STA 114: Statistics

Notes 14. Elicitation for the conjugate two-parameter normal model

Analyzing weekly food expenditure of Duke students

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 $\xi(\mu, \sigma^2) = \mathsf{N}\chi^{-2}(m, k, r, s)$. What m, k, r and s should I work with?

Eliciting m, k, r and s

Because $\xi(\mu, \sigma^2)$ is supposed to reflect our prior beliefs (knowledge + assumptions we are willing to make) about (μ, σ^2), we should first quantify our prior beliefs and then pick m, k, r and s so that the corresponding $\mathsf{N}\chi^{-2}(m, k, r, s)$ distribution gives a reasonable match to our quantified beliefs.

But we must pause here to think about this point. The variables (μ, σ^2) have no physical interpretation – they index a collection of pdfs and are just mathematical quantities. It is unlikely that we can do a good job of expressing our beliefs about variables that are not tangible. Human mind is not very good at that.

However, the variables (μ, σ^2), through our model, determine the behavior of the actual observable data $X_1, \cdots, X_n$. And because these represent quantities we can easily relate to, it is much more appealing to quantify beliefs about these variables. Our beliefs about model parameters are implicit in our beliefs about data.

So we shall think of hypothetical students randomly chosen from the population and wonder about the numbers $Y_1, Y_2, \cdots$ they are likely to report as their weekly food expenditure averaged over last 5 weeks. Since these variables are same as what we'd collect as data, we must also model $Y_j \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ with prior belief $\xi(\mu, \sigma^2) = \mathsf{N}\chi^{-2}(m, k, r, s)$.

General strategy: bisection approach

Suppose we want to quantify our beliefs about a scalar variable Z and then choose a pdf/pmf $f(z)$ that matches these quantified beliefs. There are various things we can quantify about Z, e.g., its center, spread, a range that is likely to contain most possible values, whether it is likely to be asymmetrically distributed around its center and so on. It is known, through experimentation, that we are fairly good at quantifying beliefs about "central values", but not so good at quantifying beliefs about spread or range. In particular, the question that we can most reliably answer is:

> *What is the number q_1 that we think Z is equally likely to be larger or smaller than?*

By "reliably answer" I mean that in answering this question, what we really believe and what we say we believe are usually close [psychologists have ways of figuring out what we *really believe*, or so they claim.]

Once we identify q_1, we must restrict our choice of $f(z)$ to pdfs that have q_1 as their median (i.e., 0.5-th quantile, i.e., $P(Z < q_1) = P(Z > q_1) = 0.5$ under these pdfs). Clearly, there are infinitely many pdfs that satisfy this. To make further progress, we need to answer more questions about our beliefs on Z. Now that the median has already been quantified, how can we talk about "centers" any more?

There is a fairly clever approach. We next ask this,

> *Imagine we are told $Z > q_1$ (recall q_1 is already identified). Given this information, what is the number q_2 that Z is equally likely to be larger or smaller than?*

Once we identify q_1, q_2, our choice of $f(z)$ must satisfy conditions: $P(Z < q_1) = 1/2$ and $P(Z < q_2) = 3/4$ under this pdf. So q_2 gives the 0.75-th quantile of $f(z)$.

We can repeat this on the left side, given the information $Z < q_1$ identify q_3 that Z is equally likely to be larger or smaller than. Then q_3 is the 0.25-th quantile $f(z)$. Continuing like these, we can identify the 0.875-th, the 0.125-th, the 0.9375-th, the 0.0625-th,... quantiles of $f(z)$.

Of course we can't continue forever. Pretty soon we start answering "I don't know", "I really don't know", "leave me alone"... Wherever we stop, we'd still have a large collections of pdfs that will match the quantities we have identified as the desired quantiles. At this point, we usually choose the one (among the matching ones) that is convenient to work with[1].

