Solutions Chapter 10 - Evaluation of model fit and hypothesis testing

Alexandra Hotti

November 2019

10.1 WHO's reported novel disease outbreaks

Suppose that you are interested in modelling the number of outbreaks of novel diseases that the WHO reports each year. Since these outbreaks are of new diseases, you assume that you can model the outbreaks as independent events, and hence decide to use a Poisson likelihood; $X_t \sim Poisson(\lambda)$, where X_t is the number of outbreaks in year t, and λ is the mean number of outbreaks.

Problem 10.1.1

You decide to use a $\Gamma(3, 0.5)$ prior for the mean parameter (λ) of your Poisson likelihood (where a $\Gamma(\alpha, \beta)$ is defined to have a mean of α/β). Graph this prior.

Answer:

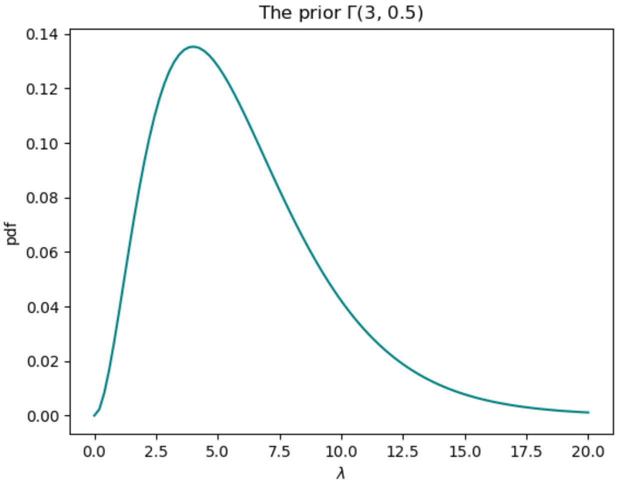

Figure 1: The prior distribution.

Problem 10.1.2

Suppose that the number of new outbreaks over the past 5 years is $X = (3, 7, 4, 10, 11)$. Using the conjugate prior rules for a Poisson distribution with a gamma prior, find the posterior and graph it.

Answer:

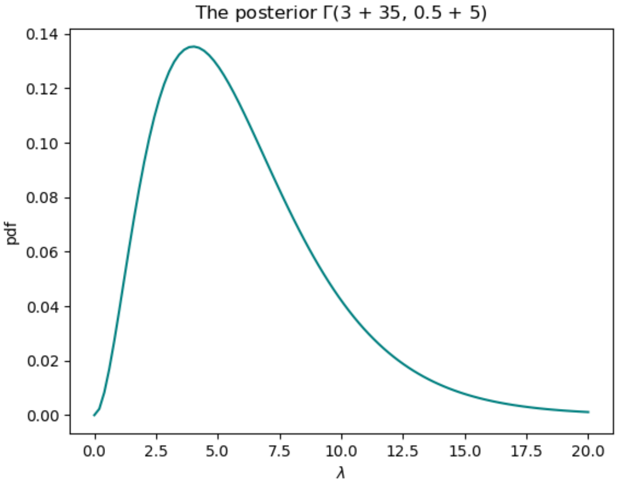

Figure 2: The posterior distribution.

Problem 10.1.3.

Generate 10,000 samples from the posterior predictive distribution, and graph the distribution.

Answer:

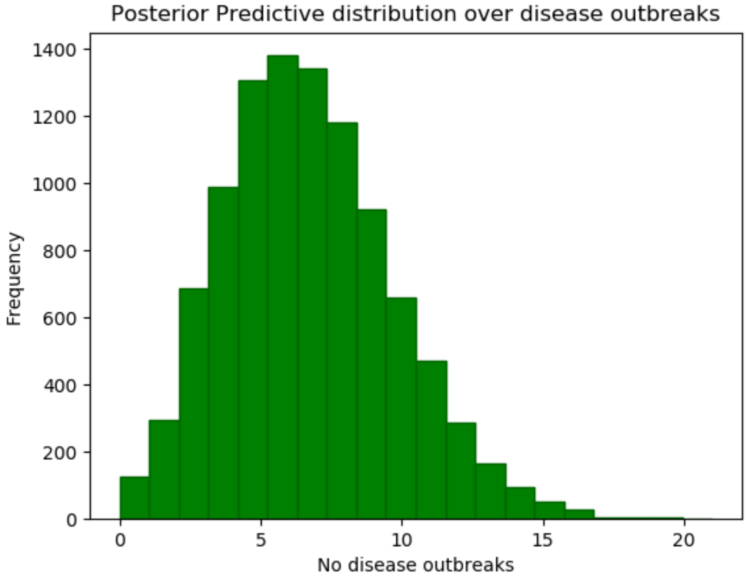

Figure 3: An estimate of the posterior predictive distribution.

