18)$$

$$= \frac{\rho_\lambda^{x_1}}{x_1! x_2!} \cdot \frac{\Gamma(\alpha)}{\beta^\alpha} \quad (19)$$

$$= \frac{\Gamma(x_1 + x_2 + 2)}{\Gamma(x_1 + 1) \Gamma(x_2 + 1)} \cdot \frac{\rho_\lambda^{x_1}}{(\rho_\lambda + 1)^{x_1 + x_2 + 2}} \quad (20)$$

$$= \frac{(x_1 + x_2 + 1)!}{x_1! x_2!} \cdot \rho_\lambda^{x_1} \cdot (\rho_\lambda + 1)^{-(x_1 + x_2 + 2)} \quad (21)$$

The mode of the distribution can be easily obtained finding the maximum (of the log) of the pdf, thus getting

$$\text{mode}(\rho_\lambda) = \frac{x_1}{x_2 + 2}, \quad (22)$$

in agreement with what we have got in Figs. 5 and 6 by Monte Carlo (indeed, done there in a fast and rather rough way – see Appendix B.2).

In order to get expected value and standard deviation, we need to evaluate the relevant integrals<sup>14</sup>

- First we check that  $f(\rho_\lambda | x_1, x_2)$  is properly normalized. Indeed the integral  $\int_0^\infty f(\rho_\lambda | x_1, x_2) d\rho_\lambda$  is equal to unity for ‘all possible’  $x_1$  and  $x_2$ .<sup>15</sup>

---

<sup>14</sup>Work done on a Raspberry Pi3, thanks to Mathematica 12.0 generously offered by Wolfram Inc. to the Raspbian system.

<sup>15</sup>To be precise, the condition is  $x_1 > -1$  and  $x_2 > -1$ , but, given the role of the two variables in our context, it means for all possible counts (including  $x_1 = x_2 = 0$ , for which the pdf becomes  $1/(1 + \rho_\lambda)^2$ , having however infinite mean and variance.)



- The expected value is equal to

$$E(\rho_\lambda | x_1, x_2) = \frac{x_1 + 1}{x_2} \quad (\mathbf{x}_2 > \mathbf{0}), \quad (23)$$

in perfect agreement with what we can read from the Monte Carlo results of Figs. 5 and 6.

- The expected value of  $\rho_\lambda^2$  is given by

$$E(\rho_\lambda^2 | x_1, x_2) = \frac{(x_1 + 1) \cdot (x_1 + 2)}{x_2 \cdot (x_2 - 1)} \quad (\mathbf{x}_2 > \mathbf{1}), \quad (24)$$

from which we evaluate (subtracting to it the square of the expected value)

$$\text{Var}(\rho_\lambda | x_1, x_2) = \frac{x_1 + 1}{x_2} \cdot \left( \frac{x_1 + 2}{x_2 - 1} - \frac{x_1 + 1}{x_2} \right) \quad (\mathbf{x}_2 > \mathbf{1}), \quad (25)$$

from which the standard deviation follows, that we rewrite in a more compact form as

$$\sigma(\rho_\lambda) = \sqrt{\mu_{\rho_\lambda} \cdot \left( \frac{x_1 + 2}{x_2 - 1} - \mu_{\rho_\lambda} \right)} \quad (\mathbf{x}_2 > \mathbf{1}), \quad (26)$$

having indicated by  $\mu_{\rho_\lambda}$  the expected value of  $\rho_\lambda$ . For the values  $x_1$  and  $x_2$  used in Figs. 5 and 6, we get, **starting from  $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{2}$**  in increasing order, the following standard deviations: 1.936, 1.247, 0.507, 0.269 and 0.143, in agreement with the Monte Carlo results (or the other way around).

The detailed comparison between closed expression of the pdf and the Monte Carlo outcome is shown in Fig. 7 for the toughest case we have met, that is  $x_1 = x_2 = 1$ .

## 4 Inferring $r_1$ and $r_2$ ( $T_1$ possibly different from $T_2$ )

After having been playing with  $\lambda$ 's and their ratios, from which we started for simplicity, let us move now to the rates  $r_1$  and  $r_2$  of the two Poisson processes, i.e. to the case in which the observation times  $T_1$  and  $T_2$  might be different.

But, before doing that, let us spend a few words on the reason of the word 'deducing', appearing in the title of the previous section. Let us start framing what we have been doing in the past section in the graphical model of Fig. 8, known as a *Bayesian network* (the reason for the adjective will be clear in the sequel). The *solid* arrows from the *nodes*  $\lambda_i$  to the *nodes*  $X_i$  indicate that the *effect*  $X_i$  is *caused*

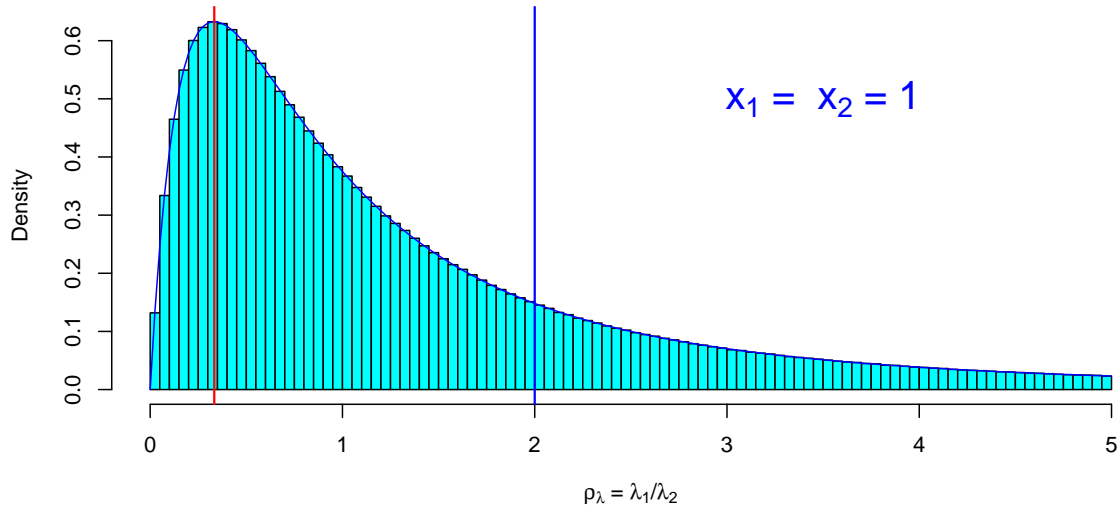


Figure 7: Comparison of the distribution of  $\rho_\lambda = \lambda_1/\lambda_2$  obtained by the closed expression (21) with that estimated by Monte Carlo (same as top plot of Fig. 5). The vertical lines indicate mode and expected value, evaluated using Eqs. (22) and (23), equal to  $1/3$  and  $2$ , respectively. (Note that none of these values is close to  $1$ , that is what one would naively expect for the ratio of the rates – indeed, only for  $x_1$  and  $x_2$  above  $\mathcal{O}(100)$  mode, expected value and ratio of the observed counts become approximately equal).

by  $\lambda_i$ , although in a probabilistic way (more properly,  $X_i$  is *conditioned* by  $\lambda_i$ , since, as it is well understood *causality is a tough concept*<sup>16</sup>). The dashed arrows indicate, instead, *deterministic* links (or deterministic ‘cause-effect’ relations, if you wish). For this reason we have been talking about ‘deduction’: each couple of values  $(\lambda_1, \lambda_2)$  provides a unique value of  $\rho_\lambda$ , equal to  $\lambda_1/\lambda_2$ , and any uncertainty about the  $\lambda$ ’s is ‘directly propagated’ into uncertainty about  $\rho_\lambda$ . The same will happen with  $\rho = r_1/r_2$ .

Navigating back along the solid arrows, that is from  $X_i$  to  $\lambda_i$ , is often called a problem of ‘inverse probability’, although nowadays many experts do not like this expression, which however gives an idea of what is going on. More precisely, it is an *inferential problem*, notoriously tackled for the first time in mathematical terms by Thomas Bayes [19] and Simon de Laplace, who was indeed talking about “*la probabilité des causes par les événements*” [20]. Nowadays the probabilistic tool to perform this ‘inversion’ goes under the name of *Bayes’ theorem*, or *Bayes’ rule*, whose essence, in terms of the possible causes  $C_i$  of the observed effect  $E$ , is given, besides

---

<sup>16</sup>For a historical review and modern developments, with implication on Artificial Intelligence application, see Ref. [18], an influential book on which I have however reservations when the author talks about causality in Physics (I have the suspicion he has never really read Newton or Laplace [20] or Poincaré [21], and perhaps not even Hume [16]).

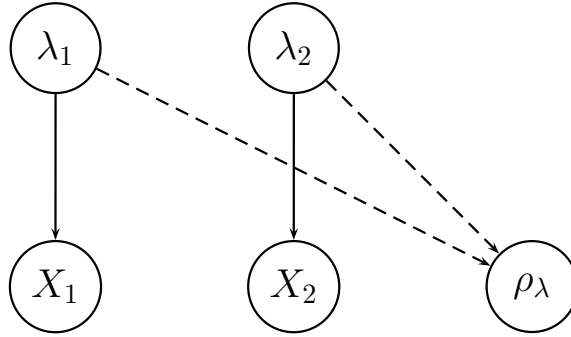


Figure 8: Graphical model showing the model underlying the inference of  $\lambda_1$  and  $\lambda_2$  from the observed numbers of counts  $X_1$  and  $X_2$ , followed by the deduction of  $\rho$ .

normalization, by this very simple formula

$$P(C_i | E, I) \propto P(E | C_i, I) \cdot P(C_i | I), \quad (27)$$

having indicated again by  $I$  the background state of information.  $P(C_i | I)$  quantifies our belief that  $C_i$  is true taking into account all available information, except for  $E$ . If also the *hypothesis that  $E$  is occurred* is considered to be true,<sup>17</sup> then the ‘prior’  $P(C_i | I)$  is turned into the ‘posterior’  $P(C_i | E, I)$ .

We shall come back on the question of the priors, but now let us move to infer  $r_1$ ,  $r_2$  and their ratio  $\rho = r_1/r_2$ .

#### 4.1 Inferring $r$ , having observed $x$ counts in the measuring time $T$

Being  $r$  equal to  $\lambda/T$ , we can obtain its pdf by a simple change of variables.<sup>18</sup> But, having practiced a bit with the Gamma distribution, we can reach the identical result observing that, using again a *flat prior* and neglecting irrelevant factors, the pdf of  $r$

<sup>17</sup>Usually we say ‘if  $E$  has occurred’, but, indeed, in probability theory there are, in general, ‘hypotheses’, to which we associate a degree of belief, being the states TRUE and FALSE just the limits, mapped into  $P = 0$  and  $P = 1$ .

<sup>18</sup>Starting from  $f(\lambda | x)$  given by Eq. (8) we get

$$\begin{aligned} f(\lambda | x) d\lambda &= \frac{\lambda^x \cdot e^{-\lambda}}{x!} d\lambda = \frac{(r \cdot T)^x \cdot e^{-rT}}{x!} \cdot T dr = f(r | x, T) dr \\ f(r | x, T) &= \frac{T^{x+1} \cdot r^x \cdot e^{-Tr}}{x!}. \end{aligned}$$

in which we recognize a Gamma pdf with  $\alpha = x + 1$  and  $\beta = T$ .

is given by

$$f(r|x, T) \propto f(x|r, T) \quad (28)$$

$$\propto (r \cdot T)^x \cdot e^{-rT} \quad (29)$$

$$\propto r^x \cdot e^{-Tr}, \quad (30)$$

in which we recognize, besides the normalization factor, a Gamma pdf for the variable  $r$  with  $\alpha = x + 1$  and  $\beta = T$ , and hence

$$f(r|x, T) = \frac{\beta^\alpha \cdot r^{\alpha-1} \cdot e^{-\beta r}}{\Gamma(\alpha)} \quad (31)$$

$$= \frac{T^{x+1} \cdot r^x \cdot e^{-Tr}}{x!}. \quad (32)$$

Mode, expected value and standard deviation of  $r$  are then (see Appendix A)

$$\begin{aligned} \text{mode}(r) &= \frac{\alpha - 1}{\beta} = \frac{x}{T} \\ \mathbf{E}(r) &= \frac{\alpha}{\beta} = \frac{x + 1}{T} \\ \sigma(r) &= \frac{\sqrt{\alpha}}{\beta} = \frac{\sqrt{x + 1}}{T}, \end{aligned}$$

as also expected from the ‘summaries’ of  $f(\lambda|x)$  and making use of  $r = \lambda/T$ .

[Note that the pdf (32) assumes, as explicitly written in the condition, a precise value of  $T$ . If this is not the case and  $T$  is uncertain, then, similarly to what we have seen in footnote 3, the pdf of  $r$  is evaluated as  $f(r|x, I) = \int_0^\infty f(r|x, T, I) \cdot f(T|I) dT$ .]

## 4.2 Role of the priors and sequential update of $f(r)$ as new observations are considered

The very essence of the so called probabilistic inference (‘Bayesian inference’) is given by Eq. (27). The rest is just a question of normalization and of extending it to the continuum, that in our case of interest is

$$f(r|x, T, I) \propto f(x|r, T, I) \cdot f(r|I). \quad (33)$$

It is evident the symmetric role of  $f(x|r, T, I)$  and  $f(r|I)$ , if the former is seen as a mathematical function of  $r$  for a given (‘observed’)  $x$ , that is  $x$  playing the role of a parameter. This function is known as *likelihood* and commonly indicated by

$\mathcal{L}(r; x, T)$ .<sup>19</sup> Indicating the second factor of Eq. (33), that is the ‘infamous’ prior that causes so much anxiety in practitioners [8], by  $f_0(r)$ , we get (assuming  $I$  implicit, as it is usually the case)

$$f(r | x, T, I) \propto \mathcal{L}(r; x, T) \cdot f_0(r), \quad (34)$$

which makes it clear that we have two mathematical functions of  $r$  playing **symmetric** and **peer** roles. Stated in different words, *each of the two has the role of ‘reshaping’ the other* [1]. In usual ‘routine’ measurements (as watching your weight on a balance) the information provided by  $\mathcal{L}(\dots)$  is so much narrower, with respect to  $f_0(\dots)$ ,<sup>20</sup> that we can neglect the latter and absorb it in the proportionality factor, as we have done above in Sec. 3.1. Employing a uniform ‘prior’ is then usually a good idea to start with, unless  $f_0(\dots)$  arises from previous measurements or from strong theoretical prejudice on the quantity of interest. It is also very important to understand that *‘the reshaping’ due to the priors can be done in a second step*, as it has been pointed out, with practical examples, in Ref. [1].

Let us now see what happens when, in our case, the Bayes rule is applied in sequence in order to account for several results on the same rate  $r$ , that is assumed to be stable. Imagine we start from rather vague ideas about the value of  $r$ , such that  $f_0(r) = k$  is, in practice, the best practical choice we can do. After the observation of  $x_1$  counts during  $T_1$  we get, as we have learned above,

$$f(r | x_1, T_1) = \frac{T_1^{x_1+1} \cdot r^{x_1} \cdot e^{-T_1 r}}{x_1!}. \quad (35)$$

Then we perform a *new* campaign of observations and record  $x_2$  counts in  $T_2$ . It is clear now that in the second inference we have to use as ‘prior’ the piece of knowledge derived from the first inference. So, we have, all together, besides irrelevant factors,

$$f(r | x_1, T_1, x_2, T_2) \propto f(x_2 | r, T_2) \cdot f(r | x_1, T_1) \quad (36)$$

$$\propto r^{x_2} \cdot e^{-T_2 r} \cdot r^{x_1} \cdot e^{-T_1 r} \quad (37)$$

$$\propto r^{x_1+x_2} e^{-(T_1+T_2)r}, \quad (38)$$

---

<sup>19</sup>The real issue with the ‘likelihood’ is not just replacing in Eq. (33)  $f(x | r, T, I)$  by  $\mathcal{L}(r; x, T)$ , but rather the fact, that, being this a function of  $r$ , it is perceived as ‘the likelihood of  $r$ ’. The result is that it is often (almost always) turned by practitioners into ‘probability of  $r$ ’, being ‘likelihood’ and ‘probability’ used practically as synonyms in the spoken language. It follows, for example, that the value that maximizes the likelihood function is perceived as the ‘most probable’ value, in the light of the observations.

<sup>20</sup>This is true unless the balance shows a ‘clear anomaly’, and then you stick to what you believed your weight should be. But you still learn something from the measurement, indeed: the balance is broken [13].

that is exactly as we had done a single experiment, observing  $x_{tot} = x_1 + x_2$  counts in  $T_{tot} = T_1 + T_2$ . The only real physical *strong* assumption is that the intensity of Poisson process was the same during the two measurements, i.e. *we have being measuring the same thing*.

This teaches us immediately how to ‘combine the results’, an often debated subject within experimental teams, if we have sets of counts  $x_i$  during times  $T_i$  (indicated all together by ‘ $\underline{x}$ ’ and ‘ $\underline{T}$ ’):

$$f(r | \underline{x}, \underline{T}) = \frac{(\sum_i T_i)^{\sum_i x_i + 1} \cdot r^{\sum_i x_i} \cdot e^{-(\sum_i T_i)r}}{(\sum_i x_i)!}, \quad (39)$$

without imaginative averages or fits. But this does not mean that we can blindly sum up counts and measuring times. Needless to say, it is important, whenever it is possible, to make a detailed study of the behavior of  $f(r | x_i T_i)$  in order to be sure that the intensity  $r$  is compatible with being constant during the measurements. But, once we are confident about its constancy (or that there is no strong evidence against that hypothesis), the result is provided by Eq. (39), from which all summaries of interest can be derived.<sup>21</sup>

### 4.3 Relative belief updating ratio

Let us consider again Eq. (34) and focus on the role of the likelihood to reshape  $f_0(r)$ . Being multiplicative factor irrelevant, it can be useful to rewrite that equation as

$$f(r | x, T, I) \propto \frac{\mathcal{L}(r; x, T)}{\mathcal{L}(r_R; x, T)} \cdot f_0(r) \quad (40)$$

with  $r_R$  a reference value, in principle arbitrary, but conceptually very interesting if properly chosen. In fact, the ratio in the above formula acquires the meaning of *relative belief update factor* [13, 17, 23],<sup>22</sup> and the updating Bayes’ rule can be rewritten as

$$f(r | x, T, I) \propto \mathcal{R}(r; x, T, r_R) \cdot f_0(r). \quad (41)$$

---

<sup>21</sup>It is perhaps important to remind that in probability theory the *full result* of the inference is the probability distribution (typically a pdf, for continuous quantities) of the quantity of interest as inferred from the data, the model and all other pertinent pieces of information. Mode, mean, median, standard deviation and probability intervals are just useful numbers to summarize with few numbers the distribution, with should always be reported, unless it is (with some degree of approximation) as simple as a Gaussian, so that mean and standard deviation provide the complete information. For example, the shape of a not trivial pdf can be expressed with coefficients of a suitable fit made in the region of interest. Or one can provide several moments of a distribution, from which the pdf can be reobtained (see e.g. Ref. [22]).

<sup>22</sup>Note how at that time we wrote Eq. (40) in a more expanded way, but the essence of this *factor* is given by Eq. (41). For recent developments and applications see Refs. [24, 25, 26].

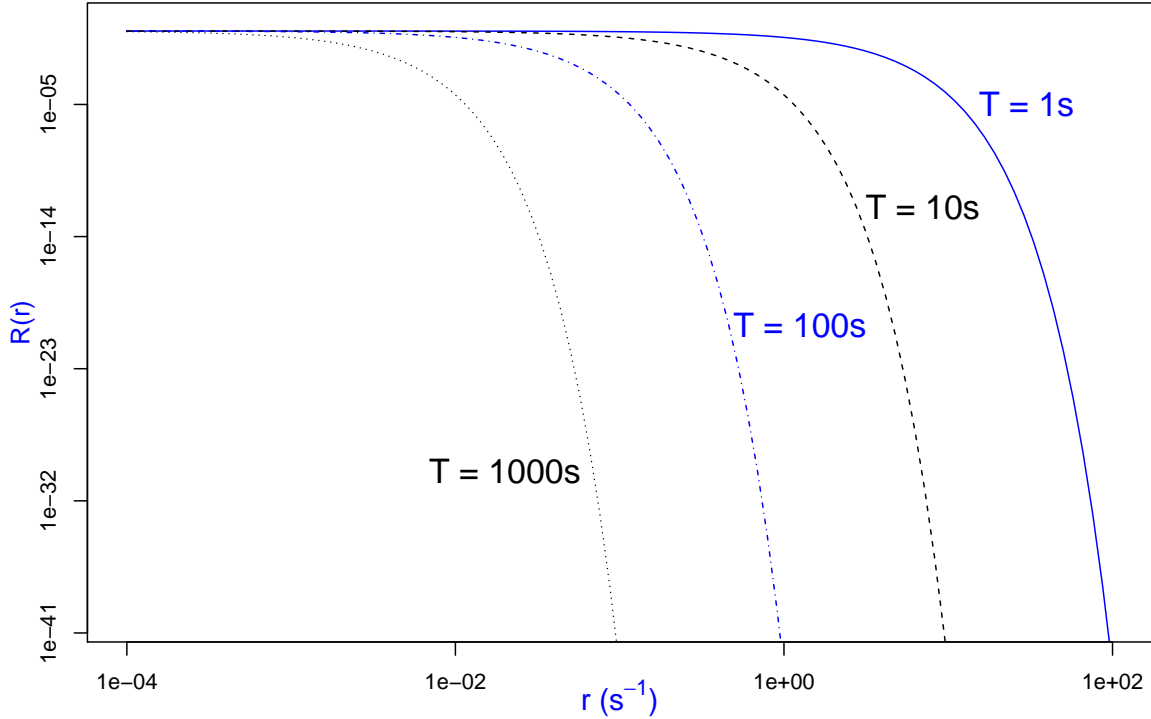


Figure 9: Relative belief updating factor  $\mathcal{R}(r; x = 0, T, r_0 = 0)$  for different observation times.

