Assignment for Chapter 6 part II

The taxicab problem

In section 6.20, the “taxicab problem” is mentioned. It concerns a person who wakes up in some unknown town, seeing a taxicab with a number on it. The premise is that all taxicabs have numbers from 1 to N, where N is the total number of taxicabs (unknown to the person). When asked for a guess on how large N is, the person answers $N_{est} = 2 \times x$, where x is the number observed on the cab. Although such a guess cannot possibly be very precise, it seems, at least to me, to have some sense and with a gun to my head I would probably answer in the same way.

As we have seen, and something that Jaynes repeatedly point out, is that the choice of prior can have a strong influence on the inferences made (here these are summarized as a point estimate) when we have very little data, as in this case. So does usually also the choice of estimation method (conditional mean vs median etc.).

Instead of going through the original problem, Jaynes analyses a very similar problem involving continous variables and the taxicab problem is left aside a bit. Here we shall go back to the original discrete problem and we play around with some priors and estimation methods and compute point estimates to see what underlying assumptions the person on the train might have been using when making the intuitive guess (assuming he reasons using the Bayesian machinery).

(A) Complete the solution to the taxicab problem for arbitrary number of observations, n, when choosing the point estimate to be

1. Conditional mean.
2. Median.

Assign a rectangular discrete prior in $[N_{00}, N_1]$ and assume that the largest observed cabnumber, $x_{\max}$, is larger than N_{00}. Can you back up the above intuitive guess in any way? (i.e., take a particular look at the case n=1).
Give a closed form approximate analytical expression for the conditional mean.
Examine, for different n, how it depends on the parameters N_{00} and N_1 in the prior. For which n does the choice of N_1 no longer influence the solution significantly? Comments?

(B) Maybe a rectangular prior for N is not adequate for this problem. After all, there are much fewer large cities than there are middle- or small sized towns. Thus, it is more likely to have a relatively small maximum number of cabs than a large number. A prior that favors small N should therefore be more natural in this problem. Try Jeffreys’ prior and repeat the above exercises. Comments? What about the intuitive guess now?

Hints:

- To obtain analytical approximate expressions for your estimates you will have use of intgral approximation of sums. Typically you will bump into sums over negative powers that can be approximated as[1]

[1] With a little give and take for the exact position of the integration limits to get the best approximation.

$$\sum_{m=m0}^{m1} \frac{1}{m^k} \approx \int_{m0}^{m1} \frac{1}{x^k} dx$$

which we have simple formulas for. In this way "our" solutions will also be fairly similar to Jaynes' continous version.

- For your own sake, plot the posteriors you derive for some of the cases just to check out that your calculations seem to be ok.