More intelligently, we can start with a collection of convenient pdfs (like a conjugate family of prior pdfs for a Bayesian analysis) and keep quantifying $q_1, q_2, \cdots$ until a member of this family is uniquely determined as the only one that provides a match. If the collection of pdfs is indexed by k many unknown quantities, then we are likely to get a unique, exact match by the time we have quantified k quantiles.

Belief quantification for Y_1 and $Y_1 - Y_2$

For our normal model, we focus on the collection of prior pdfs $\{\mathsf{N}\chi^{-2}(m, k, r, s) : -\infty < m < \infty, k > 0, r > 0, s > 0\}$. Since this collection is indexed by 4 quantities, we need four pieces of quantification on $Y_1, Y_2, \cdots$. We will quantify three quantities for Y_1 and one quantity for $Y_1 - Y_2$. Under our model, the pdfs of these variables indeed depend on the choice of m, k, r, s. In particular:

$$\frac{Y_1 - m}{\sqrt{s(1 + 1/k)}} \sim t(r), \quad \text{and} \quad \frac{Y_1 - Y_2}{\sqrt{2s}} \sim t(r).$$

[1] If we are more careful, we choose a few of such pdfs and perform our analysis under each, and then present all. If we are lucky the reports are close. Otherwise, we say there is too much prior uncertainty to come up with a singular analysis.

This is because of the following result (this is related to Result 2 from notes 10/19).

RESULT 1. If $(W, V) \sim \mathsf{N}\chi^{-2}(m, k, r, s)$ and $U|(W = w, V = v) \sim \mathsf{Normal}(aw, bv)$ then $\frac{U - am}{\sqrt{s(b + a^2/k)}} \sim t(r)$.

Proof. For U to have that conditional distribution, it must equal $U = aW + \sqrt{bV}Z$ for a $Z \sim \mathsf{Normal}(0, 1)$ that is independent of W and V. Sp given $V = v$, $U = aW + \sqrt{bv}Z$. But given $V = v$, $\sqrt{b}Z \sim \mathsf{Normal}(0, bv)$ because $Z \sim \mathsf{Normal}(0, 1)$ no matter what V is, and $aW \sim \mathsf{Normal}(am, a^2v/k)$ by properties of $\mathsf{N}\chi^{-2}(m, k, r, s)$ distributions, and these two are independent. So, given $V = v$, $U \sim \mathsf{Normal}(am, a^2v/k + bv) = \mathsf{Normal}(am, v/k')$ where $1/k' = b + a^2/k$. Hence $(U, V) \sim \mathsf{N}\chi^{-2}(am, k', r, s)$, from which the result follows. □

We will start by quantifying the median, the 0.75-th and the 0.875-th quantiles q_1, q_2, q_3 for Y_1. This follows the bisection approach discussed above, but only on one side (we do not get 0.25-th quantile, etc.). This is because we are restricted only to pdfs of Y_1 that are symmetric around the median. We also apply the bisection method on $Y_1 - Y_2$ to quantify its 0.75-th quantile (the median must be quantified 0, by symmetry of Y_1 and Y_2).

Solving for m, k, r, s

Because $\frac{Y_1 - m}{\sqrt{s(1+1/k)}} \sim t(r)$, for any fraction $u \in (0, 1)$ the u-th quantile of Y_1 must equal $m + \sqrt{s(1 + 1/k)}\Phi_r^{-1}(u)$ where $\Phi_r^{-1}(u)$ is the u-th quantile of the $t(r)$ pdf. First note that $\Phi_r^{-1}(0.5) = 0$ for any r. So

$$q_1 = m + \sqrt{s(1 + 1/k)}\Phi_r^{-1}(0.5) = m$$

and so $\boxed{m = q_1}$.