This way of rewriting the Bayes' rule is particularly convenient when the likelihood is not 'closed', that is it does not go to zero when the quantity of interest is 'very large' or 'very small'.

To be clear, let us make the example of *having observed zero counts*, that is the experiment was indeed performed, but no event of interest was found during the measurement time  $T$ . If we use a flat prior and only stick to the summaries, we have that the most probable value is zero, with  $E(r) = \sigma(r) = 1/T$ : the larger is the measuring time, the more the distribution of  $r$  is squeezed towards zero. But this does not give a complete picture of what is going on. Since  $\mathcal{L}(r; x = 0, T)$  goes to 1 for  $r \rightarrow 0$ , the likelihood is *opened in the left side*. Figure 9 shows  $\mathcal{R}$  functions for this case, for different  $T$ , although in this very simple case  $\mathcal{R}$  is mathematically equivalent to the likelihood.<sup>23</sup> If our beliefs about  $r$  were above  $O(100 \text{ s}^{-1})$ , the observation of zero events *practically* rule them out, even with  $T = 1 \text{ s}$  ('1 s' is arbitrarily chosen in this hypothetical example, just to remind that both  $T$  and  $r$  have physical dimensions).

If we run the experiment longer and longer, keeping observing zero events, the possible values of  $r$  gets smaller and smaller. What is mostly interesting, in this plot,

<sup>23</sup>The more interesting case, originally taken into account in Refs. [17, 23], is when some events are observed, which could be, however, also attributed to irreducible background.

is the region in which  $\mathcal{R}$  is flat: it means that if our beliefs are concentrate there, then the experiment does not teach us more than what we already believed: *the experiment looses sensitivity* in that region and then *reporting ‘probabilistic’ upper limits makes no sense* and it can be highly misleading (even more reporting ‘C.L. upper limits’) [13, 27].

## 4.4 Conjugate priors

At this point a technical remark is in order. The reason why the Gamma appears so often is that the expression of the Poisson probability function, seen as a function of  $\lambda$  and neglecting multiplicative factors, that is  $f(\lambda) \propto \lambda^x \cdot \exp(-\lambda)$ , has the same structure of a Gamma pdf. The same is true if the variable  $r$  is considered, that is  $f(r) \propto r^x \cdot \exp(-T \cdot r)$ . If then we have a Gamma distribution as prior, with parameters  $\alpha_0$  and  $\beta_0$ , the ‘final’ distributions is still a Gamma:

$$f(\lambda | x) = \lambda^x \cdot e^{-\lambda} \cdot \lambda^{\alpha_0 - 1} \cdot e^{-\beta_0 \lambda} = \lambda^{\alpha_0 + x - 1} \cdot e^{-(\beta_0 + 1) \cdot \lambda} \quad (42)$$

$$\propto \lambda^{\alpha_f - 1} \cdot e^{-\beta_f \cdot \lambda} \quad (43)$$

$$f(r | x, T) = r^x \cdot e^{-T \cdot r} \cdot r^{\alpha_0 - 1} \cdot e^{-\beta_0 r} = r^{\alpha_0 + x - 1} \cdot e^{-(\beta_0 + T) \cdot r} \quad (44)$$

$$\propto r^{\alpha_f - 1} \cdot e^{-\beta_f \cdot r} \quad (45)$$

This kind of distributions, such that the ‘posterior’ belongs to the same family of the ‘prior’, with *updated parameters*, are called *conjugate priors* for obvious reasons, as it is rather obvious how convenient they are in applications, *provided they are flexible enough to describe ‘somehow’ the prior belief*.<sup>24</sup> This was particularly important at the times when the monstrous computational power nowadays available was not even imaginable (also the development of logical and mathematical tools has a strong relevance). Therefore a quite rich collection of conjugate priors is available in the literature (see e.g. Ref. [30]).

In sum, these are the *updating rules* of the Gamma parameters for our cases of interest (the subscript ‘f’ is to remind that is the parameter of the ‘final’ distribution):

$$\textbf{Inferring } \lambda: \quad \alpha_f = \alpha_0 + x \quad (46)$$

$$\beta_f = \beta_0 + 1 \quad (47)$$

$$\textbf{Inferring } r: \quad \alpha_f = \alpha_0 + x \quad (48)$$

$$\beta_f = \beta_0 + T \quad (49)$$

---

<sup>24</sup>Remember that, as Laplace used to say, “*the theory of probabilities is basically just common sense reduced to calculus*”, that “*All models are wrong, but some are useful*” (G.Cox) and that even Gauss was ‘sorry’ because ‘his’ error function could not be strictly true [28] (see quote in footnote 9 of Ref. [29]).



(Note that in the case of  $r$  the parameter  $\beta$  has the dimension of a time, being  $r$  a rate, that is counts per unit of time.) A flat prior distribution is recovered for  $\alpha_0 = 1$  and  $\beta_0 \rightarrow 0$ . Technically, for  $\alpha = 1$  a Gamma distribution turns into a negative exponential: if then the ‘rate parameter’  $\beta$  is chosen to be very small, the exponential becomes ‘essentially flat’ in the region of interest.

Once we have learned the updating rules (46)-(47) and (48)-(49), it might be convenient to turn a prior expressed in terms of mean  $\mu_0$  and standard deviation  $\sigma_0$  into  $\alpha_0$  and  $\beta_0$ , inverting the expressions of expected value and standard deviation of a Gamma distributed variable (see Appendix A), thus getting

$$\alpha_0 = \mu_0^2 / \sigma_0^2 \tag{50}$$

$$\beta_0 = \mu_0 / \sigma_0^2. \tag{51}$$

For example, if we have good reason to think that  $r$  should be  $(5 \pm 2) \text{ s}^{-1}$ , the parameters of our initial Gamma distribution are  $\alpha_0 = 6.25$  and  $\beta_0 = 1.25 \text{ s}$ . This is equivalent to having started from a flat prior and having observed (rounding the numbers) 5 counts in about 1.2 seconds. This gives a clear idea of the ‘strength’ of the prior – not much in this case, but it certainly excludes the possibility of  $r = 0$ . This happens in fact as soon as  $\alpha_0$  is larger than 1, implying  $r^{\alpha_0-1}$  vanishing at  $r = 0$ . This observation can be used as a trick to forbid a vanishing value of  $\lambda$  or of  $r$ , if we have good physical reason to believe that they cannot be zero, although we are highly uncertain about even their order of magnitude: just choose a prior  $\alpha_0$  slightly larger than one.

## 5 Ratio of Gamma distributed variables

Having inferred the two rates, we can now evaluate the distribution of  $\rho = r_1/r_2$ , which is technically just a problem of ‘direct probabilities’, that is getting the pdf  $f(\rho | x_1, T_1, x_2, T_2)$  from  $f(r_1 | x_1, T_1)$  and  $f(r_2 | x_2, T_2)$  (the Bayesian network that relates the variables of interest is shown in Fig. 10). We just need to repeat what it has been done in Sec. 3.3, taking the advantage of having understood that  $f(\lambda_1 | x_1)$  and  $f(\lambda_2 | x_2)$  appearing in Eq. (9) are indeed Gamma distributions. Therefore, we start evaluating the probability distribution of the ratio of generic Gamma variables, denoted as  $Z_1$  and  $Z_2$  (and their possible occurrences  $z_1$  and  $z_2$ ) in order to avoid confusion with  $X$ ’s, associated so far to measured counts:

$$Z_1 \sim \text{Gamma}(\alpha_1, \beta_1) \tag{52}$$

$$Z_2 \sim \text{Gamma}(\alpha_2, \beta_2). \tag{53}$$

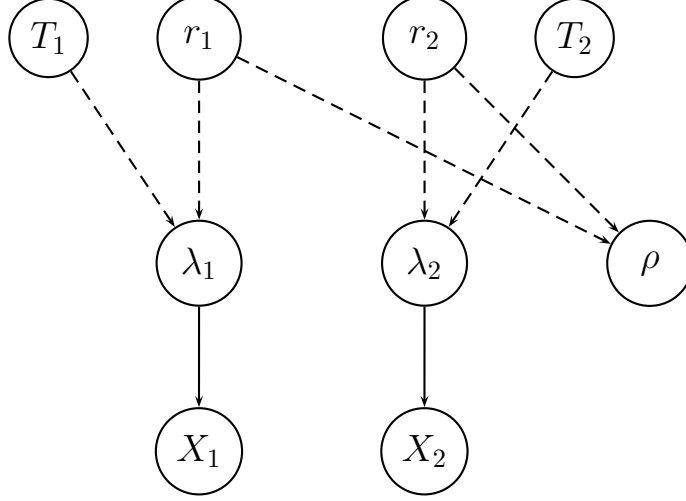


Figure 10: Graphical model relating the physical quantities (rates and measurement times) to the observed numbers of events.

The pdf of  $Z_1/Z_2$  is the given by

$$f(\rho_z | \alpha_1, \beta_1, \alpha_2, \beta_2) = \int_0^\infty \int_0^\infty \delta\left(\rho_z - \frac{z_1}{z_2}\right) \cdot f(z_1 | \alpha_1, \beta_1) \cdot f(z_2 | \alpha_2, \beta_2) dz_1 dz_2, \quad (54)$$

in which we have indicated by  $\rho_z$  their ratio. In detail, taking benefit of what we have learned in Sec. 3.3,

$$\begin{aligned} f(\rho_z | \dots) &= \int_0^\infty \int_0^\infty z_2 \cdot \delta(z_1 - \rho_z \cdot z_2) \cdot \frac{\beta_1^{\alpha_1} \cdot z_1^{\alpha_1-1} \cdot e^{-\beta_1 z_1}}{\Gamma(\alpha_1)} \cdot \frac{\beta_2^{\alpha_2} \cdot z_2^{\alpha_2-1} \cdot e^{-\beta_2 z_2}}{\Gamma(\alpha_2)} dz_1 dz_2 \\ &= \frac{\beta_1^{\alpha_1} \beta_2^{\alpha_2}}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \cdot \int_0^\infty z_2 \cdot (\rho_z \cdot z_2)^{\alpha_1-1} \cdot e^{-\beta_1 (\rho_z \cdot z_2)} \cdot z_2^{\alpha_2-1} \cdot e^{-\beta_2 z_2} dz_2 \quad (55) \end{aligned}$$

$$= \frac{\beta_1^{\alpha_1} \cdot \beta_2^{\alpha_2}}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \cdot \rho_z^{\alpha_1-1} \int_0^\infty z_2^{\alpha_1+\alpha_2-1} \cdot e^{-(\beta_2+\rho_z \beta_1) \cdot z_2} dz_2. \quad (56)$$

Writing  $\alpha_1 + \alpha_2$  as  $\alpha_*$  and  $\beta_2 + \rho_z \cdot \beta_1$  as  $\beta_*$ , we get

$$f(\rho_z | \alpha_1, \beta_1, \alpha_2, \beta_2) = \frac{\beta_1^{\alpha_1} \cdot \beta_2^{\alpha_2}}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \cdot \rho_z^{\alpha_1-1} \int_0^\infty z_2^{\alpha_*-1} \cdot e^{-\beta_* \cdot z_2} dz_2 \quad (57)$$

$$= \frac{\beta_1^{\alpha_1} \cdot \beta_2^{\alpha_2}}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \cdot \rho_z^{\alpha_1-1} \cdot \frac{\Gamma(\alpha_*)}{\beta_*^{\alpha_*}} \quad (58)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \cdot \beta_1^{\alpha_1} \cdot \beta_2^{\alpha_2} \cdot \frac{\rho_z^{\alpha_1-1}}{(\beta_2 + \rho_z \cdot \beta_1)^{\alpha_1+\alpha_2}} \quad (59)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2)} \cdot \beta_1^{\alpha_1} \cdot \beta_2^{\alpha_2} \cdot \rho_z^{\alpha_1-1} \cdot (\beta_2 + \rho_z \cdot \beta_1)^{-(\alpha_1+\alpha_2)}. \quad (60)$$

## 5.1 Application to $\rho_\lambda = \lambda_1/\lambda_2$ and to $\rho = r_1/r_2$

In the case of ratio of  $\lambda$ 's, and starting from uniform prior, as done in Sec. 3.3, we get, applying Eq. (60) and writing the conditions in terms of the Gamma parameters,

$$\begin{aligned} f(\rho_\lambda = \frac{\lambda_1}{\lambda_2} | x_1 + 1, 1, x_2 + 1, 1) &= \frac{\Gamma(x_1 + x_2 + 2)}{\Gamma(x_1 + 1) \cdot \Gamma(x_2 + 1)} \cdot \rho_\lambda^{x_1} \cdot (1 + \rho_\lambda)^{-(x_1 + x_2 + 2)} \\ &= \frac{(x_1 + x_2 + 1)!}{x_1! x_2!} \cdot \rho_\lambda^{x_1} \cdot (1 + \rho_\lambda)^{-(x_1 + x_2 + 2)}, \end{aligned} \quad (61)$$

re-obtaining exactly Eq. (21).

As far as the ratio of rates, starting again from a uniform prior, implying then  $\alpha_i = x_i + 1$  and  $\beta_i = T_i$ , we get, writing, as in Eq. (61), the conditions in terms of the Gamma parameters,

$$f(\rho = \frac{r_1}{r_2} | x_1 + 1, T_1, x_2 + 1, T_2) = \frac{(x_1 + x_2 + 1)!}{x_1! x_2!} \cdot T_1^{x_1 + 1} \cdot T_2^{x_2 + 1} \cdot \rho^{x_1} \cdot (T_2 + T_1 \cdot \rho)^{-(x_1 + x_2 + 2)},$$

that is, without redundant details,<sup>25</sup>

$$f(\rho | x_1, T_1, x_2, T_2) = \frac{(x_1 + x_2 + 1)!}{x_1! x_2!} \cdot T_1^{x_1 + 1} \cdot T_2^{x_2 + 1} \cdot \rho^{x_1} \cdot (T_2 + T_1 \cdot \rho)^{-(x_1 + x_2 + 2)}, \quad (62)$$

from which we re-obtain Eqs. (21) and (61) in the special case  $T_1 = T_2$ , as it has to be. Mode, expected value and standard deviation can be obtained quite easily from Eqs. (22)-(26), just noting that

$$\rho = \frac{r_1}{r_2} = \frac{\lambda_1/T_1}{\lambda_2/T_2} = \frac{\lambda_1}{\lambda_2} \cdot \frac{T_2}{T_1} = \frac{T_2}{T_1} \cdot \rho_\lambda,$$

---

<sup>25</sup>Another way to arrive to Eq. (62) is to start from Eq. (21), applying the transformation of variables  $\rho = \rho_\lambda \cdot T_2/T_1$ :

$$\begin{aligned} f(\rho_\lambda | \dots) d\rho_\lambda &= \frac{(x_1 + x_2 + 1)!}{x_1! x_2!} \cdot \rho_\lambda^{x_1} \cdot (1 + \rho_\lambda)^{-(x_1 + x_2 + 2)} d\rho_\lambda \\ &= \frac{(x_1 + x_2 + 1)!}{x_1! x_2!} \cdot \left(\frac{T_1}{T_2}\right)^{x_1} \cdot \rho^{x_1} \cdot \left(1 + \frac{T_1}{T_2} \cdot \rho\right)^{-(x_1 + x_2 + 2)} \cdot \frac{T_1}{T_2} d\rho \\ &= \frac{(x_1 + x_2 + 1)!}{x_1! x_2!} \cdot T_1^{x_1} \cdot T_2^{-x_1} \cdot T_2^{x_1 + x_2 + 2} \cdot \rho^{x_1} \cdot (T_2 + T_1 \cdot \rho)^{-(x_1 + x_2 + 2)} \cdot \frac{T_1}{T_2} d\rho \\ &= \frac{(x_1 + x_2 + 1)!}{x_1! x_2!} \cdot T_1^{x_1 + 1} \cdot T_2^{x_2 + 1} \cdot \rho^{x_1} \cdot (T_2 + T_1 \cdot \rho)^{-(x_1 + x_2 + 2)} d\rho \\ f(\rho | \dots) &= \frac{(x_1 + x_2 + 1)!}{x_1! x_2!} \cdot T_1^{x_1 + 1} \cdot T_2^{x_2 + 1} \cdot \rho^{x_1} \cdot (T_2 + T_1 \cdot \rho)^{-(x_1 + x_2 + 2)}. \end{aligned}$$

and then

$$\text{mode}(\rho) = \frac{T_2}{T_1} \cdot \text{mode}(\rho_\lambda) = \frac{T_2}{T_1} \cdot \frac{x_1}{x_2 + 2} = \frac{x_1/T_1}{(x_2 + 2)/T_2} \quad (63)$$

$$\text{E}(\rho) = \frac{T_2}{T_1} \cdot \text{E}(\rho_\lambda) = \frac{T_2}{T_1} \cdot \frac{x_1 + 1}{x_2} = \frac{(x_1 + 1)/T_1}{x_2/T_2} \quad (\mathbf{x}_2 > \mathbf{0}) \quad (64)$$

$$\sigma(\rho) = \frac{T_2}{T_1} \cdot \sigma(\rho_\lambda) = \frac{T_2}{T_1} \cdot \sqrt{\frac{x_1 + 1}{x_2} \cdot \left( \frac{x_1 + 2}{x_2 - 1} - \frac{x_1 + 1}{x_2} \right)} \quad (\mathbf{x}_2 > \mathbf{1}) \quad (65)$$

that we can rewrite in a more compact form, in terms of  $\mu_\rho \equiv \text{E}(\rho)$ , as

$$\sigma(\rho) = \sqrt{\mu_\rho \cdot \left( \frac{T_2}{T_1} \cdot \frac{x_1 + 2}{x_2 - 1} - \mu_\rho \right)}, \quad (\mathbf{x}_2 > \mathbf{1}) \quad (66)$$

Some examples are provided in Figs. 11 and 12, for low and relatively high numbers of counts, respectively. Each plot shows both the curve of the pdf, calculated with the closed formulae just derived, and the histogram of Monte Carlo simulation (the script to reproduce these plots is given in Appendix B.3). The value of mode, expected value and standard deviation calculated from exact formulae are reported too, together with ‘mean’ and ‘std’ (‘empirical standard deviation’) evaluated from the sampling. The excellent agreement can be considered a cross check of the exact formulae, derived above for the purpose.