Problem 10.1.4.

Compare the actual data with your 10,000 posterior predictive samples. Does your model fit the data?
Answer: Data maximum: 11
Data minimum: 3

The most extreme points of the data are the years with 3 and 11 disease outbreaks. By comparing the data wit the posterior predictive distribution I get the following PPC measurements:

$$Pr(T(fake) >= T(actual)_{max}|data) = 0.1066, Pr(T(fake) <= T(actual)_{min}|data) = 0.1073 \tag{1}$$

Since these values are far away from 0 and 1, based on these two measurements the model is quite a good fit. However, this is a very limited way of comparing the posterior predictive distribution to the actual data.

Problem 10.1.6.

The WHO issues a press release where they state that the number of novel disease outbreaks for this year was 20. Use your posterior predictive samples to test whether your model is a good fit to the data.

Answer:

$$Pr(T(fake) >= 20|data) = 0.0003 = 0.03\%$$

Thus, now it seems like there is a model misfit. This is a test of out of-sample predictive capability, and so we would expect this p value to be more extreme than the within-sample ones that we calculated before.

Problem 10.1.7.

By using your previously determined posterior as a prior, update your posterior to reflect the new datum. Graph the PDF for this new distribution.

Answer:

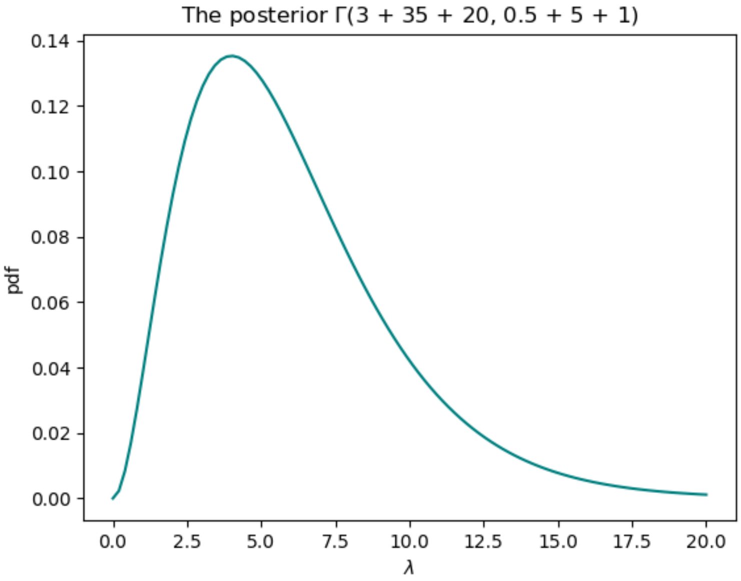

Figure 4: An estimate of the posterior predictive distribution with the new data point included.

Problem 10.1.8.

Generate posterior predictive samples from your new posterior and use it to test the validity of your model.

$$Pr(T(fake) >= 20|data) = 0.002 = 0.2\% \tag{2}$$

This within sample predictive capability is quite small, so it seems like there still is a model misfit.

Problem 10.1.9.

Would you feel comfortable using this model to predict the number of disease outbreaks next year?

No. It seems like the modle we are using is not quite right. It could be that disease outbreaks are in fact not at all independent and thus we could try using a negative binomial likelihood instead of a Poisson.

10.3 Discoveries Data

The file *evaluation_discoveries.csv* contains data on the numbers of “great” inventions and scientific discoveries in each year from 1860 to 1959 [1]. The aim of this problem is for you to build a statistical model that provides a reasonable approximation to this series. As such, you will need to choose a likelihood, specify a prior on any parameters, and go through and calculate a posterior. Once you have a posterior, you will want to carry out posterior predictive checks to see that your model behaves as desired.

Answer: Since the data consists of counts of discrete events with a fixed time frame in between events it sounds like the likelihood should either be a Poisson or a Negative Binomial distribution. However, I start with plotting the data and calculating statistics from it. In the left plot in Figure 5 it looks like there could be autocorreltaion between the number of discoveries made between consecutive years. In the histogram we see that the variance of the data exceeds the mean, more specifically the mean is: 3.1 and the variance is: 5.03.

