American Journal of Engineering and Applied Sciences

Original Research Paper

Bayesian Statistical Inference for Number Counting Experiments

Diego Casadei

School of Engineering, University of Applied Sciences and Arts Northwestern Switzerland, Windisch, Switzerland

Article history
Received: 14-10-2015
Revised: 15-10-2015
Accepted: 13-11-2015

Email: diego.casadei@fhnw.ch

Abstract: Statistical inference describes how to infer about the true but unknown population from the measured sample and is a fundamental ingredient in scientific data analysis. Often one knows the probability model and wishes to estimate its parameters. The Bayesian approach provides a solution in terms of the posterior probability density function of the parameters of interest, given the model, the experimental result and our prior knowledge about the parameters. Number counting experiments are very often performed, in the assumption that the order in which results appear does not matter. Examples are the binomial model, arising when one investigates about the efficiency of a given selection process, and the Poisson model that describes how often a given outcome may show up. Here we provide analytic solutions for the Bayesian inference for both models, in case some or no prior information is available.

Keywords: Statistics, Data Analysis, Binomial, Poisson, Bayes

Introduction

Statistical inference describes how to infer about the true but unknown population from the measured sample and is a fundamental ingredient in scientific data analysis. Often one knows the probability model and wishes to estimate its parameters. We consider here a couple of very important discrete probability models, the binomial and Poisson distributions. The model expresses the probability of each outcome, given the values of each parameter. When considered as a function of the parameters for a fixed outcome, the model is no more a probability distribution (as it is no more normalized to one) and is called the likelihood function. Many common methods address statistical inference by maximizing the likelihood function. This works well when the sample is large, because asymptotically the likelihood behaves as a (multi-dimensional) Gaussian distribution peaked at the true value of the parameters. On the other hand, often only small samples are realistic, for example because repeating the experiment is too expensive or takes too long time. But for small sample one can not rely on asymptotics. In these circumstances, the data do not “speak by themselves” and one has to face the problem posed by a small sample size.

The Bayesian approach is very useful in this case, because it makes it explicit how the solution depends on the information available prior to performing the experiment, as detailed in the classical book by Bernardo and Smith (1994). This approach provides a solution in terms of the posterior probability density function of the parameters of interest, given the model, the experimental result and our prior knowledge about the parameters. The posterior is obtained by multiplying the likelihood function of the model parameters by their prior probability densities, encoding for example our best guess of their values and the uncertainty about such values. Furthermore, as not all parameters are interesting for the user, in the Bayesian approach one integrates over all other (nuisance) parameters, in order to obtain a lower-dimensional (marginal) probability density. Sometimes, the need for multidimensional integration is a formidable obstacle to address. On the other hand, here we focus on problems for which the analytic solution is known for the (marginal) posterior.

Number counting experiments are very commonly performed in science and industrial production, in the assumption that the order in which results appear does not matter. A common example is counting how many products, among a sample with fixed size, show some defect. This is described by the binomial model, arising when one investigates about the efficiency of a given selection process. Another way of checking for defects is to count how many products are not acceptable in a

certain amount of time. In this case one has the Poisson model, describing how often a given outcome may show up, in the assumption that each occurrence is independent from the others.

Here we provide analytic solutions for the Bayesian inference for binomial and Poisson models, in case some or no prior information is available about the parameter of interest. The analytic solution is easy to find and only requires to encode the prior information into a density belonging to the class of conjugate priors to the considered model. In this case, the posterior also belongs to the same class and the posterior parameters are simple functions of the initial parameter values and of the measured data. With large samples, the posterior peak (i.e., the posterior mode) coincides with the maximum-likelihood solution. However, with small samples the two methods give different estimators.

The Binomial Model

Let's assume that we examine a set of n products and select k defective items among them. Our goal is to estimate the probability ε that the production process is imperfect. Alternatively, we can interpret ε as the selection efficiency. The probability model is given by the binomial distribution:

$$P(k \mid \varepsilon, n) = \binom{n}{k} \varepsilon^k (1-\varepsilon)^{n-k} \tag{1}$$

When both k and n are large and the observed frequency $f = k/n$ is far from zero and one, the efficiency is well estimated by the frequency, which is also the Maximum Likelihood Estimator (MLE): $\varepsilon \simeq k/n$. In this case, the variance of the MLE is given by the Cramér-Rao lower bound: $V = \varepsilon(1-\varepsilon)/n$. However ε is unknown, hence one needs to replace it with its estimator f, obtaining $V = k(n-k)/n^3$. The square root of this expression provides the very widely (ab)used value for the standard deviation of ε. One clear problem with this expression for the variance is that, when $k = 0$ or $k = n$, one gets the same dispersion (zero) independently from the sample size. However, one intuitively expects that counting 0 defective items out of 10 products is not as "precise" as counting 0 defective items out of 100 products. Anyway, this is not the only limitation of this approximation.

We emphasize again that using the MLE and the approximate value of the Cramér-Rao lower bound to represent the efficiency and its variance is good *only* if the conditions stated above hold: Both k and n must be large and f must be far from zero and one, in the sense that f has to be several standard deviations away from the boundaries. The reason is that such approximation holds when the likelihood function, that is Equation 1 when considered a function of ε for given k and n (instead of the probability of k for given ε and n), is Gaussian. On the other hand, a Gaussian is defined over the entire real line, whereas the likelihood is defined for ε in the range [0,1]. Hence it is clear that the Gaussian approximation only works well if it is very narrow (small variance), compared to the distance of the peak to the nearest boundary.

When the probability ε is very small, which is hopefully the case when dealing with defective products, a precise measurement is hardly feasible, as it would require a very large sample size n. This may be a problem, for example when the inspection process prevents the item to be commercialized, or when the inspection itself is expensive. At the same time, if the sample is small we cannot pretend the data to "speak by themselves", such that our inference will depend also on whatever additional source of information is available beyond the experimental outcome. The smaller is n, the larger is the importance of such auxiliary information. For example, at the limit of $n = 0$ the latter completely dominates. For this reason, it is fundamental to work within a framework that allows for a coherent treatment of all sources of information, including the experimental outcome and any prior knowledge, for example coming from a previous test or analysis.

This motivates our choice of the Bayesian paradigm, in which the posterior probability density $p(\varepsilon|k,n)$ is obtained from the Bayes' theorem as follows:

$$p(\varepsilon \mid k,n) \propto P(k \mid \varepsilon, n)\pi(\varepsilon) \tag{2}$$

where, $\pi(\varepsilon)$ is the prior probability density of ε, the likelihood function is given by Equation 1 and the proportionality sign means that the normalization constant (the inverse of the integral of the r.h.s.) is not explicitly given by Equation 2. This is not a problem in general, because the posterior density can be normalized to one after having computed (possibly by numerical methods) the product on the r.h.s. of (2).

Numerical methods are not required, if the prior density is chosen among the family of conjugate priors to the binomial model. This is the class of Beta distributions:

$$\mathrm{Be}(x \mid a,b) = x^{a-1}(1-x)^{b-1} / B(a,b) \tag{3}$$

where, the normalization factor is given by Euler's Beta function:

$$B(a,b) = \Gamma(a)\Gamma(b) / \Gamma(a+b) \tag{4}$$

When the prior density belongs to the family of conjugate priors, also the posterior belongs to the same

family. For the binomial model, by choosing $\pi(\varepsilon) = \mathrm{Be}(\varepsilon|a,b)$ the posterior resulting from Equation 2 is:

$$p(\varepsilon \mid k,n) = \mathrm{Be}(\varepsilon \mid k+a, n-k+b) \tag{5}$$

In particular, the posterior expectation is:

$$\mathrm{E}[\varepsilon] = (k+a)/(n+a+b) \tag{6}$$

and, when both $k+a > 1$ and $n-k+b > 1$, the posterior mode (i.e., the peak position) is:

$$\mathrm{m}[\varepsilon] = (k+a-1)/(n+a+b-2) \tag{7}$$

Finally, the posterior variance is:

$$\mathrm{V}[\varepsilon] = \frac{(k+a)(n-k+b)}{(n+a+b)^2(n+a+b+1)} \tag{8}$$

When the prior information about ε is summarized by its prior expectation E and variance V, the prior Beta parameters can be determined with the method of moments, by requiring that:

$$\begin{cases} E = a/(a+b) \\ V = \dfrac{ab}{(a+b)^2(a+b+1)} \end{cases} \tag{9}$$

and by solving for a and b.

When there is no prior information beyond the knowledge of the binomial model itself, one should use the reference prior $\pi(\varepsilon) = \mathrm{Be}(\varepsilon|0.5,0.5)$, obtaining the reference posterior $p(\varepsilon|k,n) = \mathrm{Be}(\varepsilon|k+0.5,n-k+0.5)$ as described by Casadei (2012a). The reference posterior expectation, mode and variance are then:

$$\mathrm{E}[\varepsilon] = (k+0.5)/(n+1) \tag{10}$$

$$\mathrm{m}[\varepsilon] = (k-0.5)/(n-1) \quad \text{(only for } k>0 \text{ and } n>k\text{)} \tag{11}$$

$$\mathrm{V}[\varepsilon] = \frac{(k+0.5)(n-k+0.5)}{(n+1)^2(n+2)} \tag{12}$$

It is easy to notice that $\mathrm{E}[\varepsilon]$ and $\mathrm{m}[\varepsilon]$ given by Equation 10 and 11 "bracket" the MLE estimator k/n of ε, with a difference that goes to zero for increasing n. If one has to report a single "best" value for the efficiency, it is recommended to provide $\mathrm{m}[\varepsilon]$, the most probable value (unless it is not defined, in which case one should report $\mathrm{E}[\varepsilon]$ and comment about the absence of a peak). On the other hand, if the efficiency is needed for further computation, where the uncertainty is going to be computed following the classical recipe for the "propagation of errors", then the best value is the expectation $\mathrm{E}[\varepsilon]$, together with the variance from Equation 12. One may notice that the latter is never zero and it decreases with increasing n, whatever is the value of k. Thus, it is more "natural" to use than the formula obtained in the Gaussian approximation, which is valid only in the asymptotic regime (in which the two expressions give the same numerical value). Additional details can be found in Casadei (2012a) and references therein.

The Poisson Model

Another way of investigating about production errors is to ask how often a defective item is produced. A measurement is performed by counting how many objects show problems in a given amount of time. Assuming that defects are not statistically correlated, the probability of counting m items when μ is the unknown expectation is given by the Poisson distribution:

$$P(m \mid \mu) = \mathrm{e}^{-\mu}\mu^{-m}/m! \tag{13}$$

As $\mathrm{E}[m] = \mathrm{V}[m] = \mu$, the most natural estimate of the unknown parameter μ is given by the observed number of events and its square root is typically used as its standard deviation. Indeed, this is the most common practice. However, there are subtleties here that should not be overlooked. The Poisson distribution is asymmetric, unless the value of the parameter μ is large, when it can be well approximated by a Gaussian distribution (already good when μ>30). However the latter is a continuous distribution defined over the whole real line, whereas the former is a discrete distribution defined on all non-negative integers. Even when Equation 13 is considered as a function of μ at fixed m, i.e., when it expresses the likelihood function, it is significantly asymmetric unless m is large. Indeed, μ is a non-negative real number, hence a Gaussian provides a good approximation only when its peak is several standard deviations away from zero. This condition is never met for low counts, hence again we need a paradigm in which the impact of prior information is explicitly taken into account.

The Bayesian solution to the inference problem about μ is represented by the posterior density:

$$p(\mu \mid m) \propto P(m \mid \mu)\pi(\mu) \tag{14}$$

where, $\pi(\mu)$ is the prior probability density of μ and the likelihood function is given by Equation 13. Once again, we have left the normalization constant unspecified, as it can be computed once the product on the r.h.s. of Equation 14 is known, by requiring that the integral over the positive real line is equal to one.

By choosing a conjugate prior for μ, we obtain the posterior easily, without the need for numerical integration. For the Poisson model, the conjugate family contains all Gamma densities with shape parameter $a > 0$ and rate parameter $b > 0$:

$$\mathrm{Ga}(x \mid a,b) = b^a x^{a-1} e^{-bx} / \Gamma(a) \tag{15}$$

When the number of items m is known from the measurement and the prior for μ is $\mathrm{Ga}(\mu|a,b)$, the posterior in Equation 14 has the following explicit form:

$$p(\mu \mid m) = \mathrm{Ga}(\mu \mid m + a, 1 + b) \tag{16}$$

with posterior expectation and variance:

$$\mathrm{E}[\mu] = (m + a) / (1 + b) \tag{17}$$

$$\mathrm{V}[\mu] = (m + a) / (1 + b)^2 \tag{18}$$

The difference with the approximation mentioned earlier, $\mathrm{E}[m] = \mathrm{V}[m] = \mu$, becomes negligible for very large values of m, but it is clear that for small counts the impact of the prior information is important. Such information may come from another test or analysis.

In the simple but frequent case in which the prior knowledge about μ is summarized by its prior expectation E and variance V, the Gamma parameters are determined with the method of moments by imposing $E = a/b$ and $V = a/b^2$. This gives $b = E/V$ and $a = bE$. Alternatively, one could start from the prior most probable value [the Gamma mode is at the point $(a-1)/b$ for $a > 1$] and variance, or from the knowledge of intervals covering given prior probabilities (e.g., 68.3% or 95% probability; but this requires a numerical treatment to find a and b), or from any set of conditions which is sufficient to determine the shape and rate parameters.

When there is no prior information beyond the knowledge of the Poisson model itself, one should use the reference prior $\pi(\mu) = \mathrm{Ga}(\mu|0.5,0)$, obtaining the reference posterior $p(\mu|m) = \mathrm{Ga}(\mu|m+0.5,1)$. The reference posterior expectation, mode and variance are then:

$$\mathrm{E}[\mu] = (m + 0.5) \tag{19}$$

$$\mathrm{m}[\mu] = (m - 0.5) \text{ if } m \geq 1 \tag{20}$$

$$\mathrm{V}[\mu] = (m + 0.5) \tag{21}$$

The variance is a bit larger than the approximation discussed above, which is thus a bit optimistic. However, the difference in the uncertainty (the square root of the variance) is not very significant, apart for very small counts m. At the same time, the variance from Equation 21 is never zero, even when zero counts are observed. This makes sense, as counting no event brings real information and is not equivalent to performing no experiment. The approximated variance instead is zero if $m = 0$, implicitly stating that there is no expected rate (as $\mathrm{E}[\mu] = 0$ in this approximation) with perfect certainty, which clearly makes no sense. On the other hand, the small bias of the reference posterior expectation given by Equation 19 for the measurement $m = 0$ gives an estimate of $\mu = 0.5$ with standard deviation 0.7 (please note that writing $\mu = 0.5 \pm 0.7$ makes no sense here, as the Poisson distribution is very asymmetric for small μ and no negative value is allowed for μ), i.e., it points to a non-null expected rate even if no counts are observed, although with a very large uncertainty.

Sometimes, the identification of defective objects is a procedure with some non-zero false-positive rate. Assuming no correlation between detection of false-positive and truly defective items, the model is a Poisson process with parameter $(\mu + \nu)$ given by the sum of the expected number μ of truly defective items and the expected number ν of false-positives. The joint posterior for μ and ν is then Equation 22:

$$p(\mu,\nu \mid n) \propto s^{-\mu-\nu} \frac{(\mu+\nu)^n}{n!} \pi(\mu)\pi(\nu) \tag{22}$$

where, $\pi(\mu)$ and $\pi(\nu)$ are the prior densities.

As we are interested only into μ, we shall integrate over the nuisance parameter ν. This can be done on the r.h.s. of Equation 22 leaving $\pi(\mu)$ out of the integral. This reduces the inference problem to a 1-dimensional problem with (marginal) posterior:

$$p(\mu \mid n) \propto \pi(\mu) \int s^{-\mu-\nu} \frac{(\mu+\nu)^n}{n!} \pi(\nu)\, \mathrm{d}\nu \tag{23}$$

where, the integral defines the marginal model $p(n|\mu)$. Assuming some prior knowledge of the expected number ν of false-positives, encoded into a Gamma prior with the form $\pi(\mu) = \mathrm{Ga}(\mu|a,b)$, the integral in (23) can be computed analytically as shown by Casadei (2012b). The result is Equation 24:

$$p(n \mid \mu) = \left(\frac{b}{1+b}\right)^a e^{-\mu} f(\mu; n, a, b) \tag{24}$$

where the polynomial:

$$\begin{aligned} f(\mu; n, a, b) &= \sum_{k=0}^{n} \binom{a+k-1}{k} \frac{\mu^{n-k}}{(n-k)!(1+b)^k} \\ &= x^n [(b+1)x]^a U(a, a+n+1, (b+1)x) / n! \end{aligned} \tag{25}$$

behaves as $(\mu+a/b)^n$ when both a and b are very large while their ratio a/b remains finite. The last equality is obtained by Wolfram's *Mathematica* version 10.1 and features the confluent hypergeometric function of the second kind $U(a,b;z)$.

If there is prior information about μ encoded into a Gamma density of the form $\pi(\mu) = \mathrm{Ga}(\mu|A,B)$, the marginal posterior in Equation 23 can be computed analytically and has the form of a weighted average of $n+1$ Gamma densities. Its explicit form is provided in appendix B of Casadei (2014).

When there is no prior information about μ beyond the knowledge of the probability model, the use of the reference prior for μ computed by Casadei (2012b) is encouraged. Although also in this case the analytic solution for the reference (marginal) posterior is known and at least one implementation is freely available in the Bayesian Analysis Toolkit (BAT; Caldwell *et al*., 2009), the solution involves some numerical computation because it is not expressed in closed form in terms of known functions. On the other hand, Casadei (2014) has shown that an approximate reference prior can be used, which differs only a bit from the full reference prior and is very quick to compute. It corresponds to the limit of perfect prior knowledge about the nuisance parameter ν and has the form:

$$\pi_0(\mu) = \sqrt{\frac{\nu_0}{\mu+\nu_0}} \tag{26}$$

where ν_0 is the prior expectation for ν. In this case, the approximated reference posterior is a truncated Gamma density:

$$p_0(\mu \mid n) = \frac{1}{C}\mathrm{Ga}(\mu+\nu_0 \mid n+0.5,1) \tag{27}$$

and the normalization constant C is expressed in terms of the regularized Gamma function (Casadei, 2014):

$$C = 1 - \frac{\gamma(n+0.5,1)}{\Gamma(n+0.5)} \tag{28}$$

although C may be also computed numerically from Equation 27.

Discussion

We have addressed two models that have wide applicability. The binomial model describes all selection processes without memory, that is with the assumption that different occurrences of the desired phenomenon are statistically uncorrelated. It also applies to any binary classification scheme, with the same assumption. Thus, it can describe the result of a screening process of a production chain, as well as the selection of particular physical processes happening in particle collisions. The other model is described by the Poisson distribution, which describes the probability of counting a certain number of events in an experiment in which the expected yield is given. Once again, this model holds when there is no memory, i.e., when each occurrence happens randomly. A typical example is the number of radioactive decays in a given observation time, when the decay rate is known. Here we considered an example in which the number of defective products is counted in a certain amount of time.

A parametric probability model like the binomial or the Poisson distribution gives the probability of counting a certain number of events when the parameters are known. On the other hand, when performing an experiment one typically wants to infer about the values of the parameters, once the observed number of events is known. Often the asymptotic form of the likelihood function is exploited, in order to provide easy recipes to estimate the parameters. However, when the actual count of events is not very large, the departure of the likelihood from the Gaussian distribution becomes important. Furthermore, at very low (and possibly zero) counts such asymptotic approximation completely break apart, giving absurd results. We can overcome this problem by addressing statistical inference in the Bayesian framework, in which the auxiliary information available prior to performing the experiment is explicitly and quantitatively taken into account. The result of the inference is encoded into the posterior probability density of the parameters of interest, which is valid for any observed number of counts, even when observing no count at all.

Historically, although the Bayesian approach was the first to be proposed in the attempt to solve "inverse probability" (i.e., statistical inference) problems, the need for multidimensional integration prevented the systematic application of these methods. In practice, they have been replaced by the maximum-likelihood approach until recent times, because it requires to find (local) extrema of a multidimensional function, a much simpler computational problem than integration. Today we have powerful computers and sophisticated algorithms that allow for numerically solving complex multidimensional integrations. Despite from the availability of several software packages that are able to implement such numerical methods, it is still remarkable that analytic solutions exist for a number of widely applicable models. They require basically no CPU time in modern computers and can be implemented in different programming languages without much effort. This advantage also

characterizes the solutions that we have proposed for the binomial and Poisson models.

When speaking about Bayesian methods, one cannot avoid to discuss the choice of the prior distribution. The first thing to notice is that classical methods based on maximum-likelihood or profile likelihood approaches are actually not free from "priors": Although they are not called with this name, they still feature in such methods in the form of constraints on the model parameters that reflect our knowledge of their domain, "best" or "typical" value and uncertainty. Next, one should realize that the two most widely (ab)used priors, the (truncated) Gaussian and the flat distribution, are not always a good choice.

The Gaussian is typically taken as "the" informative prior, also when its domain clearly does not match the domain of the parameter of interest (think for a moment about the efficiency ε of a binomial model, defined in [0,1]). When this happen, it is clear that it has to be truncated on one or both sides. Strictly speaking, this makes its peak different from the mean and its "sigma" parameter different from its standard deviation (although almost nobody seems to pay attention to this fact in practice). When the Gaussian is a good approximation (which does happen in many cases), it still requires numerical methods to compute the posterior density. Given that the analytic solution is readily available in the models considered here, it seems absurd to adopt a Gaussian in place of the conjugate prior (Beta density for the binomial model; Gamma for Poisson), even when the latter closely mimics a Gaussian.

On the other hand, the flat prior is typically chosen as "non-informative" prior. The sole advantage of this choice is that the value of the posterior mode becomes identical to the maximum-likelihood estimator in this case (but their interpretation is different). However, one must be aware that, for the models considered here, the uniform prior is not well justified. For the binomial model, it is actually an informative prior (the non-informative choice being the reference prior). Anyway, as the flat distribution is $\mathrm{Be}(x|1,1)$, the Beta posterior computed with the uniform prior has parameters that differ only by half unit from those of the reference posterior. In many cases, the difference is small and can be neglected for practical purposes. On the other hand, the situation is very different for the Poisson model. As its parameter is defined over the positive real line, its domain is unbounded on the right and the flat prior is not normalizable. Thus, it can not represent an informative prior. At the same time, it is not the same as the reference prior, which is allowed to be an improper density, because it is only a mathematical tool that allows to obtain the reference posterior, defined as the posterior that maximizes the amount of missing prior information. This means that there is no mathematical justification for the use of a flat prior in the Poisson model: strictly speaking, it is forbidden because it is mathematically ill-defined.

Conclusion

In conclusion, for both binomial and Poisson models, if we encode the prior information into a density belonging to the conjugate family, we obtain the analytic form of the posterior density for the parameter of interest. This is true both in case of prior information about such parameter and in case of no additional information beyond the knowledge of the probability model. In the latter case, the reference posterior should be used (or perhaps a good approximation to it, when this simplifies the problem significantly, without noticeably affecting the result).

Ethics

This article is original and contains unpublished material. Author declares that are not ethical issues that may arise after the publication of this manuscript.

References

Bernardo, J.M. and A.F.M. Smith, 1994. Bayesian Theory. 1st Edn., Wiley, New York, ISBN-10: 0471924164, pp: 586.

Caldwell, A., D. Kollar and K. Kröninger, 2009. BAT-the Bayesian analysis toolkit. Comput. Phys. Commun., 180: 2197-2197. DOI: 10.1016/j.cpc.2009.06.026

Casadei, D., 2012a. Estimating the selection efficiency. J. Instrument., 7: P08021-P08021. DOI: 10.1088/1748-0221/7/08/P08021

Casadei, D., 2012b. Reference analysis of the signal + background model in counting experiments. J. Instrument., 7: P01012-P01012. DOI: 10.1088/1748-0221/7/01/P01012

Casadei, D., 2014. Reference analysis of the signal + background model in counting experiments II. Approximate reference prior. J. Instrument., 9: T10006-T10006. DOI: 10.1088/1748-0221/9/10/T10006