The counts and the measuring times have been chosen such that  $(x_1/T_1)/(x_2/T_2)$  are equal to one in all cases. Therefore the plots are comparable to those of Figs. 5 and 6 reporting  $\rho_\lambda$  for several values of  $\lambda_1 = \lambda_2$  (but in that case all summaries were evaluated from sampling, having, at that stage of the work, not yet derived the closed formulae of interest). As we can again see, for small numbers of counts the distribution of the ratio of rates is strongly asymmetric, with mode and expected value systematically below and above, respectively, the ratio calculated naively as  $(x_1/T_1)/(x_2/T_2)$ . This value is reached asymptotically, as we see in Fig. 12, and as expected by the fact that for high numbers of counts we get

$$\text{mode}(\rho) = \frac{x_1/T_1}{(x_2 + 2)/T_2} \xrightarrow{x_2 \gg 2} \frac{x_1/T_1}{x_2/T_2} \quad (67)$$

$$\text{E}(\rho) = \frac{(x_1 + 1)/T_1}{x_2/T_2} \xrightarrow{x_1 \gg 1} \frac{x_1/T_1}{x_2/T_2} \quad (68)$$

$$\sigma(\rho) = \frac{T_2}{T_1} \cdot \sqrt{\frac{x_1 + 1}{x_2} \cdot \left( \frac{x_1 + 2}{x_2 - 1} - \frac{x_1 + 1}{x_2} \right)} \xrightarrow{x_1, x_2 \rightarrow \infty} 0. \quad (69)$$

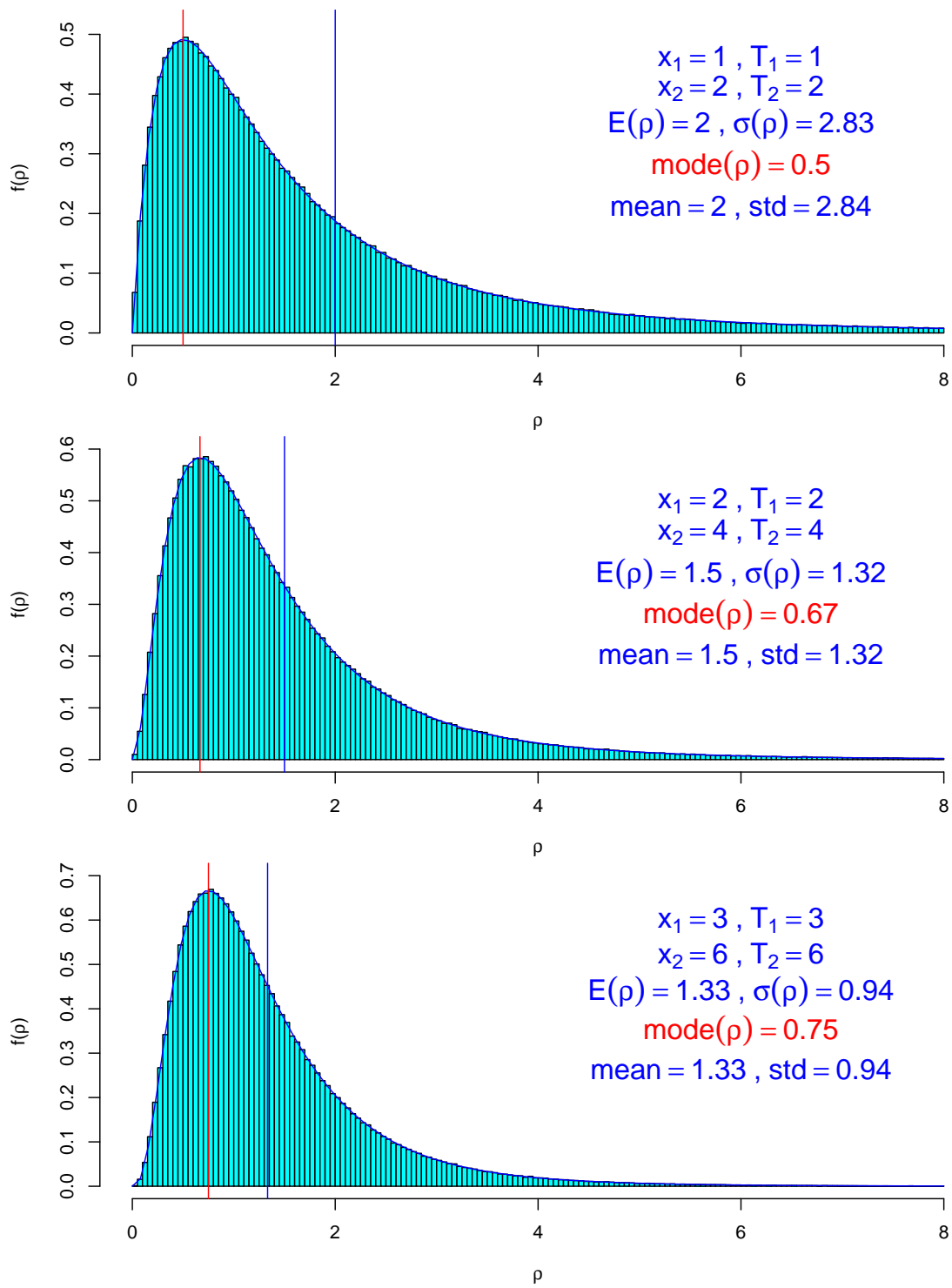


Figure 11: Ratios of rates, given counts and times.

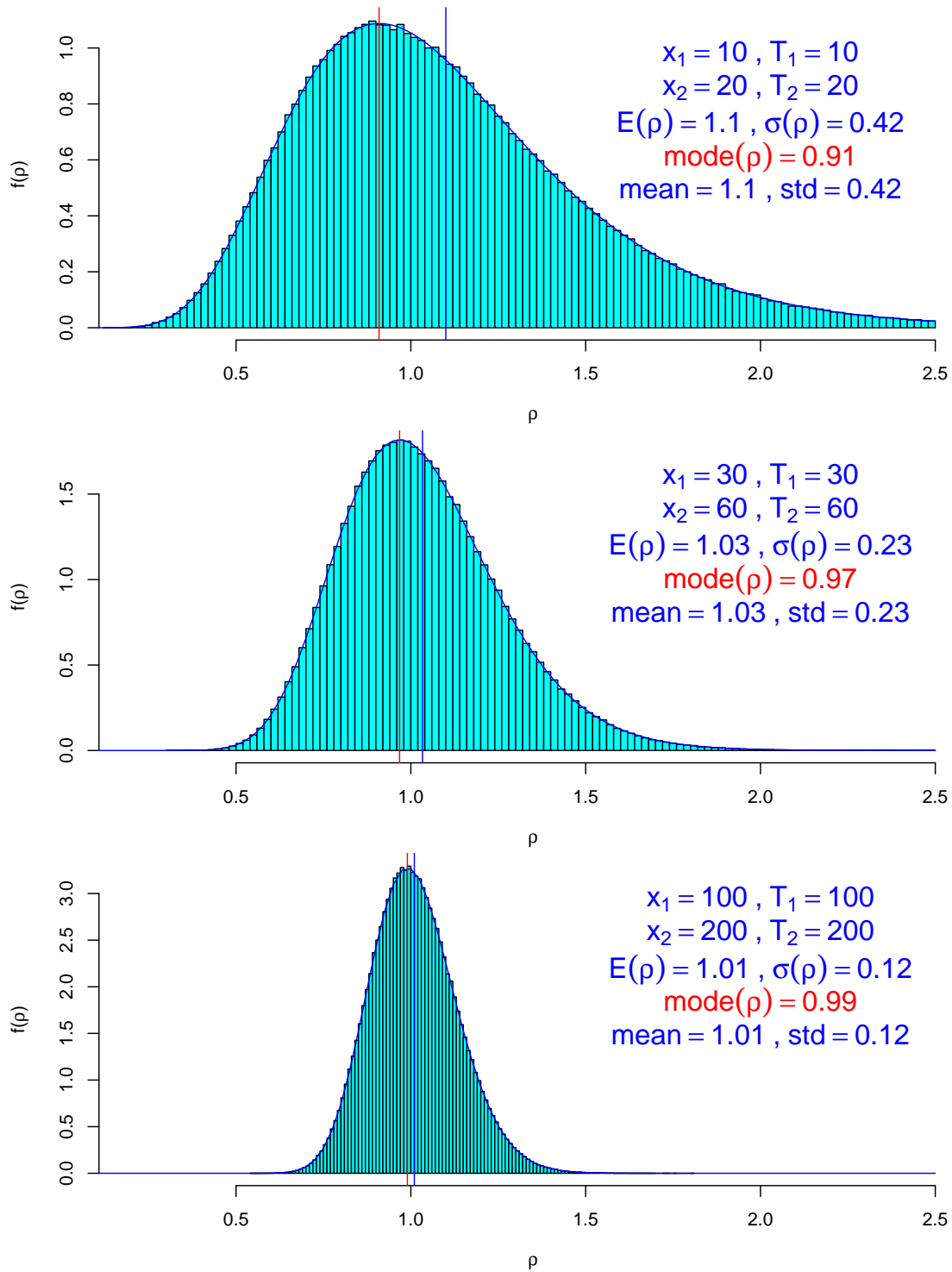


Figure 12: As Fig. 11 for larger values of counts, observed in proportionally larger times.

## 5.2 More on the ratio of Gamma distributed variables

Keeping the notation  $\rho_z$  for the ratio of the generic Gamma distributed variables  $Z_1$  and  $Z_2$ , the pdf of Eq. (60) can be further simplified reminding that the *beta* special function (or *Euler integral of the first kind* [33]), defined as

$$B(r, s) = \int_0^1 t^{r-1} \cdot (1-t)^{s-1} dt, \quad (70)$$

can be written as

$$B(r, s) = \frac{\Gamma(r) \cdot \Gamma(s)}{\Gamma(r+s)}. \quad (71)$$

We can then rewrite the combination of three gamma functions appearing in Eq. (60) as  $1/B(\alpha_1, \alpha_2)$ , thus getting

$$f(\rho_z | \alpha_1, \beta_1, \alpha_2, \beta_2) = \frac{1}{B(\alpha_1, \alpha_2)} \cdot \beta_1^{\alpha_1} \cdot \beta_2^{\alpha_2} \cdot \rho_z^{\alpha_1-1} \cdot (\beta_2 + \rho_z \cdot \beta_1)^{-(\alpha_1+\alpha_2)}. \quad (72)$$

As far as mode, expected value and variance are concerned, they can be obtained, without direct calculations, just transforming those of  $\rho = r_1/r_2$ , seen above, remembering that, starting from a flat prior,  $r_i \sim \text{Gamma}(\alpha_i = x_i + 1, \beta_i = T_i)$ . We get then

$$\text{mode}(\rho_z) = \frac{\beta_2}{\beta_1} \cdot \frac{\alpha_1 - 1}{\alpha_2 + 1} \quad (73)$$

$$E(\rho_z) = \frac{\beta_2}{\beta_1} \cdot \frac{\alpha_1}{\alpha_2 - 1} \quad (\alpha_2 > 1) \quad (74)$$

$$\text{Var}(\rho_z) = \frac{\beta_2^2}{\beta_1^2} \cdot \left[ \frac{\alpha_1}{\alpha_2 - 1} \cdot \left( \frac{\alpha_1 + 1}{\alpha_2 - 2} - \frac{\alpha_1}{\alpha_2 - 1} \right) \right] \quad (\alpha_2 > 2). \quad (75)$$

Moreover, just for completeness, let us mention the special case  $\beta_1 = \beta_2 = 1$ , that written for the generic variable  $X$ , becomes

$$f(x | \alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} \cdot \rho_z^{\alpha_1-1} \cdot (1 + \rho_z)^{-(\alpha_1+\alpha_2)}, \quad (76)$$

‘known’ (certainly not to me before I was attempting to write these subsection) as *Beta prime distribution* [34], with parameters  $\alpha$  and  $\beta$ :

$$f(x | \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \cdot x^{\alpha-1} \cdot (1+x)^{-(\alpha+\beta)}. \quad (77)$$

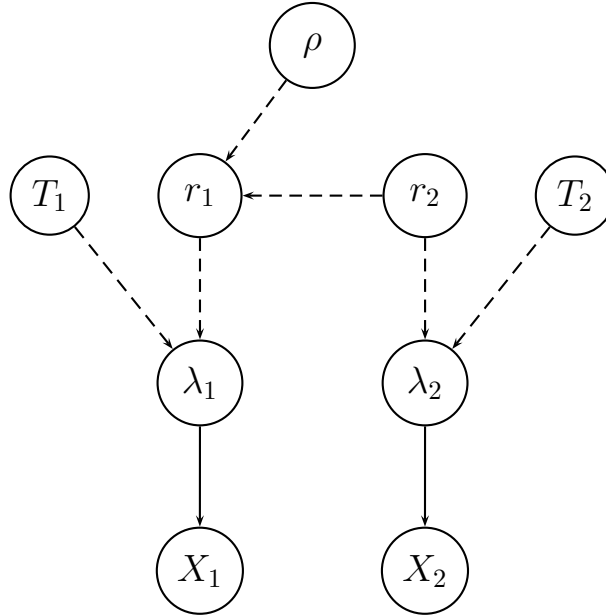


Figure 13: Alternative graphical model to that of Fig. 10.

The name of the distribution is clearly due to the special function resulting from normalization. It is ‘prime’ in order to distinguish it from the more famous (and more important as far as practical applications are concerned) *Beta distribution* which arises quite ‘naturally’ when inferring the parameter  $p$  of the Bernoulli trials, in the light of  $x$  successes in  $n$  trials (essentially the original problem tackled by Bayes [19] and Laplace [20]), and then used as conjugate prior of the binomial distribution (see e.g. Ref. [30] as well as Ref. [1] for practical applications). The Beta prime distribution is actually what has been independently derived in Sec. 3.3 to describe  $\rho_\lambda$ , although the beta special function was not used there, nor in Sec. 5.

## 6 Direct inference of the rate ratio $\rho$ (and of $r_2$ )

We have remarked several times that  $r_1$  and  $r_2$  are inferred from the observed numbers of events  $X_1$  and  $X_2$  (we assume  $T_1$  and  $T_2$  can be exactly known), and that the possible values of their ratio  $\rho$  are successively evaluated (‘deduced’) from each possible pair of values of the rates. This logical scheme is represented by the graphical model of Fig. 10. But this is not the only way to approach the problem. An alternative model is shown in Fig. 13, in which the node  $\rho$  appears ‘at the top’ of the



network and it is then really *inferred*<sup>26</sup> (indeed also  $r_2$  is ‘at the top’, having above it no *parents nodes* from which to depend).

Writing one diagram or another one is not just a question of drawing art. Indeed, the network reflects the supposed causal model (‘what depends from what’) and therefore the choice of the model can have an effect on the results. It is therefore important to understand in what they differ. In the model of Fig. 10 the rates  $r_1$  and  $r_2$  assume a primary role. We infer their values and, *as a byproduct*, we get  $\rho$ . In this new model, instead, it is  $\rho$  to have a primary role, together with one of the two rates (they cannot be both at the same level because there is a constraint between the three quantities). Our choice to make  $r_1$  depend on  $r_2$  is due to the fact that  $r_2$ , appearing at the denominator, can be seen as a ‘baseline’ to which the other rate is referred (obviously, here  $r_1$  and  $r_2$  are just names, and therefore the choice of their role depend on their meaning).

The strategy to get  $f(\rho | \dots)$  is then different, being this time  $\rho$  directly inferred using the Bayes theorem applied to the entire network. A strong advantage of this second model is that, as we shall see, its prior can be factorized (see also Ref. [1], especially Appendix A there, in which there is a summary of the formulae we are going to use).

In analogy to what has been done in detail in Ref. [1], the pdf of  $\rho$  is obtained in two steps: first infer  $f(\rho, r_1, r_2 | x_1, x_2, T_1, T_2)$ ; then get the pdf of  $\rho$  by marginalization. For the first step we need to write down the joint distribution of all variables in the network (apart from  $T_1$  and  $T_2$  which we consider just as fixed parameters, having usually negligible uncertainty) using the most convenient *chain rule*, obtained navigating bottom up the graphical model. Indicating, as in Ref. [1], with  $f(\dots)$  the joint pdf of all relevant variables, we obtain from the chain rule

$$f(\dots) = f(x_2 | r_2, T_2) \cdot f_0(r_2) \cdot f(x_1 | r_1, T_1) \cdot f(r_1 | r_2, \rho) \cdot f_0(\rho) \quad (78)$$

from which we can get, besides a normalization constant, the pdf’s of interest as

$$f(\rho | x_1, T_1, x_2, T_2) \propto \int_0^\infty \int_0^\infty f(\dots) dr_1 dr_2 \quad (79)$$

$$f(r_2 | x_1, T_1, x_2, T_2) \propto \int_0^\infty \int_0^\infty f(\dots) d\rho dr_1, \quad (80)$$

or the joint pdf  $f(r_2, \rho | x_1, T_1, x_2, T_2)$ , integrating only over  $r_1$ . Using explicit expressions of the pdf’s, of which  $f(r_1 | r_2, \rho)$  is just the Dirac delta  $\delta(r_1 - \rho \cdot r_2)$ ,<sup>27</sup> and

---

<sup>26</sup>For those who have doubts about the meaning of ‘deduction’ and ‘induction’, Ref. [35] is highly recommended (and they will discover that Sherlock Holmes was indeed not *deducing* explanations).

<sup>27</sup>It is interesting to note that there is an alternative way to get Eq. (9), starting from the joint

ignoring multiplicative factors, we can then only focus on

$$\tilde{f}(\dots) \propto r_2^{x_2} \cdot e^{-T_2 r_2} \cdot f_0(r_2) \cdot r_1^{x_1} \cdot e^{-T_1 r_1} \cdot \delta(r_1 - \rho \cdot r_2) \cdot f_0(\rho), \quad (81)$$

having indicated by  $\tilde{f}()$  the unnormalized pdf.

## 6.1 Inferred distribution of $\rho$

Let us start with pdf of  $\rho$ . Our starting point is

$$f(\rho | x_1, T_1, x_2, T_2) \propto \int_0^\infty \int_0^\infty \tilde{f}(\dots) dr_1 dr_2, \quad (82)$$

from which it follows

$$f(\rho | x_1, T_1, x_2, T_2) \propto \int_0^\infty \int_0^\infty r_2^{x_2} e^{-T_2 r_2} \cdot r_1^{x_1} \cdot e^{-T_1 r_1} \cdot \delta(r_1 - \rho \cdot r_2) \cdot f_0(r_2) \cdot f_0(\rho) dr_1 dr_2 \quad (83)$$

$$\propto \left[ \int_0^\infty r_2^{x_2} e^{-T_2 r_2} \cdot (\rho \cdot r_2)^{x_1} \cdot e^{-T_1 \rho r_2} \cdot f_0(r_2) dr_2 \right] \cdot f_0(\rho) \quad (84)$$

$$\propto \left[ \rho^{x_1} \cdot \int_0^\infty r_2^{x_1+x_2} \cdot e^{-(T_2+T_1 \cdot \rho) \cdot r_2} \cdot f_0(r_2) dr_2 \right] \cdot f_0(\rho), \quad (85)$$

in which we have explicitly factorized  $f_0(\rho)$ . Again, besides  $f_0(r_2)$ , we recognize in the integrand something proportional to a Gamma pdf. If we then model also  $f_0(r_2)$  by a Gamma of parameters  $\alpha_0$  and  $\beta_0$ , and again neglect irrelevant factors, we get

$$f(\rho | x_1, T_1, x_2, T_2) \propto \left[ \rho^{x_1} \cdot \int_0^\infty r_2^{x_1+x_2} \cdot e^{-(T_2+T_1 \cdot \rho) \cdot r_2} \cdot r_2^{\alpha_0-1} \cdot e^{-\beta_0 r_2} dr_2 \right] \cdot f_0(\rho) \quad (86)$$

$$\propto \left[ \rho^{x_1} \cdot \int_0^\infty r_2^{\alpha_0+x_1+x_2-1} \cdot e^{-(\beta_0+T_2+T_1 \cdot \rho) \cdot r_2} dr_2 \right] \cdot f_0(\rho). \quad (87)$$

Indicating, in analogy to what done to obtain Eq. (58), the power of  $r_2$  as  $\alpha_* - 1 = \alpha_0 + x_1 + x_2 - 1$ , and the factor multiplying  $r_2$  at the exponent as  $\beta_* = \beta_0 + T_2 + T_1 \cdot \rho$ ,

distribution  $f(\rho_\lambda, \lambda_1, \lambda_2 | x_1, x_2)$  and then marginalizing it. In fact, using the chain rule, we get

$$\begin{aligned} f(\rho_\lambda, \lambda_1, \lambda_2 | x_1, x_2) &= f(\rho_\lambda | \lambda_1, \lambda_2) \cdot f(\lambda_1 | x_1, x_1) \cdot f(\lambda_2 | x_1, x_2) \\ &= f(\rho_\lambda | \lambda_1, \lambda_2) \cdot f(\lambda_1 | x_1) \cdot f(\lambda_2 | x_2). \end{aligned}$$

But, being  $\rho_\lambda$  deterministically related to  $\lambda_1$  and  $\lambda_2$ ,  $f(\rho_\lambda | \lambda_1, \lambda_2)$  is nothing but  $\delta(\rho_\lambda - \lambda_1/\lambda_2)$  (see also other examples in Ref. [1]). Integrating then  $f(\rho_\lambda, \lambda_1, \lambda_2 | x_1, x_2)$  over  $\lambda_1$  and  $\lambda_2$  we get Eq. (9).

we get

$$f(\rho | x_1, T_1, x_2, T_2) \propto \left[ \rho^{x_1} \cdot \int_0^\infty r_2^{\alpha_* - 1} \cdot e^{-\beta_* \cdot r_2} dr_2 \right] \cdot f_0(\rho) \quad (88)$$

$$\propto \left[ \rho^{x_1} \cdot \frac{\Gamma(\alpha_*)}{\beta_*^{\alpha_*}} \right] \cdot f_0(\rho) \quad (89)$$

$$\propto \left[ \rho^{x_1} \cdot (\beta_0 + T_2 + T_1 \cdot \rho)^{-(\alpha_0 + x_1 + x_2)} \right] \cdot f_0(\rho). \quad (90)$$

What is interesting with this result is that we can consider the term inside the square brackets as an effective likelihood (remember that multiplicative factors are irrelevant), and therefore we can rewrite Eq. (90) as

$$f(\rho | x_1, T_1, x_2, T_2) \propto \mathcal{L}(\rho; x_1, T_1, x_2, T_2, \alpha_0, \beta_0) \cdot f_0(\rho). \quad (91)$$

For this reason *we can serenely proceed assuming a flat prior* about  $\rho$ , because we can reshape in a second step the result (see Ref. [1] for details). So, assuming  $f_0(\rho) = k$  and comparing the expression inside the square bracket of Eq. (90) with Eq. (60) we get the normalization just by analogy, thus getting

$$f(\rho | x_1, T_1, x_2, T_2) = \frac{\Gamma(\alpha_0 + x_1 + x_2)}{\Gamma(x_1 + 1) \cdot \Gamma(\alpha_0 + x_2 - 1)} \cdot T_1^{x_1 + 1} \cdot (\beta_0 + T_2)^{\alpha_0 + x_2 - 1} \cdot \rho^{x_1} \cdot (\beta_0 + T_2 + T_1 \cdot \rho)^{-(\alpha_0 + x_1 + x_2)}, \quad (92)$$

or

$$f(\rho | x_1, T_1, x_2, T_2) = \frac{T_1^{x_1 + 1} \cdot (\beta_0 + T_2)^{\alpha_0 + x_2 - 1}}{\mathbb{B}(x_1 + 1, \alpha_0 + x_2 - 1)} \cdot \rho^{x_1} \cdot (\beta_0 + T_2 + T_1 \cdot \rho)^{-(\alpha_0 + x_1 + x_2)}, \quad (93)$$

that, for a flat prior about  $r_2$ , i.e.  $\alpha_0 = 1$  and  $\beta_0 = 0$ , becomes

$$f(\rho | x_1, T_1, x_2, T_2) = \frac{\Gamma(x_1 + x_2 + 1)}{\Gamma(x_1 + 1) \cdot \Gamma(x_2)} \cdot T_1^{x_1 + 1} \cdot T_2^{x_2} \cdot \rho^{x_1} \cdot (T_2 + T_1 \cdot \rho)^{-(x_1 + x_2 + 1)} \quad (94)$$

$$= \frac{(x_1 + x_2)!}{x_1! \cdot (x_2 - 1)!} \cdot T_1^{x_1 + 1} \cdot T_2^{x_2} \cdot \rho^{x_1} \cdot (T_2 + T_1 \cdot \rho)^{-(x_1 + x_2 + 1)} \quad (95)$$

The comparison of this result with Eq. (62), obtained using flat priors for  $r_1$  and  $r_2$ , is at least surprising: the structures of the pdf's are the same, but  $x_2$  in Eq. (62) is

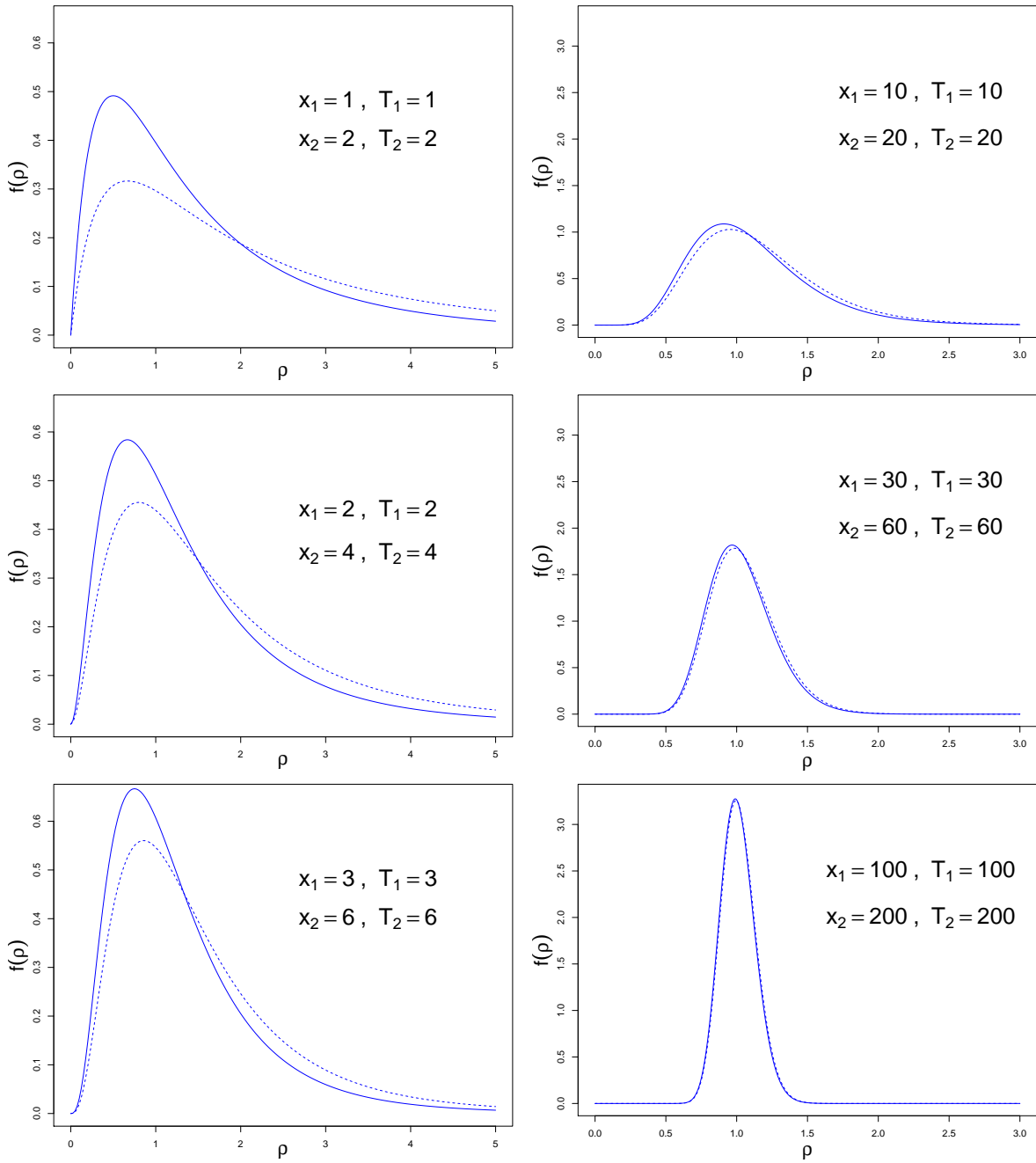


Figure 14: Dependence of the inference of  $\rho$  from the priors. Solid lines: flat priors on  $r_1$  and  $r_2$ , as in Figs. 11 and 12, following from the causal model depicted in Fig. 10. Dashed lines: flat priors on  $r_2$  and  $\rho$  (causal model of Fig. 13.)

replaced by  $x_2 - 1$  in Eq. (95). Obviously, the two results will coincide for large  $x_2$ , and also for small  $x_2$  there is not a dramatic difference, as we can see from Fig. 14. As far as the summaries of the distribution are concerned, we get

$$\text{mode}(\rho) = \frac{x_1/T_1}{(x_2 + 1)/T_2} \quad (96)$$

$$E(\rho) = \mu_\rho = \frac{(x_1 + 1)/T_1}{(x_2 - 1)/T_2} \quad (\mathbf{x}_2 > \mathbf{1}) \quad (97)$$

$$\sigma(\rho) = \sqrt{\mu_\rho \cdot \left( \frac{T_2}{T_1} \cdot \frac{x_1 + 2}{x_2 - 2} - \mu_\rho \right)} \quad (\mathbf{x}_2 > \mathbf{2}). \quad (98)$$

At this point, instead of taking comfort for the fact that the differences are irrelevant in practical cases, or *tout court* ‘rejecting Bayesian methods because of their dependence of priors’, it is interesting to try to understand the origin of this effect, certainly related to the priors.

But, before proceeding, let us not forget that Eq. (93) was obtained assuming a flat prior about  $\rho$  and that in that model this prior can be factorized. Therefore the more general pdf of the rate ratio for the model of Fig. 13 is

$$f(\rho | x_1, T_1, x_2, T_2) = \frac{1}{B(x_1 + 1, \alpha_0 + x_2 - 1)} \cdot T_1^{x_1 + 1} \cdot (\beta_0 + T_2)^{\alpha_0 + x_2 - 1} \cdot \rho^{x_1} \cdot (\beta_0 + T_2 + T_1 \cdot \rho)^{-(\alpha_0 + x_1 + x_2)} \cdot f_0(\rho), \quad (99)$$

having only the limitation (but in reality almost irrelevant, given the flexibility of the Gamma distribution) of depending on the chosen parametrization for  $f_0(r_2)$ .

## 6.2 Cross-influences of priors

One might say that in the first case, that of Fig. 10, yielding Eq. (62) starting from  $f_0(r_1) = f_0(r_2) = k$  there were no priors on  $\rho$ . But this is quite not true, because the flat priors on  $r_1$  and  $r_2$  impinge on the prior on  $\rho$ , due to the relation  $\rho = r_1/r_2$ . The easiest way to see what is going on is by Monte Carlo, that is, in R,

```
n = 10^7
rM = 100
r1 = runif(n, 0, rM)
r2 = runif(n, 0, rM)
rho = r1/r2
rho.h <- rho[rho<5]
hist(rho.h, nc=200, col='blue', freq=FALSE)
abline(v=1, col='red')
```

where the selection of the values below  $\rho=5$  is to visualize the more interesting region, shown in the top plot of Fig. 15 (a more complete script, which also performs the correct normalization of the histogram, is shown in Appendix B.4). The histogram is characterized by a plateau till  $\rho = 1$ , followed by a slow decreasing. Curiously, the histogram does not depend on the maximum value  $r_M$ .

Although it might be bizarre, this histogram shows in essence the prior on  $\rho$  we have been tacitly assumed, when flat priors on  $r_1$  and  $r_2$  were chosen (as a cross check, the commented instructions of the script of Appendix B.4, executed one by one, plot the distribution of  $r_1$  assuming a flat prior for  $r_2$  and the curious distribution of the top plot of Fig. 15 for  $\rho$ ).

In order to have a better insight of what is going on, the bottom plot of the same figure shows the histogram of  $\log_{10} \rho$ . The maximum is at  $\log \rho = 0$  and it decreases symmetrically, exponentially,<sup>28</sup> as  $|\log \rho|$  increases. This symmetry indicates that the probabilities to get a value of  $\rho$  below or above 1 are the same. The same conclusion, within the uncertainties due to sampling, can be drawn from the histogram in linear scale, since  $f(\rho)$  is ‘about 1/2’ for  $0 \leq \rho \leq 1$ . Similarly, from the comparison of the two histograms we can evaluate, by symmetry arguments, that the probability that  $\rho$  is between 0.1 and 10 is equal to 90% (exact value, indeed as we shall see in a while).

It is interesting to get the distribution shown in the top plot of Fig. 15 making a transformation of variables, as we have done in Eq. (9) and following equations:<sup>29</sup>

$$f(\rho) = \int_0^{r_M} \int_0^{r_M} \delta(\rho - r_1/r_2) \cdot f(r_1) \cdot f(r_2) dr_1 dr_2 \quad (100)$$

$$= \int_0^{r_M} \int_0^{r_M} r_2 \cdot \delta(r_1 - \rho \cdot r_2) \cdot \frac{1}{r_M} \cdot \frac{1}{r_M} dr_1 dr_2, \quad (101)$$

where  $r_M$  is the maximum value of  $r_1$  and  $r_2$ .<sup>30</sup>

---

<sup>28</sup>Empirically, we can evaluate, taking two points from the histogram of Fig. 15, the following exponential:  $f(\log(r)) \approx 1.16 \times \exp(-2.30 \cdot |\log(r)|)$ .

<sup>29</sup>Perhaps it is worth noticing again (see footnote 27), since this observation seems raised for the first time in this paper, that Eq. (100) can be seen not only as an extension to continuous variables of Eq. (2), but also as the joint pdf  $f(\rho, r_1, r_2)$  obtained by the chain rule, that is  $f(\rho, r_1, r_2) = f(\rho | r_1, r_2) \cdot f(r_1) \cdot f(r_2)$ , where  $f(\rho | r_1, r_2) = \delta(\rho - r_1/r_2)$ , followed by marginalization.

<sup>30</sup>If you like to reproduce the final result, given by Eq. (103) with *Mathematica*, here are the commands to get it, although the output will appear a bit cryptic (but you will recognize the resulting plot):

```
rM = 10
frho := Integrate[r2*DiracDelta[r1 - rho*r2]/rM^2, {r1, 0, rM}, {r2, 0, rM}]
frho
Plot[frho, {rho, 0, 5}]
```

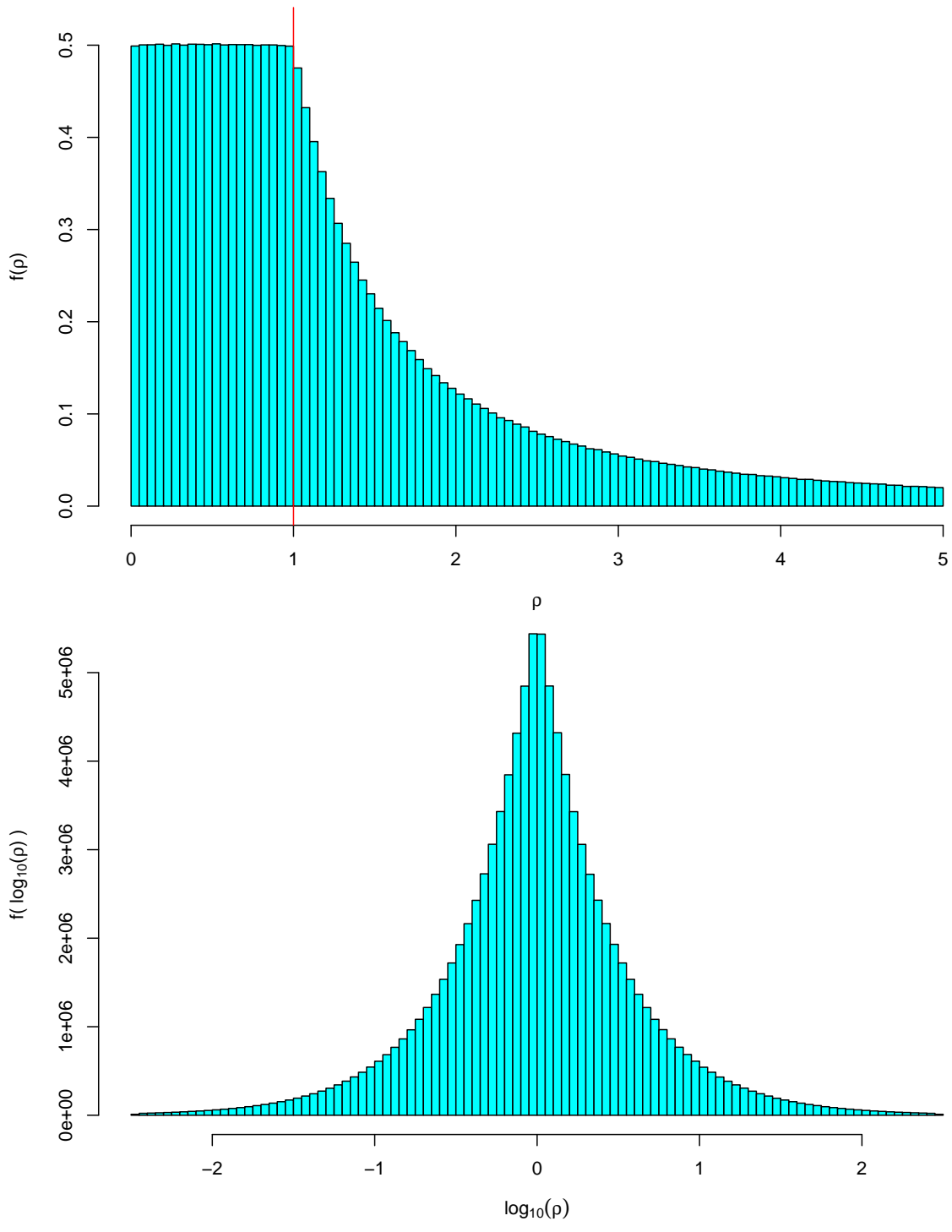


Figure 15: Distribution of  $\rho$  implied by flat priors on  $r_1$  and  $r_2$  in linear and log scale. The vertical line in the upper plot shows the discontinuity of the distribution at  $\rho = 1$ .

At this point, some care is needed with the limits of the integral over  $r_2$ , due to its ‘natural’ upper limit at  $r_M$  and to that given by the constraint  $\rho \cdot r_2 \leq 1$ , i.e.  $r_2 \leq 1/\rho$ . Therefore, after the trivial integration over  $r_1$ , we are left with

$$f(\rho) = \frac{1}{r_M^2} \cdot \int_0^{r_2^u} r_2 \, dr_2, \quad (102)$$

where the upper limit  $r_2^u$  depends on  $\rho$  in the following way:

$$\begin{aligned} \rho \leq 1 &\longrightarrow r_2^u = r_M \\ \rho > 1 &\longrightarrow r_2^u = r_M/\rho. \end{aligned}$$

and therefore

$$\begin{aligned} \rho \leq 1 &\longrightarrow f(\rho) = \frac{1}{r_M^2} \cdot \int_0^{r_M} r_2 \, dr_2 = \frac{1}{r_M^2} \cdot \frac{r_M^2}{2} = \frac{1}{2} \\ \rho > 1 &\longrightarrow f(\rho) = \frac{1}{r_M^2} \cdot \int_0^{r_M/\rho} r_2 \, dr_2 = \frac{1}{r_M^2} \cdot \frac{r_M^2}{2\rho^2} = \frac{1}{2\rho^2}, \end{aligned}$$

that we summarize as<sup>31</sup>

$$f\left(\rho \mid f(r_1)=\frac{1}{r_M}, f(r_2)=\frac{1}{r_M}\right) = \begin{cases} \frac{1}{2} & (0 \leq \rho \leq 1) \\ \frac{1}{2\rho^2} & (\rho > 1), \end{cases} \quad (103)$$

which, indeed, does not depend on the the maximum values of  $r_1$  and  $r_2$ , as we had already learned playing with Monte Carlo simulations.<sup>32</sup>

For completeness, let also make the game of seeing how flat priors on  $r_2$  and  $\rho$  (up to  $r_{2M}$  and  $\rho_M$ , respectively) are reflected into  $r_1$  in the model of Fig.13:

$$f(r_1) = \int_0^{\rho_M} \int_0^{r_{2M}} \delta(r_1 - \rho r_2) \cdot f(\rho) \cdot f(r_2) \, d\rho \, dr_2 \quad (104)$$

$$f(r_1) = \int_0^{\rho_M} \int_0^{r_{2M}} \frac{\delta(\rho - r_1/r_2)}{r_2} \cdot \frac{1}{\rho_M} \cdot \frac{1}{r_{2M}} \, d\rho \, dr_2 \quad (105)$$

$$= \frac{1}{\rho_M \cdot r_{2M}} \cdot \int_{r_{2L}}^{r_{2U}} \frac{1}{r_2} \, dr_2 \quad (106)$$

where the extremes of integration are  $r_{2L} = r_1/\rho_M$  and  $r_{2U} = r_{2M}$ .

---

<sup>31</sup>We can check that  $P(1/10 \leq \rho \leq 10) = 9/10$ , as previously guessed from symmetry arguments.

<sup>32</sup>Curiously, this distribution has the property that  $f(1/\rho) = f(\rho)$ . I wonder if there are others.