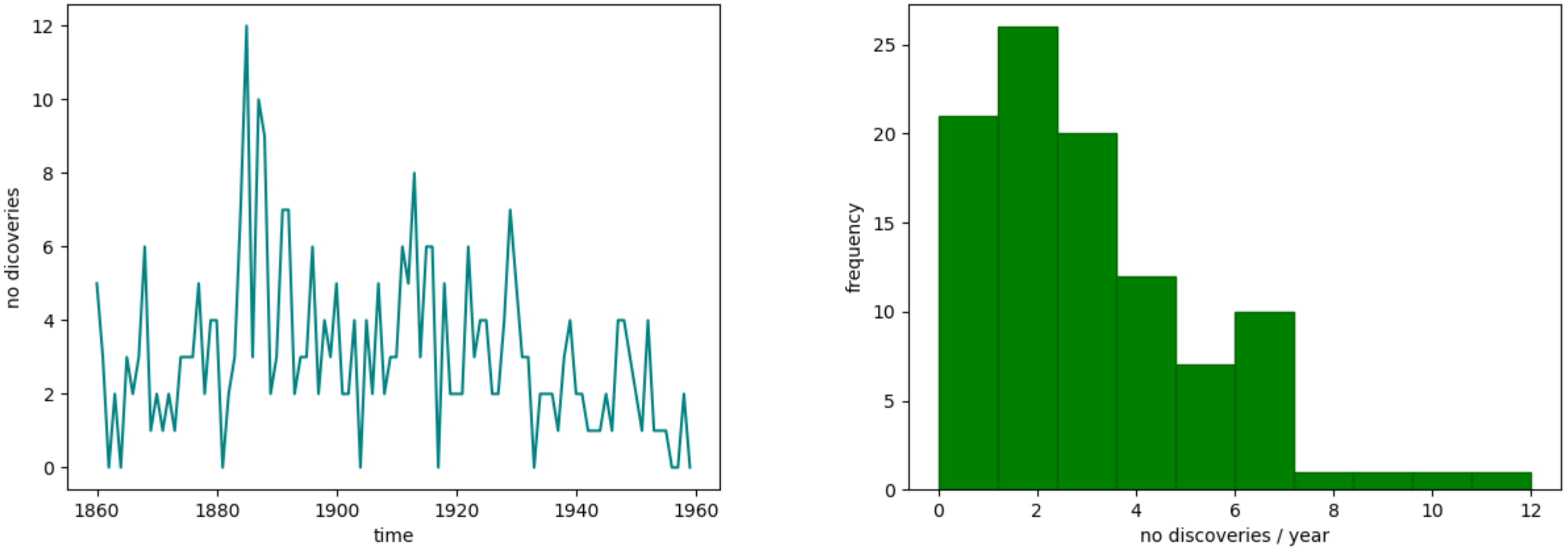

Figure 5: Number of discoveries of time.

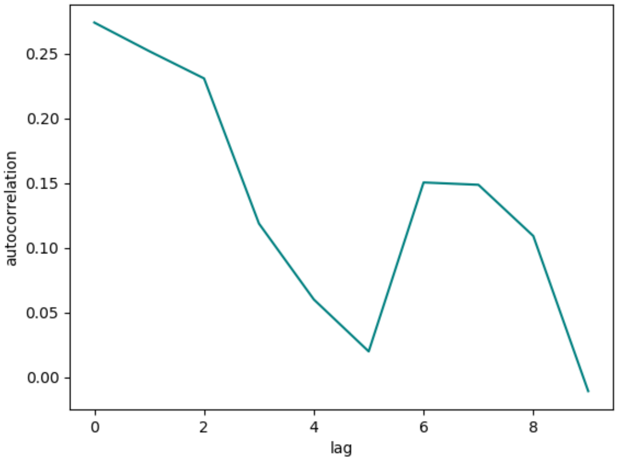

Figure 6: Autocorrelation between no discoveries per year.

The horizontal axis of Figure 6 shows the size of the lag between the elements of the time series. For example, the autocorrelation with lag 2 is the correlation between the time series elements and the corresponding elements that were observed two time periods earlier. Thus, it seems as though the autocorrelation has reached zero after 10 years. Perhaps a certain large scientific discovery makes it more likely for someone else to make a new large discovery based on the previous discovery.

Despite our better judgment we start by making the assumption that the events being modeled are independent and identically distributed over time. By making this assumption we can use a Poission likelihood with a Gamma prior. Another option would be to use the negative binomial likelihood which would allow the variance to exceed the mean and for events to be dependent. However, the possion distribution allows for individual events to have their own discovery rate, while the negative binominal likelihood assume that all event over time occur at a common rate. This assumption might not be valid for our data though.

10.4 Marginal likelihood of voting

Suppose that we collect survey data where respondents are asked to indicate for whom they will vote in an upcoming election. Each poll consists of a sample size of 10 and we collect the following data for 20 such polls:
$\{2, 7, 4, 5, 4, 5, 6, 4, 4, 4, 5, 6, 5, 7, 6, 2, 4, 6, 6, 6\}$. We model each outcome as having been obtained from a $X_i \sim B(10, \theta)$ distribution.

Problem 10.4.1.

Find the posterior distribution where we specify $\theta \sim beta(a, 1)$ as a prior. Graph how the posterior changes as $a \in [1, 10]$.

Answer: