STA 250: Statistics

Notes 10. Prior Selection: Contextual and Default Choices

Book chapters: –

1 How do we choose a prior for our analysis?

In a Bayesian analysis, the only choice the analyst has to make is the prior selection. We will avoid, for now, discussions on the issue of model choice, which is pertinent to both classical and Bayesian statistics. How does she choose a prior pdf/pmf? There are two possible approaches toward this. First, the analyst may carefully compile all background information she (or the expert she is working with) has about the problem and also think of what she is willing to believe about pertinent quantities, and then choose a prior pdf/pmf that matches with her information and beliefs. Or, she could take Laplace's path of reasoning "I have no reason to prefer one value of the parameter over another", and seek a uniform prior pdf/pmf. The first approach is often called a "subjective" or an "informed" prior selection. But it's better to call it a "contextual" selection. The second approach is often referred to as an "objective" or a "non-informative" or a "default" selection. Neither approach guarantees a unique choice. There maybe many prior pdfs/pmfs that match the expert's information + beliefs. And certainly there are more than one choices of default priors.

2 Contextual choice

Contextual choices are necessarily made in a case-by-case manner. There is no universal recipe. But there are certain broad guiding principles. These principles involve some research from psychology dealing with how people tend to express uncertainty in a quantitative way. By design we are not very good at this, but we do have a general ability to express uncertainty in a qualitative manner.

For example, assuming that you have not looked up the height of Duke Chapel, I could ask what you thought about it being taller than 50 feet. You may answer "more likely to be taller than 50 feet than not" (or the other way). Probing this to a limit I could ask you to give me a number q_1 such that you think the chapel is equally likely to be taller or shorter than q_1 feet. I will then equate q_1 with the "median of *your pdf* on the Chapel height".

Let's suppose your $q_1 = 200$. Next I could ask what you thought about the Chapel height, if you were told that it is taller than 200 feet. Is it taller than 250 feet? You may answer "more likely to be shorter than taller". I could ask you for a number q_2 such that you believed the Chapel to be equally likely to be taller or shorter than q_2 feet. I would then equate this with the "3rd quartile of *your pdf* on the Chppel height".

I could carry this on many steps each time getting a quantile of *your pdf* on the Chapel height (also probing the other side of q_1, i.e., asking you to think about the Chapel height

if you were told it was shorter than 200 feet). Once I get a dense enough set of quantile points, I can more or less trace out what your pdf should look like. But the deeper I probe, the more shaky (and tired) you will be in carrying out the thought experiment, and your answers will be less reliable. So I will have to stop, usually after only a handful of questions. At that point I will have a handful of quantiles determined, and obviously I could find a large number of pdfs matching those quantiles.

To resolve this, it helps to decide on a smaller collection of pdfs to work with. For example, if I had decided to match your belief to a $\mathsf{Normal}(a, b^2)$, then I could stop with only q_1 and q_2 and uniquely determine a and b. Indeed, the $\mathsf{Normal}(a, b^2)$ pdf has median $= a$ and 3rd quartile $= a + b\Phi^{-1}(0.75) = a + 0.67b$. So if you answered $q_1 = 200$ and $q_2 = 240$ then $a = 200$ and $b = (240 - 200)/0.67 = 59.7$.

For standard Bayesian analyses, such "smaller collections" are presented by the conjugate families. Here is a worked out example.

3 Weekly food expenditure analysis

Suppose I want to create an information booklet for incoming Duke students. Among other things, I want to include the dollar amount a student is likely to spend on food every week. My data would be the numbers I get from STA114 students reporting their weekly expenditure on food averaged over last 5 weeks. I'd model the data $X = (X_1, \cdots, X_n)$ as $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ with a conjugate prior $\xi(\mu, \sigma^2) = \mathsf{N}\chi^{-2}(m, k, r, s^2)$. What m, k, r and s should I work with? I will ask the following four questions to my "expert" (you!):

First, imagine a student randomly selected from campus, suppose her weekly food expenditure is X^*

1. Give me q_1 such that you believe $X^* < q_1$ & $X^* > q_1$ are equally likely
2. Give me q_2 such that you believe $X^* < q_2$ & $X^* > q_2$ are equally likely if you were told X^* is larger than the q_1 you gave me earlier
3. Give me q_3 such that you believe $X^* < q_3$ & $X^* > q_3$ are equally likely if you were told X^* is larger than the q_2 you have me earlier

Now imagine another randomly selected students and suppose his weekly food expenditure is Y^*. Think about $D^* = |X^* - Y^*|$.

4. Give me q_4 such that you believe $D^* < q_4$ & $D^* > q_4$ are equally likely

I then equate q_1, q_2, q_3 with the median, 3rd quartile and 87.5% quantile of X^* and equate q_4 with the median of D^*. With $\xi(\mu, \sigma^2) = \mathsf{N}\chi^{-2}(m, k, r, s^2)$, we have

$$\frac{X^* - m}{s\sqrt{1 + 1/k}} \sim \chi^2(r) \quad \text{and} \quad \frac{X^* - Y^*}{s\sqrt{2}} \sim \chi^2(r) \quad \text{(see HW 5)}.$$

From this link-up we can solve for m, k, r, s (run code).

