Bayesian Statistics, Some More Problems

Friday, Nov. 8, 2024

---

1. Best Fit Using Least Squares — Sewage Outflow Event

A winter rainstorm has overwhelmed the sewage system in a lovely coastal location where you are mayor. During the day of the rainstorm, which we will call Day 0, a large amount of untreated sewage had to be dumped into the ocean.

Over the next three days, which we will call Days 1, 2, and 3, the fecal bacteria concentration in the ocean steadily declined, and you allowed people to go back into the water at the beaches.

Using historical records from previous events, you have good evidence that the fecal bacteria concentration, $c(d)$, where d is the day number, obeys the relationship,

$$c(d) = a \left(\frac{1}{2}\right)^d$$

Your civil engineers and environmental scientists measured the following concentrations with the following error bars:

In[1]:=
```
TableForm[{{1, 47}, {2, 25}, {3, 7}}]
```

Out[1]//TableForm=

1	47
2	25
3	7

Notice that no data is given for Day 0. The problem is to do a least squares fit to find a. Once you know a, you have the concentration on Day 0.

2. More Sewage — Poisson Statistics and χ^2

Bacteria concentrations are obtained by your environmental scientists as follows:

They take a liter of water from the the ocean. They mix a nutrient into the water. The pour it into a large shallow dish, warm the water to 100 °F, and wait six hours.

Then they visually count the number of bacterial colonies in the sample. This is a “fecal coliform count.”

Knowing this procedure and your result from Part 1, what would you believe the uncertainty, σ in the Day 1 count (which was 47) to be?

Repeat for Days 2 and 3.

Use the uncertainties and your result from Part 1 to estimate the χ^2 of your fit. If it is very roughly 2, this is about what you would expect. If it is a lot more than 2, this is a poor fit. If it is a lot lower than 2, either you were just very lucky with the quality of the data, or one of your environmental engineers is faking fecal coliform data instead of going through the official procedure.

3. Hamilton vs. Madison — Done Better

We know that the analysis we did of the Hamilton vs. Madison authorship is very sensitive to the data binning. Let's do it better!

Use the tabulated data for Hamilton and Madison to get rates of upon per thousand words for each author:

Table 5.2

Rate	Hamilton	Madison
0 (exactly)	0	41
(0,1]	1	7
(1,2]	10	2
(2,3]	11	0
(3,4]	11	0
(4,5]	10	0
(5,6]	3	0
(6,7]	1	0
(7,8]	1	0
	48	**50**

Having gotten a_{Hamilton}, and a_{Madison} we now consider the paper with the unknown authorship.

It had 2 upons in 2008 words. Use Poisson statistics to calculate likelihoods with the alternative hypotheses.

4. Gaussian Data with a Gaussian Prior

We have been doing introductory problems where the prior was a flat line for some range. That made it easy to tabulate the prior. In the real world, this might have to be done, but more commonly the prior belief is captured in a more complex function.

Suppose, after studying past data, you come to the conclusion that a good prior for bacteria survival time is:

$$P(m) = \frac{1}{\sqrt{2\pi}} e^{-(m-5)^2/2}$$

In other words the survival time is most likely 5 hours, but it has a standard deviation of 1.

Along comes three new data points for the survival time, $x = 3.6$, $x = 4.7$, and $x = 5.8$ hours.

(a) If you knew m (you don't!), what is the likelihood of getting $x = 3.6$? Your answer will be a function of m.

(b) Repeat for $x = 4.7$.

(c) Repeat for $x = 5.8$.

(d) What is the joint likelihood (the product of these three functions)? Again this will be a function of m.

(e) What is the product of these three functions multiplied by the prior?

(f) I will bring my laptop and graph this function and compute the integral of this function. This will be our denominator in our final expression for the posterior, which is a new function of m.