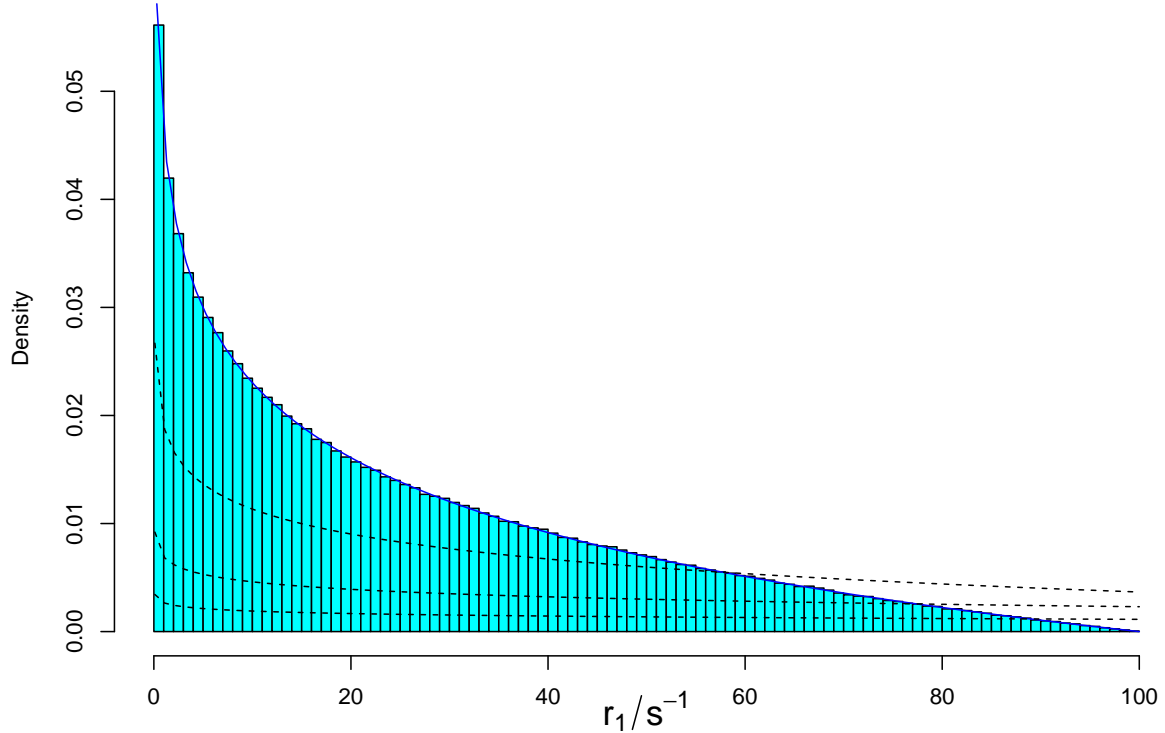


Figure 16: Histogram of  $r_1$  implied by priors on  $r_2$  and  $\rho$  flat up to  $\rho_M = r_{2M}/s^{-1} = 10$ , compared to the exact evaluation of the pdf (solid line). Dashed lines: pdf of  $r_1$  for  $\rho_M = r_{2M}/s^{-1} = 30, 100, 300$  (higher to lower, that is from steeper to flatter).

Here is, finally, the pdf of  $r_1$ , in which we have written explicitly the conditions:

$$f\left(r_1 \mid f_0(r_2)=\frac{1}{r_{2M}}, f_0(\rho)=\frac{1}{\rho_M}\right) = \frac{1}{\rho_M \cdot r_{2M}} \cdot \log\left(\frac{r_{2M} \cdot \rho_M}{r_1}\right) \quad (0 < r_1 \leq r_{2M} \cdot \rho_M) \quad (107)$$

An example with  $\rho_M = 1$  and  $r_{2M} = 10 \text{ s}^{-1}$  is reported in Fig. 16, in which the exact pdf (blue solid line) is compared with the Monte Carlo result. The plot also shows the pdf's of  $r_1$  for increasing maximum values ( $\rho_M = r_{2M}/s^{-1} = 30, 100, 300$ , from higher to lower curves). We see that for  $\rho_M \rightarrow \infty$  and  $r_{2M} \rightarrow \infty$  also the distribution of  $r_1$  becomes flat. This is an interesting result, showing that, contrary to the model of Fig. 10, the model of Fig. 13 can accommodate in practice flat prior distributions for the three quantities of interest.<sup>33</sup>

<sup>33</sup>But when measuring rates, a flat prior has more implications than one might think, as discussed in chapter 13 of Ref. [13], and therefore a full understanding of the physical case is desirable.

### 6.3 Final distributions of $r_1$ and $r_2$ (starting from flat initial distributions of $r_2$ and $\rho$ )

For completeness, let us also try to get the closed expressions of  $f(r_1 | x_1, T_1, x_2, T_2)$  and  $f(r_2 | x_1, T_1, x_2, T_2)$ , although only under the assumption of a flat prior of  $\rho$ . In this case this choice is forced from the fact that  $f_0(\rho)$  cannot be expressed in term of a conjugate prior which would then simplify the calculations. For the general case, in fact, we have to change methods, moving to Markov Chain Monte Carlo (MCMC), as done e.g. in Ref. [1] and as it will be sketched in the next section.

In order to get the pdf of  $r_1$ , we need to restart from the unnormalized joint distribution (81), proceeding then like in Eq. (82), but this time integrating over  $r_2$  and  $\rho$  and absorbing the constant priors in the proportionality factor:

$$f(r_1 | x_1, T_1, x_2, T_2) \propto \int_0^\infty \int_0^\infty \tilde{f}(\dots) d\rho dr_2 \quad (108)$$

$$\propto \int_0^\infty \int_0^\infty r_2^{x_2} \cdot e^{-T_2 r_2} \cdot r_1^{x_1} \cdot e^{-T_1 r_1} \cdot \delta(r_1 - \rho \cdot r_2) d\rho dr_2 \quad (109)$$

$$\propto \int_0^\infty \int_0^\infty r_2^{x_2} \cdot e^{-T_2 r_2} \cdot r_1^{x_1} \cdot e^{-T_1 r_1} \cdot \frac{\delta(\rho - r_1/r_2)}{r_2} d\rho dr_2 \quad (110)$$

$$\propto r_1^{x_1} \cdot e^{-T_1 r_1} \cdot \int_0^\infty r_2^{x_2-1} e^{-T_2 r_2} dr_2 \quad (111)$$

$$\propto r_1^{x_1} \cdot e^{-T_1 r_1}, \quad (112)$$

thus reobtaining, besides normalization, Eq. (35).

Similarly, we have

$$f(r_2 | x_1, T_1, x_2, T_2) \propto \int_0^\infty \int_0^\infty \tilde{f}(\dots) d\rho dr_1 \quad (113)$$

$$\propto \int_0^\infty \int_0^\infty r_2^{x_2} \cdot e^{-T_2 r_2} \cdot r_1^{x_1} \cdot e^{-T_1 r_1} \cdot \delta(r_1 - \rho \cdot r_2) d\rho dr_1 \quad (114)$$

$$\propto \int_0^\infty r_2^{x_2} \cdot e^{-T_2 r_2} \cdot (\rho \cdot r_2)^{x_1} \cdot e^{-T_1 \rho r_2} d\rho \quad (115)$$

$$\propto r_2^{x_2+x_1} \cdot e^{-T_2 r_2} \cdot \int_0^\infty \rho^{x_1} \cdot e^{-T_1 r_2 \rho} d\rho \quad (116)$$

$$\propto r_2^{x_2+x_1} \cdot e^{-T_2 r_2} \cdot \frac{\Gamma(x_1 + 1)}{(r_2 T_1)^{(x_1+1)}} \quad (117)$$

$$\propto r_2^{x_2+x_1} \cdot e^{-T_2 r_2} \cdot r_2^{-(x_1+1)} \quad (118)$$

$$\propto r_2^{x_2-1} \cdot e^{-T_2 r_2}. \quad (119)$$

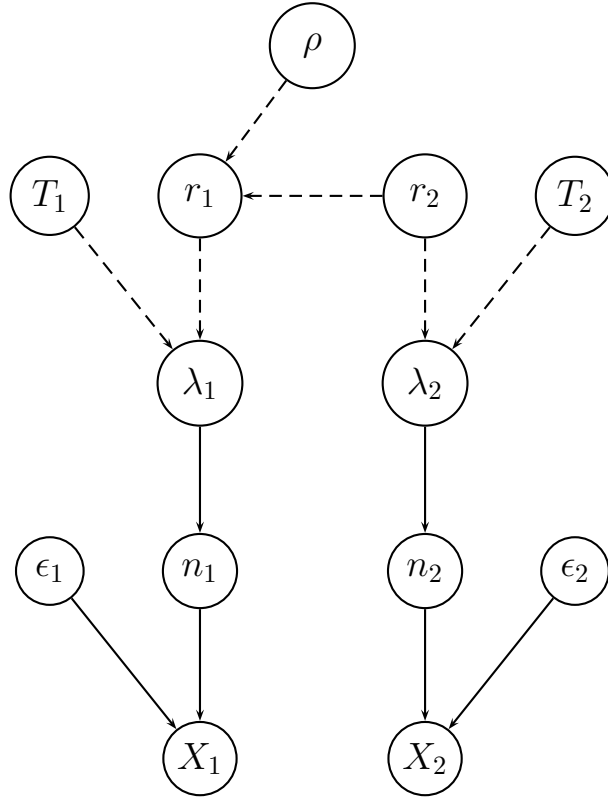


Figure 17: Extension of the model of Fig. 13 in order to include efficiencies.

We see that, differently from  $f(r_1 | x_1, T_1, x_2, T_2)$ , the power of  $r_2$  is, instead of  $x_2$ ,  $x_2 - 1$ , that is we get an effect similar to that found for the distribution of  $\rho$ . As a consequence, expected values and standard deviation of  $r_2$  are  $x_2/T_2$  and  $\sqrt{x_2}/T_2$ , respectively.

## 7 Use of MCMC methods to cross-check the closed results and to analyze extended models

So far our models have been rather simple, missing however several real life complications. For example, assuming that we do observe the number of counts due to a Poisson distribution with a given  $\lambda = r \cdot T$  clearly implies that we are neglecting *efficiency* issues. In order to include efficiencies we need to modify our graphical model of Fig. 13 (hereafter we stick to this last model), adding the relevant nodes.

The extended model is shown in Fig. 17, in which we have redefined the symbols,

keeping  $X_1$  and  $X_2$  associated to the observed counts and then calling  $n_1$  and  $n_2$  those ‘produced’ by the Poissonians. Each  $X_i$  is then binomially distributed with parameters  $n_i$  and  $\epsilon_i$ . In summary, listing the ‘causal relations’ from bottom to top, we have

$$X_i \sim \text{Binom}(n_i, \epsilon_i) \quad (120)$$

$$n_i \sim \text{Poisson}(\lambda_i) \quad (121)$$

$$\lambda_i = r_i \cdot T_i \quad (122)$$

$$r_1 = \rho \cdot r_2 \quad (123)$$

At this point we can easily build up the joint distribution of all quantities in the network, as we have done in the previous section, and then evaluate the (possibly joint) distribution of the variables of interest, conditioned by those which are observed or somehow assumed. Moreover, also the efficiencies  $\epsilon_1$  and  $\epsilon_2$  are by themselves uncertain, and then we have to integrate also over them, taking into account their probability distributions  $f(\epsilon_1)$  and  $f(\epsilon_2)$ . In fact, their value come from test experiments or, more likely, from Monte Carlo simulations of the physics process and of the detector. So we need to enlarge the model adding four other nodes, taking into account the probabilistic links

$$X_i^{(MC)} \sim \text{Binom}(n_i^{(MC)}, \epsilon_i). \quad (124)$$

We refrain from adding the four nodes in the network of Fig. 17, which will become more busy in a while. Anyway, we can just assign to  $\epsilon_1$  and  $\epsilon_2$  the parameter of the probability distribution resulting from the inferences based on Monte Carlo simulations (see Ref. [1] for details – remember that, having the nodes  $\epsilon_1$  and  $\epsilon_2$  no parents, they need priors).

What is still missing in the model of Fig. 17 is *background*. In fact, we do not only lose events because of inefficiencies, but the ‘experimentally defined class’ can get contributions from other ‘physical class(es)’ (in general there are several physical classes contributing as background). Figure 18 shows the extension of the previous model, in which each Poisson process which describes the *signal* has just one background Poisson process. All variables have subscripts  $S$  or  $B$ , depending if their are associated to signal or background (with exception of  $r_1$  and  $r_2$ , which are obviously the two signal rates). As before, the nodes needed to infer the efficiencies are not shown in the diagram, which is therefore missing eight ‘bubbles’.

At this point it is clear that trying to achieve closed formulae is out of hope, and we need to use other methods to perform the integrals of interest, namely those based on Markov Chain Monte Carlo. We show here how to use a powerful package that does the work for us. But we do it only for the two cases of which we already have

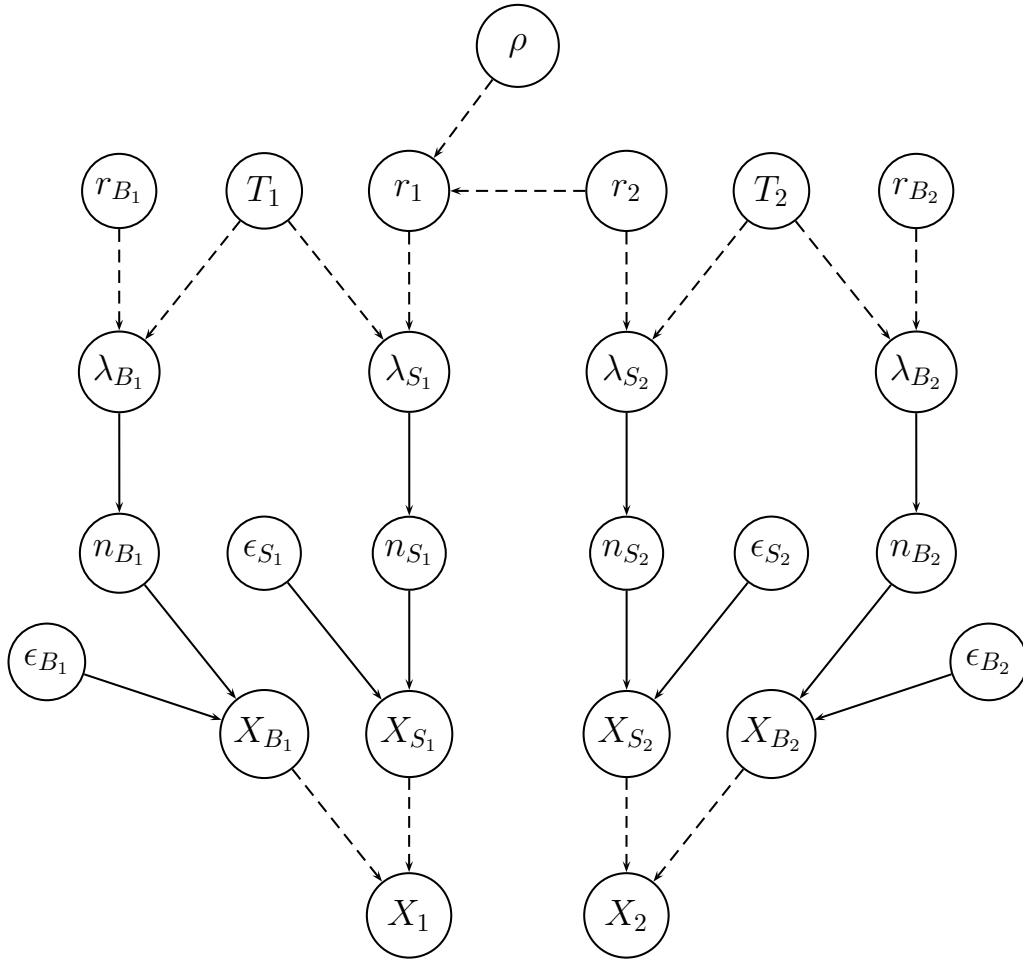


Figure 18: Extended model of Fig. 17 including also background.

closed solutions in hand, that is the models of Figs. 10 and 13 starting from uniform priors for the ‘top nodes’. The program we are going to use is *JAGS* [36] interfaced to R via the package *jags* [37].

[Introducing MCMC and related algorithms goes well beyond the purpose of this paper and we recommend Ref. [38] (some examples of application, including R scripts, are also provided in Ref. [1]). Moreover, mentioning the Gibbs Sampler algorithm applied to probabilistic inference (and forecasting) it is impossible not to refer to the *BUGS project* [39], whose acronym stands for Bayesian inference using Gibbs Sampler, that has been a kind of revolution in Bayesian analysis, decades ago limited to simple cases because of computational problems (see also Sec. 1 of Ref.[36]). In the BUGS project web site [40] it is possible to find packages with excellent Graphical User Interface, tutorials and many examples [41].]

## 7.1 Model A (Fig. 10), with flat priors on $r_1$ and $r_2$

We start from the model of Fig. 10. The code that instructs JAGS about the model is practically a transcription of the expressions to state that a variable follows a given distribution. Therefore, since we have

$$X_1 \sim \text{Poisson}(\lambda_1) \quad (125)$$

$$X_2 \sim \text{Poisson}(\lambda_2) \quad (126)$$

$$\lambda_1 = r_1 \cdot T_1 \quad (127)$$

$$\lambda_2 = r_2 \cdot T_2 \quad (128)$$

$$\rho = r_1/r_2, \quad (129)$$

we get

```
model {
  x1 ~ dpois(lambda1)
  x2 ~ dpois(lambda2)
  lambda1 <- r1 * T1
  lambda2 <- r2 * T2
  r1 ~ dgamma(1, 1e-6)
  r2 ~ dgamma(1, 1e-6)
  rho <- r1/r2
}
```

in which are also included the flat priors of  $r_1$  and  $r_2$ ,<sup>34</sup> implemented by Gamma distributions with  $\alpha = 1$  and  $\beta \lll 1$ :

$$r_1 \sim \text{Gamma}(1, 10^{-6}) \quad (130)$$

$$r_2 \sim \text{Gamma}(1, 10^{-6}) \quad (131)$$

The complete R script which calls `rjags` and shows the results is provided in Appendix B.5 (see Ref. [1] for clarifications about the structure of the R code). The values of  $(x_1 = 3, T_1 = 3 \text{ s})$  and  $(x_2 = 6, T_2 = 6 \text{ s})$  have been chosen in order to have small numbers, but with finite expected values and standard deviation, in order to make a comparison with the results of the closed formulae. The parameter determining the ‘length’ of the Markov chain has been set at  $10^5$ .

---

<sup>34</sup>Note that, since *priors are logically needed*, programs of this kind require them, even if they are flat. This can be seen as an annoyance, but it is instead a power of these programs: first they can include also non trivial priors; second, even if one wants to use flat priors, the user is forced to think on the fact that priors are unavoidable, instead of following the illusion that she is using a prior-free method [42], sometimes very dangerous, unless one does simple routine measurements characterized by a very narrow likelihood [13].

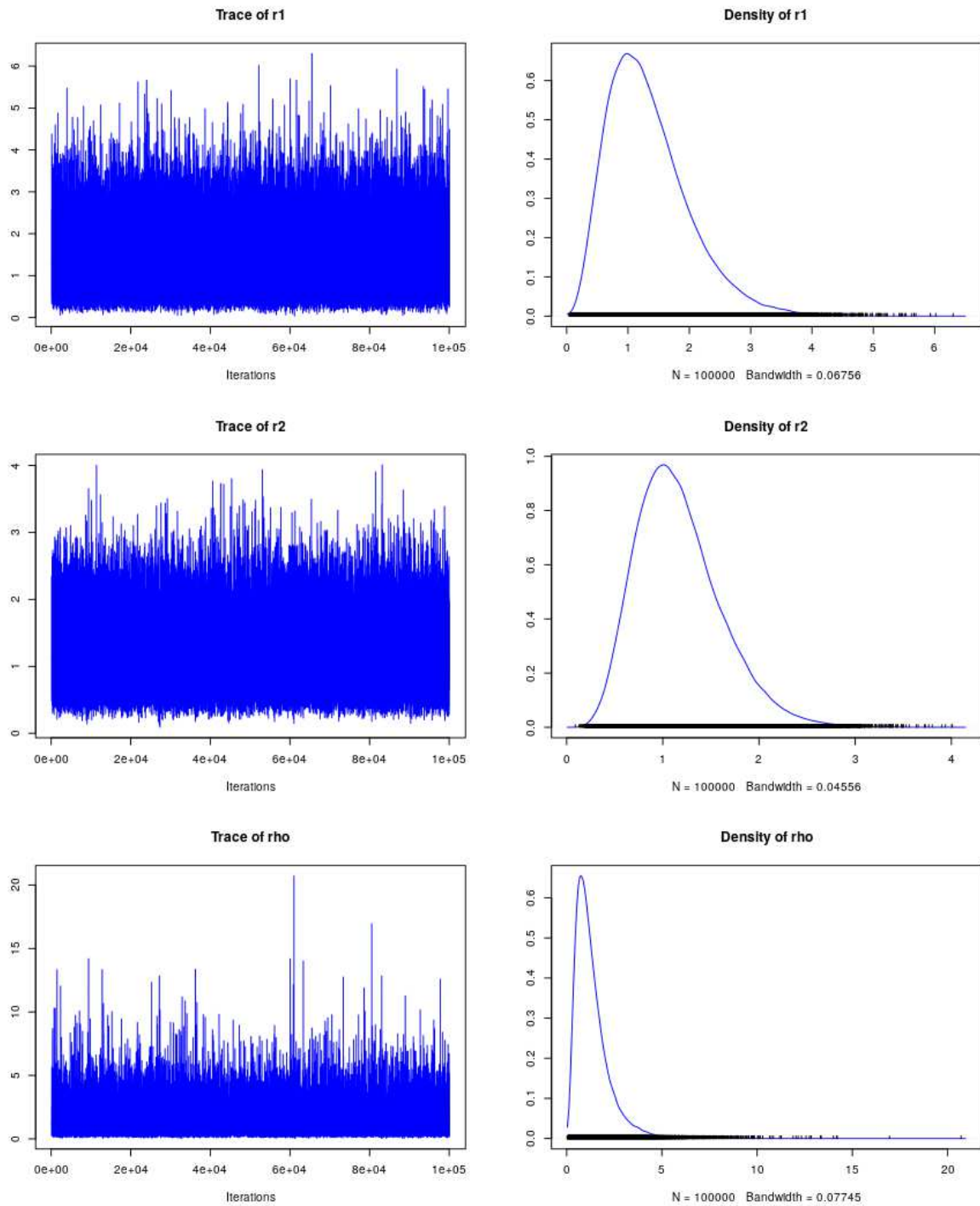


Figure 19: Graphical summary of the chain produced by the script of Appendix B.5 implementing the graphical model of Fig. 10.

The summary figure 19, drawn automatically by R when the command `plot()` is called with first argument an *MCMC chain object*, shows, for each of the three variables that we have chosen to *monitor*, the ‘trace’ and the ‘density’. The latter is a smoothed representation of the histogram of the possible occurrences of a variable in the chain. The former shows the ‘history’ of a variable during the sampling, and it is important to understand the quality of the sampling. If the traces appear quite randomic, as they are in this figure, there is nothing to worry. Otherwise we have to increase the length of the chain so that it can visit each ‘point’ (in fact a little volume) of the space of possibilities with relative frequencies ‘approximately equal’ to their probabilities (just *Bernoulli theorem*, nothing to do with the ‘frequentist definition of probability’).

Here is the relevant output of the script:

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
r1	1.334	0.6671	0.002110	0.002110
r2	1.167	0.4418	0.001397	0.001397
rho	1.333	0.9444	0.002986	0.002986

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
r1	0.3637	0.8457	1.223	1.700	2.927
r2	0.4701	0.8481	1.111	1.424	2.179
rho	0.2772	0.7048	1.102	1.684	3.770

Exact:

```

r1 = 1.333 +- 0.667
r2 = 1.167 +- 0.441
rho = 1.333 +- 0.943

```

As we can see, the agreement between the MCMC and the exact results, evaluated from Eqs. (64) and (65), is excellent (remember that ‘`r1 = 1.333 +- 0.667`’ stands for  $E(r_1) = 1.333 \text{ s}^{-1}$  and  $\sigma(r_1) = 0.667 \text{ s}^{-1}$ ).



## 7.2 Model B (Fig. 13), with flat priors on $\rho$ and $r_2$

Let us move to the model of Fig. 13, whose implementation in the JAGS language is the following:

```
model {
  x1 ~ dpois(lambda1)
  x2 ~ dpois(lambda2)
  lambda1 <- r1 * T1
  lambda2 <- r2 * T2
  r1 <- rho * r2
  r2 ~ dgamma(1, 1e-6)
  rho ~ dgamma(1, 1e-6)
}
```

The complete R script, which uses the same data ( $x_1=3$ ,  $T_1=3$ s;  $x_2=6$ ,  $T_2=6$ s) is provided in Appendix B.6. The result is shown in Fig. 20 and the details are given in the following printouts

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
r1	1.334	0.6694	0.002117	0.002117
r2	1.002	0.4068	0.001286	0.001925
rho	1.595	1.1918	0.003769	0.006058

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
r1	0.3616	0.8438	1.2234	1.704	2.941
r2	0.3671	0.7061	0.9477	1.238	1.940
rho	0.3167	0.8199	1.2923	2.012	4.638

Exact:

```
r1 = 1.333 +- 0.667
r2 = 1.000 +- 0.408
rho = 1.600 +- 1.200
```

Again, the agreement between the MCMC and the exact results is excellent.

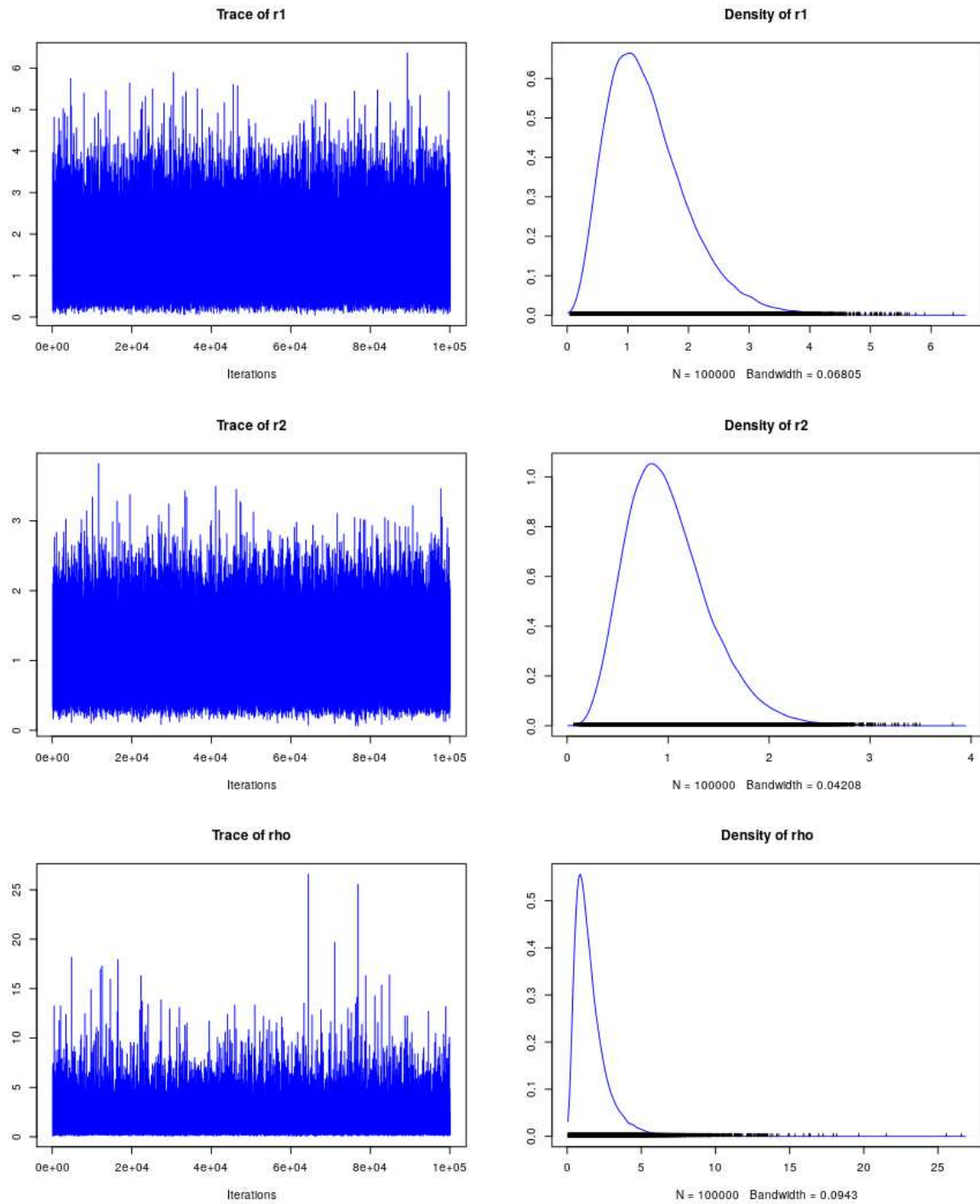


Figure 20: Graphical summary of the chain produced by the script of Appendix B.6 implementing the graphical model of Fig. 13.

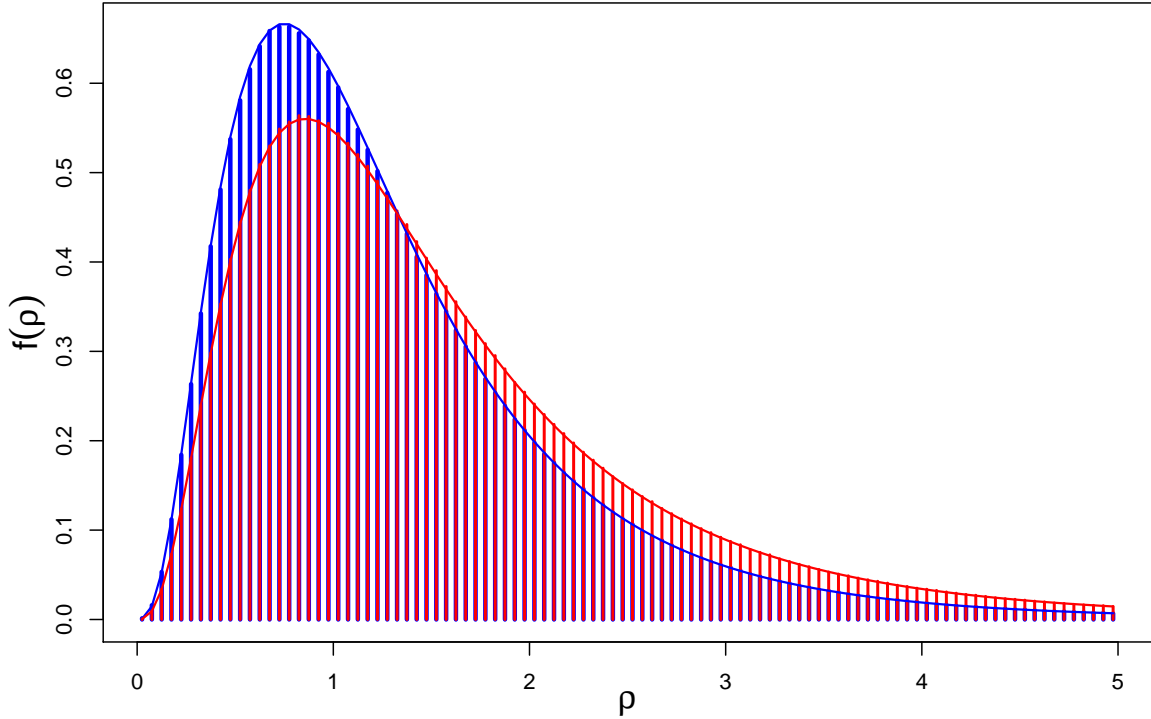


Figure 21: Comparison of the distribution of  $\rho = r_1/r_2$  obtained by the models of Fig. 10 (blue, slightly narrower) and Fig. 13 (red, slightly wider) in the case of  $(x_1 = 3, T_1 = 3 \text{ s})$  and  $(x_1 = 6, T_1 = 6 \text{ s})$  using flat priors for the top nodes. The histograms are the JAGS results and the lines come from the pdf's in closed form (see text).

### 7.3 Comparison of the results from the two models

An overall comparison of the two models, again based on the observations of 3 counts in 3s from process 1 and 6 counts in 6s from process 2, is shown in Fig. 21, while expected values and standard deviations (separated by ‘ $\pm$ ’) calculated from the closed formulae are summarized in the following table.

	Model A (Fig. 10) [ $f_0(r_1) = k$ & $f_0(r_2) = k$ ]	Model B (Fig. 13) [ $f_0(\rho) = k$ & $f_0(r_2) = k$ ]
$r_1$ ( $\text{s}^{-1}$ )	$1.33 \pm 0.67$	$1.33 \pm 0.67$
$r_2$ ( $\text{s}^{-1}$ )	$1.17 \pm 0.44$	$1.00 \pm 0.41$
$\rho$	$1.33 \pm 0.94$	$1.60 \pm 1.20$

As we have seen in Fig. 14, the second model produces a distribution of  $\rho$  with higher expected value and higher standard deviation.

## 7.4 Dependence of the rate ratio from a physical quantity

Another interesting question is how to approach the problem of a ratio of rates that depends on the value of another physical quantity. That is we assume a dependence of  $\rho$  from  $v$  (symbol for a generic *variable*),

$$\rho = g(v; \boldsymbol{\theta}_\rho), \quad (132)$$

with  $\boldsymbol{\theta}_\rho$  the set of parameters of the functional dependence. The simplest and best understood case is the linear dependence

$$\rho = m \cdot v + c, \quad (133)$$

where  $\boldsymbol{\theta}_\rho = \{m, c\}$ , treated in detail in Ref. [43] where we used the same approach we are adopting here. In analogy to what done in Fig. 1 there, we can extend the Model B of Fig. 13 to that of Fig. 22 (we continue to neglect efficiency and background issues in order to focus to the core of the problem). Moreover, as in Fig. 1 of Ref. [43], we have considered the fact that the physical quantity  $v$  is ‘*experimentally observed*’ as  $v_O$ . In the simple case of a linear dependence the model is described by the following relations among the variables

$$X_{1j} \sim \text{Poisson}(\lambda_{1j}) \quad (134)$$

$$X_{2j} \sim \text{Poisson}(\lambda_{2j}) \quad (135)$$

$$\lambda_{1j} = r_{j1} \cdot T_{1j} \quad (136)$$

$$\lambda_{2j} = r_{j2} \cdot T_{2j} \quad (137)$$

$$r_{1j} = \rho_j \cdot r_{2j} \quad (138)$$

$$\rho_j = m \cdot v_j + c \quad (139)$$

$$v_{Oj} \sim \mathcal{N}(v_j, \sigma_{E_j}), \quad (140)$$

in which we have assumed a Gaussian (‘normal’) error function of  $v_{Oj}$  around  $v_j$ , with standard deviations  $\sigma_{E_j}$ . But the description of the model provided by the above relations is not complete (besides the complications related to inefficiencies and background, that we continue to neglect). In fact, we miss priors for  $v_j$ ,  $r_{2j}$  and  $\boldsymbol{\theta}_\rho$ , as they have no parent nodes (instead, we continue to consider  $T_{1j}$  and  $T_{2j}$  ‘exactly known’, being their uncertainty usually irrelevant).

The priors which are easier to choose are those of  $v_j$ , if their values are ‘well measured’, that is if  $\sigma_{E_j}$  are small enough. We can then confidently use flat priors, as done e.g. for the ‘unobserved’  $\mu_{y_i}$  of Fig. 1 in Ref. [43].

Also the priors about  $\boldsymbol{\theta}_\rho$  can be chosen quite vague, paying however some care in order to forbid negative values of  $\rho$ . Incidentally, having mentioned the simple case

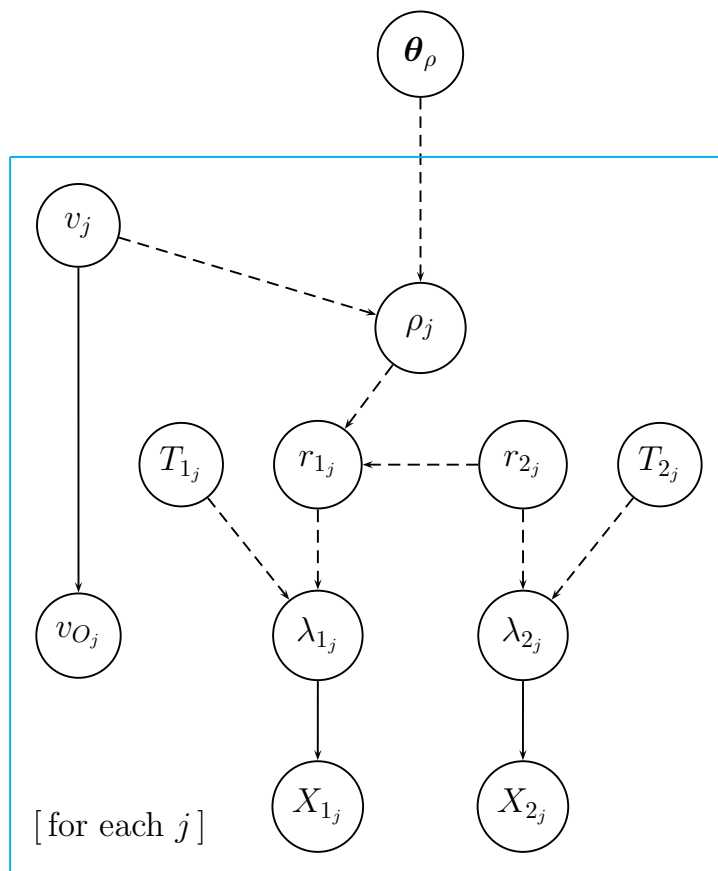


Figure 22: Further extension of the model of Fig. 13 (neglecting the ‘complications’ of the models of Figs. 17 and 18) to take into account that each value  $\rho_j$  might depend on the physical quantity  $v_j$ , measured as  $v_{m_j}$  via the set of parameters  $\theta$

of linear dependence, an important sub-case is when  $m$  is assumed to be null: the remaining prior on  $c$  becomes indeed the prior on  $\rho$ , and the inference of  $c$  corresponds to the inferred value of  $\rho$  having taken into account several instances of  $X_1$  and  $X_2$  – this is indeed the question of the ‘combination of values of  $\rho$ ’ on which we shall comment a bit more in detail in the sequel.

As far as the priors of the rates are concerned, one could think, a bit naively, that the choice of *independent* flat priors for  $r_{2j}$  could be a reasonable choice. But we need to understand the physical model underlying this choice. In fact, most likely, as the ratio  $\rho$  might depends on  $v$ , the same could be true for  $r_2$ , but perhaps with a completely different functional dependence. For example  $r_1$  and  $r_2$  could have a strong dependence on  $v$ , e.g. they could decrease exponentially, but, nevertheless,

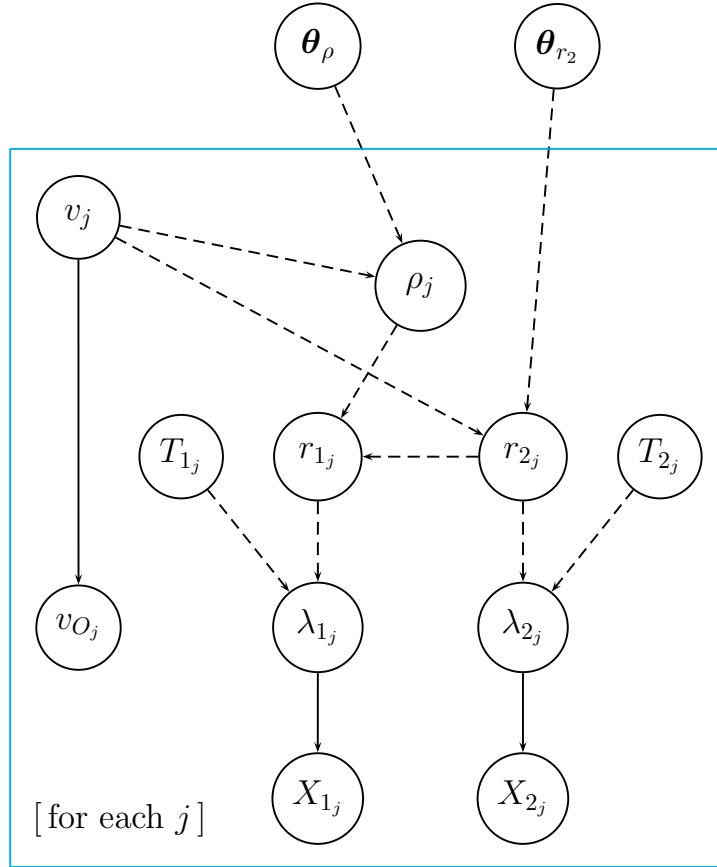


Figure 23: Extension of the model of Fig. 22, but making also  $r_2$  depend on  $v$  according to a suited law depending on the set of parameters  $\theta_{r_2}$ .

their ratio could be independent of  $v$ , or, at most, could just exhibit a small linear dependence. Therefore we have to add this possibility into the model, which then becomes as in Fig. 23, in which we have included a set of parameters  $\theta_{r_2}$ , such that

$$r_2 = h(v; \theta_{r_2}), \quad (141)$$

and, needless to say, some priors are required for  $\theta_{r_2}$ .

At this point, any further consideration goes beyond the rather general purpose of this paper, because we should enter into details that strongly depend on the physical case. We hope that the reader could at least appreciate the level of awareness that these graphical models provide. The existence of computing tools in which the models can be implemented makes then nowadays possible what decades ago was not even imaginable.

## 7.5 Combining ratios of rates

Let us end the work with the related topic of ‘combining several values’ of  $\rho$ , a problem we have already slightly touched above. Let us start phrasing it in the terms we usually hear about it. Imagine we have in hand  $N$  instances of  $(X_1, T_1, X_2, T_2)$ . From each of them we can get a value of  $\rho$  with ‘its uncertainty’. Then we might be interested in getting a single value, combining the individual ones.

The first idea that might come to the mind is to apply the well known weighted average of the individual values, using as weights the inverses of the variances. But, before doing it, it is important to understand the assumptions behind it, that is something that goes back to none other than Gauss, and for which we refer to Refs. [29, 44]. The basic idea of Gauss was to get two numbers (let us say ‘central value’ and standard deviation – indeed Gauss used, instead of the standard deviation, what he called ‘degree of precision’ and ‘degree of accuracy’ [44], but this is an irrelevant detail) such that *they contain the same information of the individual values*. In practice the rule of combination had to satisfy what is currently known as *statistical sufficiency*. Now it is not obvious at all that the weighted average using  $E(\rho)$  and  $\sigma(\rho)$  satisfies sufficiency (see e.g. the puzzle proposed in the Appendix of Ref. [44]).

Therefore, instead of trying to apply the weighted average as a ‘prescription’, let us see what comes out applying consistently the rules of probability on a suitable model, restarting from that of Fig. 23. It is clear that if we consider meaningful a combined value of  $\rho$  for all instances of  $(X_1, T_1, X_2, T_2)$  it means we assume  $\rho$  not depending on a quantity  $v$ . However,  $r_2$  could. This implies that the values of  $r_2$  are strongly correlated to each other.<sup>35</sup> Therefore the graphical model of interest would be that at the top of Fig. 24. Again, at this point there is little more to add, because what would follow depends on the specific physical model.

A trivial case is when both rates, and therefore their ratio, are assumed to be constant, although unknown, yielding then the graphical model shown in the bottom diagram of Fig. 24, whose related joint pdf, evaluated by the best suited chain rule, is an extension of Eqs. (78)-(81)

$$f(\dots) = \left[ \prod_{j=1}^N f(x_{2j} | r_2, T_{2j}) \right] \cdot f_0(r_2) \cdot \left[ \prod_{j=1}^N f(x_{1j} | r_1, T_{1j}) \right] \cdot f(r_1 | r_2, \rho) \cdot f_0(\rho), \quad (142)$$

---

<sup>35</sup>At this point a clarification is in order. When we make fits and say, again with reference to Fig. 1 of Ref. [43], that the observations  $y_i$  are independent from each other we are referring to the fact that each  $y_i$  depends only on its  $\mu_{y_i}$ , e.g.  $y_i \sim \mathcal{N}(\mu_{y_i}, \sigma_Y)$ , but not on the other  $y_{j \neq i}$ . Instead, the *true values*  $\mu_{y_i}$  are *certainly correlated*, being  $\mu_y = \mu_y(\mu_x; \theta)$ .

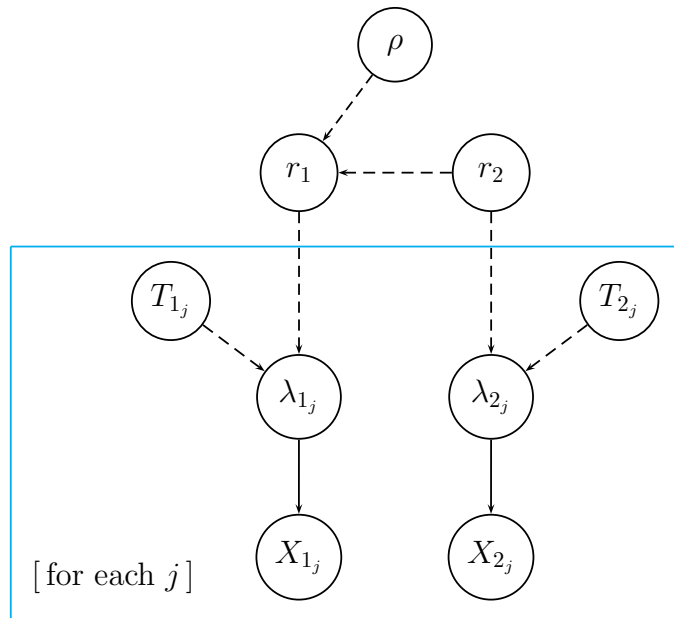
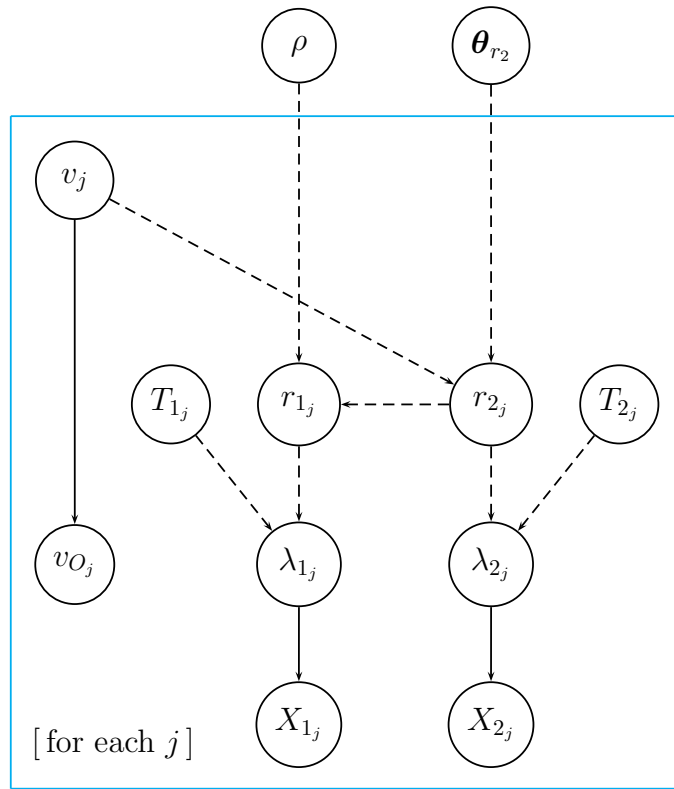


Figure 24: Possible reductions of the model of Fig. 23 for the 'combination' of  $\rho$  (see text).



from which the unnormalized joint pdf follows:

$$\tilde{f}(\dots) \propto \left[ \prod_{j=1}^N r_2^{x_{2j}} \cdot e^{-T_{2j} r_2} \right] \cdot f_0(r_2) \cdot \left[ \prod_{j=1}^N r_1^{x_{1j}} \cdot e^{-T_{1j} r_1} \right] \cdot \delta(r_1 - \rho \cdot r_2) \cdot f_0(\rho) \quad (143)$$

$$\propto \left[ r_2^{x_{2tot}} \cdot e^{-T_{2tot} r_2} \right] \cdot f_0(r_2) \cdot \left[ r_1^{x_{1tot}} \cdot e^{-T_{1tot} r_1} \right] \cdot \delta(r_1 - \rho \cdot r_2) \cdot f_0(\rho). \quad (144)$$

We recognize the same structure of Eq. (81), with  $x_1$  replaced by  $x_{1tot} = \sum_j x_{1j}$ ,  $T_1$  by  $T_{1tot} = \sum_j T_{1j}$ ,  $x_2$  by  $x_{2tot} = \sum_j x_{2j}$  and  $T_2$  by  $T_{2tot} = \sum_j T_{2j}$ . We get then the same result obtained in Sec. 6.1 if we use the total numbers of counts in the total times of measurements. This is a simple and nice result, close to the intuition, but we have to be aware of the model on which it is based.

## 8 Conclusions

In this paper we have dealt with the often debated issue of ‘ratios of small numbers of events’, approaching it from a probabilistic perspective. After having shown the difference between predicting numbers of counts (and their ratios) and inferring the Poisson parameters (and their ratios) on the base of the observed numbers of counts, the attention has been put on the latter, “*a problem in the probability of causes, . . . the essential problem of the experimental method*” [21]. Having the paper a didactic intent, the basic ideas of probabilistic inference have been reminded, together with the use of conjugate priors in order to get closed results with minimum effort. It has been also shown how to perform the so called ‘propagation of uncertainties’ in closed forms, which has required, for the purposes of this work, to derive the probability density function of the ratio of Gamma distributed variables. And, as byproducts, the ‘curious’ pdf of the ratio of two uniform variables has been derived and a new derivation of the formula to get the pdf of a function of variables has been devised.

The importance of graphical models has been stressed. In fact, they are not only very useful to form a global, clearer vision of the problem, but also to possibly take into account alternative models. In the case of rather simple models it has been shown how to write down the joint distribution of all variables, from which the pdf of the variables of interest follows. In some cases, thanks to reasonable (or at least well stated) assumptions, closed results have been obtained, but we have also seen how to use tools based on MCMC, both to check the closed results and to tackle more realistic models (samples of programming code are provided in Appendix B).

Finally, as far as the issue of ‘combination of ratios’ is concerned, it has been shown how the solution depends crucially on the physical model describing the variation of the rates and/or their ratio in function of an external variable. Therefore only general indications on how to approach the problem have been given, highly recommending

the use of MCMC tools (my preference for small problem with limited amount of data goes presently to JAGS/rjags, but particle physicists might prefer BAT [45], or perhaps the more recent, Julia [46] based, BAT.jl [47]).

I am indebted to Alfredo (Dino) Esposito for many discussions on the probabilistic and technical aspects the paper, some of which admittedly based on Ref. [1], and for valuable comments on the manuscript.

## References

- [1] G. D'Agostini and A. Esposito, *Checking individuals and sampling populations with imperfect tests*, arXiv:2009.04843 [q-bio.PE].
- [2] W. Nelson, *Confidence intervals for the ratio of two Poisson means and Poisson predictor intervals*, IEEE Trans. on Reliability, Vol. R-19 (1970) 42-49.
- [3] F. James and M Roos, *Errors on ratios of small numbers of events*, Nuclear Physics **B172** (1980) 475-470.
- [4] K.J. Coakley, D.S. Simons and A.M. Leifer, *Secondary ion mass spectroscopy measurements in isotopic ratios: corrections for time varying count rate*, Int. J. of Mass Spectrometry **240** (2005) 107-120.
- [5] K. Gu, H.K.T. Ng, M.L. Tang and W.R. Schucany, *Testing the ratio of two Poisson rates*, Biometrical Journal **50** (2008) 283-298.
- [6] R.C. Ogliore, G.R. Huss and K. Nagashima, *Ratio estimation in SIMS analysis*, Nucl. Instr. and Meth, in Phys. Reas. **B269** (2011) 1910-1918.
- [7] C.D. Coath, R.C.J. Steele and W.F. Lunnon, *Statistical bias in isotope ratios*, J. Anal. At. Spectrom., 2013, **28**, 52-58.
- [8] G. D'Agostini, *Overcoming priors anxiety, Bayesian Methods in the Sciences*, J. M. Bernardo Ed., special issue of *Rev. Acad. Cien. Madrid*, Vol. 93, Num. 3, 1999, arXiv:physics/9906048 [physics.data-an].
- [9] International Organization for Standardization (ISO), *Guide to the expression of uncertainty in measurement*, Geneva, Switzerland, 1993.
- [10] [https://en.wikipedia.org/wiki/Skellam\\_distribution](https://en.wikipedia.org/wiki/Skellam_distribution).

- [11] R Core Team (2018), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.  
<https://www.R-project.org/> .
- [12] J.W. Lewis et al., *Package ‘skellam’*,  
<https://CRAN.R-project.org/package=skellam> .
- [13] G. D’Agostini, *Bayesian Reasoning in Data Analysis. A critical Introduction*, World Scientific, 2003.
- [14] G. D’Agostini, *Bayesian reasoning versus conventional statistics in High Energy Physics*, Proc. XVIII International Workshop on Maximum Entropy and Bayesian Methods, Garching (Germany), July 1998, V. Dose et al. eds., Kluwer Academic Publishers, Dordrecht, 1999, arXiv:physics/9811046 [physics.data-an].
- [15] G. D’Agostini, *The Waves and the Sigmas (To Say Nothing of the 750 GeV Mirage)*, arXiv:1609.01668 [physics.data-an].
- [16] D. Hume, *Enquiry concerning human understanding* (1748)  
LibriVox entry (Chapter 8: *Of probability*)
- [17] P. Astone and G. D’Agostini, *Inferring the intensity of Poisson processes at the limit of the detector sensitivity (with a case study on gravitational wave burst search)*, CERN-EP/99-126, arXiv:hep-ex/9909047 .
- [18] J. Pearl, *Causality*, Cambridge University Press, 2000.
- [19] Bayes, Thomas and Price, Richard *An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S.*, Philosophical Transactions of the Royal Society of London. 53: 370–418, (1763),  
<https://doi.org/10.1098/rstl.1763.0053> .
- [20] P.S. Laplace, *Mémoire sur la probabilité des causes par les événements*”, Mémoire de l’Académie royale des Sciences de Paris (Savants étrangers), Tome VI, p. 621, 1774, <https://gallica.bnf.fr/ark:/12148/bpt6k77596b/f32> .
- [21] H. Poincaré, *“Science and Hypothesis”*, 1905 (Dover Publications, 1952).
- [22] M.G. Kendall and A. Stuart, *The advanced theory of statistics*, 1943 (C. Griffin & Co., 1969).

- [23] G. D'Agostini and G. Degrassi, *Constraints on the Higgs boson mass from direct searches and precision measurements*, Eur. Phys. J. **C10** (1999) 633, <https://arxiv.org/abs/hep-ph/9902226> .
- [24] S. Gariazzo, *Constraining power of open likelihoods, made prior-independent*, arXiv:1910.06646 [astro-ph.CO] .
- [25] P.F. de Salas, D.V. Forero, S. Gariazzo, P. Martínez-Miravé, O. Mena, C.A. Ternes, M. Tórtola, J.W.F. Valle, *2020 Global reassessment of the neutrino oscillation picture*, arXiv:2006.11237 [hep-ph] .
- [26] G. Grilli di Cortona, A. Andrea and S. Piacentini, *Migdal effect and photon Bremsstrahlung: improving the sensitivity to light dark matter of liquid argon experiments*, arXiv:2006.02453 [hep-ph] .
- [27] G. D'Agostini, *Confidence limits: what is the problem? Is there the solution?*, Workshop on Confidence Limits, CERN, Geneva, 17-18 January 2000, arXiv:hep-ex/0002055.
- [28] C.F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientum*, Hamburg 1809, [https://archive.org/details/bub\\_gb\\_ORUOAAAAQAAJ](https://archive.org/details/bub_gb_ORUOAAAAQAAJ) .
- [29] G. D'Agostini, *Skeptical combination of experimental results using JAGS/rjags with application to the  $K^\pm$  mass determination*, arXiv:2001.03466 [physics.data-an]
- [30] [https://en.wikipedia.org/wiki/Conjugate\\_prior](https://en.wikipedia.org/wiki/Conjugate_prior) .
- [31] M. Bogner, *Probability distributions*, <https://play.google.com/store/apps/details?id=com.mbogner.probdist>, <https://apps.apple.com/us/app/probability-distributions/id889106396> .
- [32] [https://en.wikipedia.org/wiki/Gamma\\_distribution](https://en.wikipedia.org/wiki/Gamma_distribution) .
- [33] [https://en.wikipedia.org/wiki/Beta\\_function](https://en.wikipedia.org/wiki/Beta_function) .
- [34] [https://en.wikipedia.org/wiki/Beta\\_prime\\_distribution](https://en.wikipedia.org/wiki/Beta_prime_distribution) .
- [35] Th. Cathcart and D. Klein, *Plato and a Platypus walk into a bar...: understanding Philosophy through jokes*, Penguin Group, 2008.

- [36] M. Plummer, *JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling*, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria. ISSN 1609-395X, <http://mcmc-jags.sourceforge.net/> .
- [37] M. Plummer, *rjags: Bayesian Graphical Models using MCMC*. R package version 4-10, <https://CRAN.R-project.org/package=rjags> .
- [38] C. Andrieu et al., *An introduction to MCMC for Machine Learning*, Machine Learning **50** 5-43 (2003), <https://doi.org/10.1023/A:1020281327116> .
- [39] D. Lunn et al., *The BUGS project: Evolution, critique and future directions*, Statistics in Medicine **28** 3049-3067 (2008), <https://doi.org/10.1002/sim.3680> .
- [40] The BUGS Project, <http://www.mrc-bsu.cam.ac.uk/software/bugs/> .
- [41] <http://www.openbugs.net/w/Examples> .
- [42] J.O. Berger and D.A. Berry, *Statistical analysis and the illusion of objectivity*, Am. Scientist **76** (1988) 159.
- [43] G. D’Agostini, *Fits, and especially linear fits, with errors on both axes, extra variance of the data points and other complications*, arXiv:physics/0511182 [physics.data-an] .
- [44] G. D’Agostini, *On a curious bias arising when the  $\chi^2/\nu$  scaling prescription is first applied to a sub-sample of the individual results*, arXiv:2001.07562 [physics.data-an] .
- [45] A. Caldwell et al., *BAT: The Bayesian Analysis Toolkit*, Comput. Phys. Comm. **180** (2009) 2197-2209; J.Phys.Conf.Ser. **219** (2010) 032013; J.Phys.Conf.Ser. **331** (2011) 072040; <https://bat.mpp.mpg.de/> .
- [46] J. Bezanson, A. Edelman, S. Karpinski, V.B. Shah, *Julia: A Fresh Approach to Numerical Computing*, arXiv:1411.1607 [cs.MS]
- [47] O. Schulz et al. *BAT.jl – A Julia-based tool for Bayesian inference*, arXiv:2008.03132 [stat.CO].

# Appendix A – From the Bernoulli process to the Poisson process: binomial, Poisson and exponential distributions (and more)

## A1. Reminder of basic formulae

Let us start reminding the well known binomial and Poisson distributions, taken verbatim from Ref. [13], just to introduce the notation used in this note.

### Binomial distribution

$X \sim \text{Binom}(n, p)$  (hereafter “ $\sim$ ” stands for “follows”);  $\text{Binom}(n, p)$  stands for *binomial* with parameters  $n$  and  $p$ :

$$f(x | n, p) = \frac{n!}{(n-x)!x!} \cdot p^x \cdot (1-p)^{n-x}, \quad \begin{cases} n = 1, 2, \dots, \infty \\ 0 \leq p \leq 1 \\ x = 0, 1, \dots, n \end{cases} .$$

Expected value, standard deviation and *variation coefficient* [ $v \equiv \sigma(X)/E(X)$ ]:

$$\begin{aligned} E(X) &= n \cdot p \\ \sigma(X) &= \sqrt{n \cdot p \cdot (1-p)} \\ v &= \frac{\sqrt{n \cdot p \cdot (1-p)}}{n \cdot p} \propto \frac{1}{\sqrt{n}} . \end{aligned}$$

### Poisson distribution

$X \sim \text{Poisson}(\lambda)$ :

$$f(x | \lambda) = \frac{\lambda^x}{x!} \cdot e^{-\lambda} \quad \begin{cases} 0 < \lambda < \infty \\ x = 0, 1, \dots, \infty \end{cases} .$$

Expected value, standard deviation and variation coefficient

$$\begin{aligned} E(X) &= \lambda \\ \sigma(X) &= \sqrt{\lambda} \\ v &= 1/\sqrt{\lambda} . \end{aligned}$$

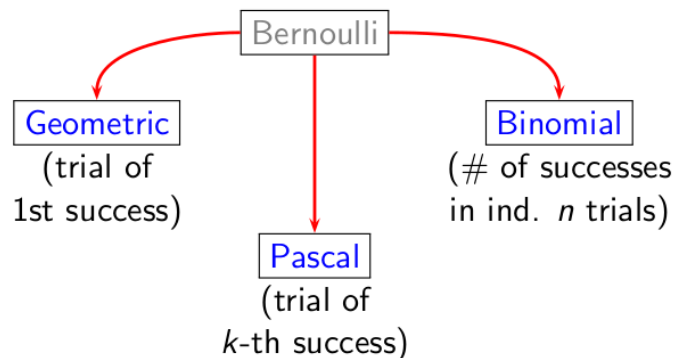
### Binomial $\rightarrow$ Poisson

$$\text{Binom}(n, p) \xrightarrow[\substack{n \rightarrow \infty \\ p \rightarrow 0 \\ (n \cdot p = \lambda)}]{} \text{Poisson}(\lambda) .$$

## A2. Bernoulli process and related distributions

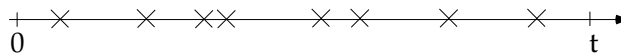
A Bernoulli process is characterized by a probability  $p$  of *success*, to which is associated the *uncertain number*  $X = 1$ , and probability  $1 - p$  of *failure*, to which is associated the uncertain number  $X = 0$ . Therefore, technically, a *Bernoulli distribution* is just a binomial with  $n = 1$ . But conceptually it is very important, because it is the basic process from which other distributions arise:

- a *binomial distribution* describes the probability of the total number of successes in  $n$  *independent* Bernoulli trials ‘having’ (or more precisely ‘believed to have’) the same probability of success  $p$ ;
- a *geometric distribution* describes the probability (again assuming independence and constant  $p$ ) of the trial *at which*<sup>36</sup> the first success occurs;
- a *Pascal distribution* (or *negative binomial distribution*) concerns finally the trial at which the  $k$ -th success occurs.<sup>37</sup>



## A3. Poisson process

Let us now imagine phenomena that might happen at random at a given instant<sup>38</sup>



<sup>36</sup>Indeed, it can also be found in the literature as the probability of *the number of failures before the first success occurs* (for example, my preferred *vademecum* of Probability Distributions, that is the homonymous *app* [31], reports both distributions).

<sup>37</sup>Also of this distribution there are two flavors, the other one describing the number of trials *before* the  $k$ -th success [31].

<sup>38</sup>One could also think at ‘things’ occurring in ‘points’ in some different space. All what we are going to say in the domain of time can be translated in other domains.

such that

- the probability of one count in  $\Delta T$  is proportional to  $\Delta T$ , with  $\Delta T$  ‘small’, that is

$$p = P(\text{“1 count in } \Delta T\text{”}) = r \Delta T$$

where the proportionality factor  $r$  is interpreted as the *intensity of the process*;

- the probability that two or more counts occur in  $\Delta T$  is much smaller than the probability of one count (the condition holds if  $\Delta T$  is small enough, that will be the case of interest):

$$P(\geq 2 \text{ counts}) \ll P(1 \text{ count});$$

- what happens in one interval does not depend on what happened (or ‘will happen’) in other intervals (if disjoint).

Let us divide a finite time interval  $T$  in  $n$  small intervals, i.e. such that  $T = n \Delta T$ . Considering the possible occurrence of a count in each small interval  $\Delta T$  as an independent Bernoulli trial, of probability

$$p = r \Delta T = r \cdot \frac{T}{n},$$

if we are interested in the total number of counts in  $T$  we get a binomial distribution, that is, indicating by  $X$  the uncertain number of interest,

$$X \sim \text{Binom}(n, p).$$

But when  $n$  is ‘very large’ ( $n \rightarrow \infty$ ) we obtain a Poisson distribution with

$$\lambda = n \cdot \left( r \cdot \frac{T}{n} \right) = r \cdot T,$$

equal to the intensity of the process times the finite time of observation. In particular, we can see that the physical quantity of interest is  $r$ , while the Poisson parameter  $\lambda$  is a kind of ancillary quantity, depending on the measurement time.

## A4 – Waiting time to observe the $k$ -event

It is clear that if we are interested in the probability that the first count occurs in the  $i$ -th time interval of amplitude  $\Delta T$ , we recover ‘in principle’ a geometric distribution. But since  $\Delta T$  can be arbitrary small, it makes no sense in numbering the intervals.



Nevertheless, thinking in terms of the  $n$  Bernoulli process can be again very useful. Indeed, the probability that the first count occurs *after* the  $x - th$  trial is equal to the probability that it never occurred in the trials from 1 to  $x$ :

$$P(X > x) = (1 - p)^x.$$

In the domain of time, indicating now by  $T$  the time at which the first event can occur, the probability that this variable is larger than the value  $t$ , the latter being  $n$  times  $\Delta T$ , is given by

$$\begin{aligned} P(T > t) &= (1 - p)^n \\ &= \left(1 - r \cdot \frac{t}{n}\right)^n \xrightarrow{n \rightarrow \infty} e^{-rt}. \end{aligned}$$

As a complement, the cumulative distribution of  $T$ , from which the probability density function follows, is given by

$$F(t|r) \equiv P(T \leq t) = 1 - P(T > t) = 1 - e^{-rt}$$

$$f(t|r) \equiv \frac{dF(t|r)}{dt} = r e^{-rt}.$$

The time at which the first count is recorded is then described by an exponential distribution having expected value, standard deviation and variation coefficient equal to

$$\begin{aligned} \mathbf{E}(T) &= 1/r \quad [\equiv \tau] \\ \sigma(T) &= 1/r = \tau \\ v &= 1, \end{aligned}$$

while the *mode* ('most probable value') is always at  $T = 0$ , independently of  $r$ .

As we can see, as it is reasonable to be, the higher is the intensity of the process, the smaller is the expected time at which the first count occurs (but note that the distribution extends always rather slowly to  $T \rightarrow \infty$ , a mathematical property reflecting the fact that such a distribution has always a 100% *standard uncertainty*, that is  $v = 1$ ). Moreover, since the choice of the instant at which we start waiting from the first event is arbitrary (this is related to the so called 'property of no memory' of the exponential distribution, which has an equivalent in the geometric one), we can choose it to be the instant at which a previous count occurred. Therefore, the same distribution describes the time intervals between the occurrence of subsequent counts.

Once we have got the probability distribution of  $k = 1$ , using probability rules we can get that of  $k = 2$ , reasoning on the fact that the associated variable is the sum of two exponentials, and so on. We shall not enter into details,<sup>39</sup> but only say that we end with the *Erlang distribution*, given by

$$f(t|r, k) = \frac{r^k}{(k-1)!} \cdot t^{k-1} \cdot e^{-rt} \quad \begin{cases} r > 0 \\ k : \text{integer}, \geq 1 \end{cases}$$

The extension of  $k$  to the continuum, indicated for clarity as  $c$ , leads to the famous *Gamma distribution* (here written for our variable  $t$ )

$$f(t|r, c) = \frac{r^c}{\Gamma(c)} \cdot t^{c-1} \cdot e^{-rt} \quad \begin{cases} r > 0 \\ c > 0 \end{cases}$$

with  $r$  the ‘rate parameter’ (and it is now clear the reason for the name) and  $c$  the ‘shape parameter’ (the special cases in which  $c$  is integer help to understand its meaning), having expected value and standard deviation equal to  $c/r$  and  $\sqrt{c}/r$ , both having the dimensions of time (this observation helps to remember their expression).

However, since in the text the symbol  $r$  is assigned to the intensity of the physical process of interest, we are going to use for the Gamma distribution the standard symbols met in the literature (see e.g. [31] and [32]) applying the following replacements:

$$\begin{aligned} c &\rightarrow \alpha \\ r &\rightarrow \beta. \end{aligned}$$

Using also the usual symbol  $X$  for generic variable, here is a summary of the most important expressions related to the Gamma distribution (we also add the mode, easily obtained by the condition of maximum<sup>40</sup>):

---

<sup>39</sup>It is indeed a useful exercise to derive the Erlang distribution starting from

$$f(t|r, k=2) = \int_0^\infty \int_0^\infty \delta(t - t_1 - t_2) \cdot f(t_1|r, k=1) \cdot f(t_2|r, k=1) dt_1 dt_2,$$

and going on until the general rule is obtained.

<sup>40</sup>Taking the log of  $f(x|\alpha, \beta)$ , we get the condition of maximum by

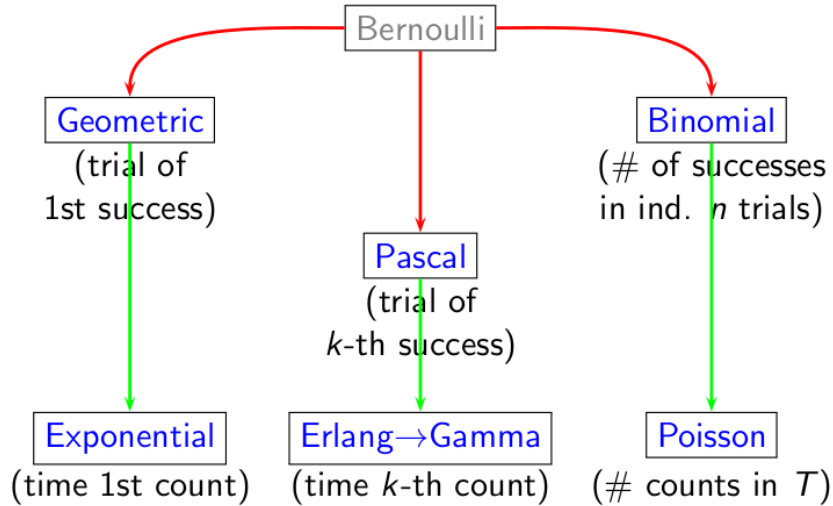
$$\frac{\partial}{\partial x} \log f(x|\alpha, \beta) = \frac{\alpha - 1}{x} - \beta = 0,$$

resulting in  $x = (\alpha - 1)/\beta$ .

$X \sim \text{Gamma}(\alpha, \beta)$ :

$$\begin{aligned}
 f(x | \alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} \cdot e^{-\beta x} && \begin{cases} \alpha > 0 \\ \beta > 0 \end{cases} \\
 \mathbb{E}(X) &= \frac{\alpha}{\beta} \\
 \text{Var}(X) &= \frac{\alpha}{\beta^2} \\
 \sigma(X) &= \frac{\sqrt{\alpha}}{\beta} \\
 \text{mode}(X) &= \begin{cases} 0 & \text{if } \alpha < 1 \\ \frac{\alpha-1}{\beta} & \text{if } \alpha \geq 1 \end{cases}
 \end{aligned}$$

Here is, finally, a summary of the distributions derived from the ‘apparently insignificant’ Bernoulli process:



For completeness, let us also remind that:

- the famous  $\chi^2$  *distribution* is technically a Gamma, with  $\alpha = \nu/2$  and  $\beta = 1/2$ ;
- most distributions appearing in this scheme, with the obvious exception of the geometric and the exponential, which have fixed shape, ‘tend to a Gaussian distribution’ for some values of the parameters. In particular, for what concerns this paper, the Poisson distribution tends to ‘normality’ for ‘large’ values of  $\lambda$ , as well known. However, it is perhaps worth remembering that, in general, such a limit applies to the cumulative distribution, and not to the probability function, defined for the Poisson distribution only for non negative integers:

$$F(x | \text{Poisson}(\lambda)) \xrightarrow{\text{‘}\lambda \rightarrow \infty\text{’}} F(x | \mathcal{N}(\lambda, \sqrt{\lambda})).$$

## Appendix B – R and JAGS code

### B.1 – Distribution of the difference of Poisson distributed counts

```
dPoisDiff <- function(d, lambda1, lambda2) {  
  xmax = round(max(lambda1,lambda2)) + 20*sqrt(max(lambda1,lambda2))  
  sum( dpois((0+d):xmax, lambda1) * dpois(0:(xmax-d), lambda2) )  
}
```

```
l1 = 1  
l2 = 1  
d = -8:8  
fd = rep(0,length(d))  
for(i in 1:length(d)) fd[i] = dPoisDiff(d[i], l1, l2)  
E.d <- sum(d*fd)  
E.d2 <- sum(d^2*fd)  
sigma.d <- sqrt(E.d2 - E.d^2)  
cat(sprintf(" d: %.3f +- %.3f ", E.d, sigma.d))  
cat(sprintf(" (exact: %.3f +- %.3f)\n", l1-l2, sqrt(l1+l2)))  
  
barplot(fd, names=d, col='cyan', xlab='d', ylab='f(d)')
```

(The function `dPoisDiff()` is simple implementation of the reasoning shown in the text. For a more professional function see footnote 5.)

### B.2 – Monte Carlo estimate of the pdf of $\rho = \lambda_1/\lambda_2$ (flat priors on $\lambda_1$ and $\lambda_2$ )

```
n=10^7  
x1 = 1  
x2 = 1  
lambda1 = rgamma(n, x1+1, 1)  
lambda2 = rgamma(n, x2+1, 1)  
rho = lambda1/lambda2  
E.rho = mean(rho)  
sigma.rho = sd(rho)  
  
max.rho.hist = 8  
rho = rho[rho<max.rho.hist]  
hist(rho, nc=150, col='cyan', freq=FALSE, xlim=c(0,max.rho.hist), main='',  
      xlab=expression(paste(rho, ' = ', lambda[1], '/', lambda[2]))) )
```

```

# dummy histogram for rough evaluation of the mode
h.rx <- hist(rho, nc=1000, plot=FALSE )
mode = h.rx$mids[which.max(h.rx$density)]

abline(v=mode, col='red')
abline(v=E.rho, col='blue')

p.overflow = (n - length(rho))/n * 100
cat(sprintf("fraction of overflows %.2f%%\n", p.overflow))
text(6,0.63-0.05,expression(paste(x[1], ' = ', x[2], ' = 1')),
     cex=1.6,col='blue')
text(6,0.54-0.05,sprintf("mean = %.2f;  std = %.2f", E.rho, sigma.rho),
     cex=1.5, col='blue')
text(6,0.46-0.05, sprintf("[ Overflow: %.2f%% ]", p.overflow),
     cex=1.5, col='gray')
text(6,0.37-0.05, sprintf("[ Mode = %.2f ]", mode), cex=1.5, col='red')

```

### B.3 – Ratio of rates: exact evaluations vs simulation

```

mode.rho <- function(a1,b1, a2,b2) ifelse(a2 >-1, b2/b1* (a1-1)/(a2+1), Inf)
E.rho     <- function(a1,b1, a2,b2) ifelse(a2 > 1, b2/b1* a1 / (a2-1), Inf)
var.rho   <- function(a1,b1, a2,b2)  ifelse(a2 > 2,
      (b2/b1)^2 * ( a1 / (a2-1) * ((a1+1)/(a2-2) - a1/(a2-1))), Inf)
sigma.rho <- function(a1,b1, a2,b2)  ifelse(a2 > 2,
      sqrt(var.rho(a1,b1, a2,b2)), Inf)
f.rho <- function(rho, a1,b1, a2,b2) {
  lf = ( a1*log(b1) + a2*log(b2) + (a1-1)*log(rho)
        + (-a1-a2)*log(b2+rho*b1) - lbeta(a1,a2) )
  return(exp(lf))
}

x1 = 1; T1 = 1
x2 = 2; T2 = 2

a1 = x1+1; b1 = T1
a2 = x2+1; b2 = T2
rho.max = 8
n = 10^6

cat(sprintf("x1,T1 = %.2f, %.2f;  ", x1, T1 ))
cat(sprintf("x2,T2 = %.2f, %.2f;  \n", x2, T2 ))

```

```

cat(sprintf("alpha1,beta1 = %d, %d; ", a1, b1 ))
cat(sprintf("alpha2,beta2 = %d, %d \n", a2, b2 ))
Erho <- E.rho(a1,b1, a2,b2)
Srho <- sigma.rho(a1,b1, a2,b2)
cat(sprintf("mode = %.3f; E() = %.3f ; sigma %.3f\n",
            mode.rho(a1,b1, a2,b2),
            Erho, Srho ))

x1.r <- rgamma(n, a1, b1)
x2.r <- rgamma(n, a2, b2)
rho.r <- x1.r/x2.r
Mrho <- mean(rho.r)
SDrho <- sd(rho.r)
cat(sprintf("MC: mean = %.3f ; sigma %.3f\n", Mrho, SDrho ))

rho.r = rho.r[rho.r<rho.max] # only for histogram!!
                        # Warning!! It changes normalization!
norma = length(rho.r)/n
h <- hist(rho.r, nc=150, plot=FALSE)
h$density <- h$density * norma
h$counts <- h$counts * norma
plot(h, col='cyan', freq=FALSE, main='', xlim=c(0,rho.max),
      ylim=c(0,0.52),
      xlab=expression(rho), ylab=expression(paste('f(',rho,')')))

rho = seq(0, rho.max, len=101)
points(rho, f.rho(rho, a1,b1, a2,b2), ty='l', col='blue')
text(6,0.46,bquote(x[1] == .(x1) ~ ", " ~ T[1] == .(T1) ), cex=1.5, col='blue')
text(6,0.41,bquote(x[2] == .(x2) ~ ", " ~ T[2] == .(T2) ), cex=1.5, col='blue')
Erho.s <- round(Erho, 2)
Srho.s <- round(Srho, 2)
mode.s <- round(mode.rho(a1,b1, a2,b2),2)
Mrho.s <- round(Mrho,2)
SDrho.s <- round(SDrho,2)
text(6,0.35,bquote(E(rho) == .(Erho.s) ~ ", " ~
                  sigma(rho) == .(Srho.s) ), cex=1.5, col='blue')
text(6,0.28,bquote(mode(rho) == .(mode.s) ), cex=1.5, col='red')
text(6,0.21,bquote("mean" == .(Mrho.s) ~ ", " ~
                  "std" == .(SDrho.s) ), cex=1.5, col='blue')
abline(v=Erho, col='blue')
abline(v=mode.rho(a1,b1, a2,b2), col='red',)

```

## B.4 – Distribution of $\rho$ implied by uniform priors on $r_1$ and $r_2$

```
n = 10^7
rM = 100
r1 = runif(n, 0, rM)
r2 = runif(n, 0, rM)
rho = r1/r2
rho.h <- rho[rho<5]      # for the histogram
norma = length(rho.h)/n  # normalization
h <- hist(rho.h, nc=100, plot=FALSE)
h$density <- h$density * norma
h$counts <- h$counts * norma
plot(h, col='cyan', freq=FALSE, main='', xlim=c(0,5),
      ylim=c(0,0.52),
      xlab=expression(rho), ylab=expression(paste('f(',rho,')')))
abline(v=1, col='red')

# r1.check = rho*r2
# hist(r1.check, nc=200, col='blue', freq=FALSE, xlim=c(0,rM))
```

## B.5 – Example of JAGS inference of rates and their ratio for the model of Fig. 10

```
#----- Data -----
x1 = 3; T1=3
x2 = 6; T2=6
nr = 1e5

#----- JAGS model -----
library(rjags)
model = "tmp_model.bug"  # name of the model file ('temporary')
write("
model {
  x1 ~ dpois(lambda1)
  x2 ~ dpois(lambda2)
  lambda1 <- r1 * T1
  lambda2 <- r2 * T2
  r1 ~ dgamma(1, 1e-6)
  r2 ~ dgamma(1, 1e-6)
  rho <- r1/r2
}
", model)
```

```

#----- JAGS call via rjags -----
data <- list(x1=x1, T1=T1, x2=x2, T2=T2)
jm <- jags.model(model, data)
update(jm, 100)
to.monitor <- c('r1', 'r2', 'rho')
chain <- coda.samples(jm, to.monitor, n.iter=nr)

#----- Results -----
print(summary(chain))
plot(chain, col='blue')

cat(sprintf("Exact: \n"))
cat(sprintf("  r1 = %.3f +- %.3f\n", (x1+1)/T1, sqrt(x1+1)/T1))
cat(sprintf("  r2 = %.3f +- %.3f\n", (x2+1)/T2, sqrt(x2+1)/T2))
mu.rho <- ((x1+1)/T1)/(x2/T2)
sigma.rho <- sqrt(mu.rho*(T2/T1*(x1+2)/(x2-1)-mu.rho))
cat(sprintf("  rho = %.3f +- %.3f\n", mu.rho, sigma.rho))

```

## B.6 – Example of JAGS inference of rates and their ratio for the model of Fig. 13

```

#----- Data -----
x1 = 3; T1=3
x2 = 6; T2=6
nr = 1e5

#----- JAGS model -----
library(rjags)
model = "tmp_model.bug" # name of the model file ('temporary')
write("
model {
  x1 ~ dpois(lambda1)
  x2 ~ dpois(lambda2)
  lambda1 <- r1 * T1
  lambda2 <- r2 * T2
  r1 <- rho * r2
  r2 ~ dgamma(1, 1e-6)
  rho ~ dgamma(1, 1e-6)
}
", model)

```



```

#----- JAGS call via rjags -----
data <- list(x1=x1, T1=T1, x2=x2, T2=T2)
jm <- jags.model(model, data)
update(jm, 100)
to.monitor <- c('r1', 'r2', 'rho')
chain <- coda.samples(jm, to.monitor, n.iter=nr)

#----- Results -----
print(summary(chain))
plot(chain, col='blue')

cat(sprintf("Exact: \n"))
cat(sprintf("  r1 = %.3f +- %.3f\n", (x1+1)/T1, sqrt(x1+1)/T1))
cat(sprintf("  r2 = %.3f +- %.3f\n", x2/T2, sqrt(x2)/T2) )
mu.rho <- ((x1+1)/T1)/((x2-1)/T2)
sigma.rho <- sqrt(mu.rho*(T2/T1*(x1+2)/(x2-2)-mu.rho))
cat(sprintf("  rho = %.3f +- %.3f\n", mu.rho, sigma.rho))

```