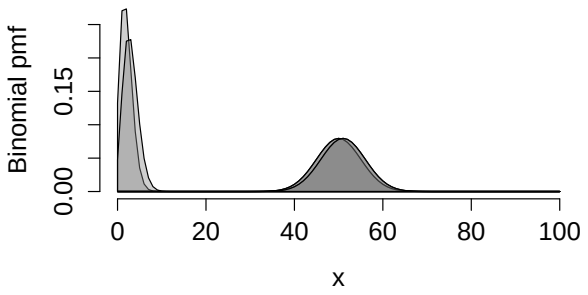
Couple of comments are due. First, we did not ask the expert about μ or σ^2. It is usually more difficult to express uncertainties about abstract quantities. We do not bet on things that we can never measure or observe. Second, after getting q_2, I could have asked "give me q_2' such that you believe $X^* < q_2'$ & $X_2^* > q_2'$ are equally likely if you were told $X^* < q_1$". I didn't ask this because the model entertains an X^* only symmetric around its median, so either $q_2' = q_2$ or we cannot work with the conjugate priors. Also, even with answers to the four questions above, we may not get an exact match (no solution) with a $\mathsf{N}\chi^{-2}(m, k, r, s^2)$. In either case, we should either ask the expert to revise her answers or we should look for a different prior family.

4 Default uniform choice

Recall that in his female birth rate analysis, Laplace used a uniform prior on the birth rate $p \in [0, 1]$. His justification was one of "ignorance" or "lack of information". He pretended that he had no (prior) reason to consider one value of $p = p_1$ more likely than another value $p = p_2$ (both values coming from the range $[0, 1]$). A uniform pdf is consistent with such a consideration. But there is a logical flaw.

Suppose a disciple of Lapalce decides to re-do the analysis. But being a betting enthusiast, he would rather analyze the log odds ratio $q = \log \frac{p}{1-p}$ of the female birth rate. Following the great man's notion of "no preference" he uses the prior density function $\tilde{\xi}(q) = 1$, $q \in (-\infty, \infty)$ (this is not a pdf, but we will ignore that for the moment). Next day he consults Laplace who tells him under the assumption of "no-preference", the right prior on p is $\mathsf{Uniform}(0, 1)$ and so the resulting right prior on q should be (by change of variable with $q = e^q/(1 + e^q)$) $\xi(q) = e^q/(1 + e^q)^2$, $q \in (-\infty, \infty)$ [this is the standard logistic pdf].

The logical flaw lies in measuring "uniformness". We want a pdf that spreads uniformly over the model space $\{\mathsf{Binomial}(n, p) : p \in (0, 1)\}$. It is wrong to assume that spreading uniformly over the model space is same as spreading uniformly over the parameter space. To see this consider two pairs of pmfs $\{\mathsf{Binomial}(n, 0.5), \mathsf{Binomial}(n, 0.51)\}$ and $\{\mathsf{Binomial}(n, 0.01), \mathsf{Binomial}(n, 0.02)\}$. For each pair, the mutual separation in the parameter space is $\Delta p = 0.01$. But the pmfs in the first pair are much more different from each other than those in the second pair.



5 Jeffreys' default prior

Harold Jeffreys recommended a general recipe for finding a default prior that is indeed uniformly spread over the model space. His recipe is widely used for regular statistical models with a scalar parameter. Suppose the model is: $X \overset{\text{IID}}{\sim} g(x_i|\theta)$, $\theta \in \Theta$ is scalar. The (single observation) Fisher information at a parameter value θ is defined as:

$$I_1^F(\theta) = -\mathrm{E}_{[X_i|\theta]}\left\{\frac{\partial^2}{\partial\theta^2}\log g(X_1|\theta)\right\}$$

which is always positive for regular models. Jeffreys recommended to take:

$$\xi^J(\theta) = \text{const} \times \sqrt{I_1^F(\theta)}.$$

For standard models, these computations can be carried out routinely. Here are some examples, along with the corresponding posterior pdfs.

Model	Jeffreys' prior	Posterior
$X \sim \mathsf{Binomial}(n,p)$ $p \in (0,1)$	$\xi^J(p) = \mathsf{Beta}(0.5, 0.5)$	$\mathsf{Beta}(x+0.5, n-x+0.5)$
$X_i \overset{\text{IID}}{\sim} \mathsf{Poisson}(\mu)$ $\mu > 0$	$\xi^J(\mu) = \text{const}/\sqrt{\mu}$	$\mathsf{Gamma}(n\bar{x}+0.5, n)$
$X_i \overset{\text{IID}}{\sim} \mathsf{Exponential}(\lambda)$ $\lambda > 0$	$\xi^J(\lambda) = \text{const}/\lambda$	$\mathsf{Gamma}(n, n\bar{x})$
$X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ $\mu \in (-\infty, \infty)$ σ known	$\xi^J(\mu) = \text{const}$	$\mathsf{Normal}(\bar{x}, \sigma^2/n)$

Often, Jeffreys' prior for an unbounded parameter spac is not a pdf. It is a density function that is positive, but it integrates to infinity, and cannot be normalized to make into a pdf. Such a prior is called an "improper prior". Analyses under improper priors are accepted as long as it produces a posterior pdf that is proper, i.e., a true pdf and not just a density function. This is true for all three improper priors in the above table.

6 Reference prior for two-parameter normal model

Jeffreys' recipe is widely accepted for scalar parameter models. But for vector valued parameters, people prefer other default choices. A popular choice is the so called "reference prior" (see http://arxiv.org/pdf/0904.0156.pdf). For the full normal model: $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, $\sigma > 0$, the reference prior is $\xi^R(\mu, \sigma^2) = \text{const}/\sigma^2$. The corresponding posterior is $\mathsf{N}\chi^{-2}(\bar{x}, n, n-1, s_x^2)$ (which is proper if $n \geq 2$).