Next, in our old notations, $\Phi_r^{-1}(0.75) = \Phi_r^{-1}(1-0.5/2) = z_r(0.5)$ and similarly, $\Phi_r^{-1}(0.875) = z_r(0.25)$ and so

$$\begin{aligned} q_2 &= m + \sqrt{s(1 + 1/k)}z_r(0.5) \\ q_3 &= m + \sqrt{s(1 + 1/k)}z_r(0.25) \end{aligned}$$

and so $\boxed{\frac{z_r(0.5)}{z_r(0.25)} = \frac{q_2 - m}{q_3 - m} = \frac{q_2 - q_1}{q_3 - q_1}}$. The ratio $z_r(0.5)/z_r(0.25)$ is a continuous, increasing function in r and ranges between 0 (for $r \to 0$) and $z(0.5)/z(0.25) = 0.5863347$ [for $r \to \infty$, as $z_r(\alpha)$ becomes $z(\alpha)$]. See Figure 1. Therefore it is important that we have $\frac{q_2 - q_1}{q_3 - q_1}$ within this range. Otherwise, there is no $\mathsf{N}\chi^{-2}(m, k, r, s)$ that matches our prior belief. In case of a mismatch we may revisit some of our answers about q_1, q_2 and q_3. The most suspect would be q_3 and a revised answer maybe considered for which a match occurs. If $\frac{q_2 - q_1}{q_3 - q_1}$ is inside the range $[0, 0.5863347]$ then we can identify r as follows.

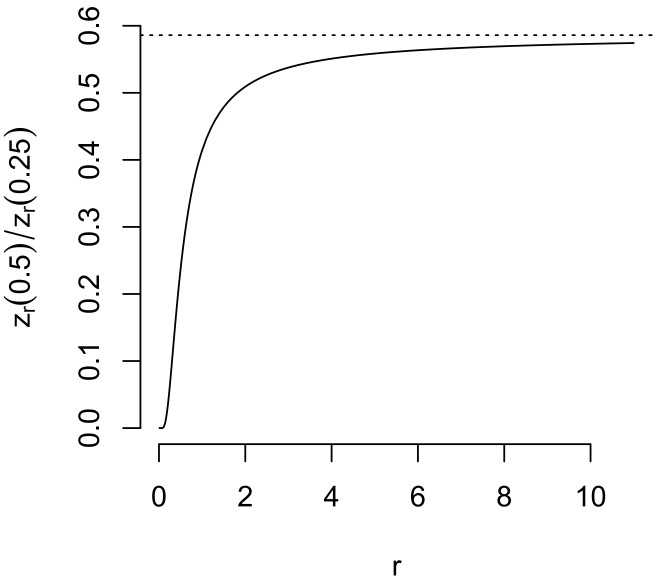

Figure 1: The ratio $\frac{z_r(0.5)}{z_r(0.25)}$ as a function of r.

```
ratio <- (q2 - q1) / (q3 - q1)
fn <- function(r) return(qt(0.75, r) / qt(0.875, r) - ratio)
r.sol <- uniroot(fn, interval = c(1e-3, 1e3))
r <- r.sol$root
```

Now that we have m and r, we can also identify $s' = s(1 + 1/k)$ from the identity $q_2 = m + \sqrt{s(1 + 1/k)} z_r(0.5)$. Namely, $\boxed{s' = \{(q_2 - q_1)/z_r(0.5)\}^2}$. But we cannot disentangle s and k from this. In fact no amount of further quantification on Y_1 can identify s and k separately from s'.

So we now turn to $Y_1 - Y_2$ whose 0.75-th quantile must equal $0 + \sqrt{2s} z_r(0.5)$. Equating this to q_4, and using the value of r that we obtained before, we can now identify s by $\boxed{s = 0.5(q_4/z_r(0.5))^2}$. Combine this with the identified value of $s' = s(1 + 1/k)$ to identify k as: $\boxed{k = s/(s' - s)}$. This is a legitimate value for k, provided $s' > s$. If we do not get this then again we need to see if we can revise our quantified beliefs.

Example (Weekly food expenditure)**.** To be done at lecture 10/21.