Prof. David Draper
Department of
Applied Mathematics and Statistics
University of California, Santa Cruz

AMS 132: Discussion Section 5

Today we'll look at a sampling distribution that was once important historically but is now not used much, so that you can see some of the sorts of problems that arise when you try to fit unfriendly distributions to data.

You've taken n IID measurements $\boldsymbol{y} = (y_1, \ldots, y_n)$ of a quantity μ that lives somewhere on the real line, and you choose the following for your sampling distribution:

$$p(y_i \,|\, \mu\, \sigma\, \mathcal{B}) = \frac{1}{2\,\sigma} \exp\left(-\frac{|y_i - \mu|}{\sigma} \right) , \tag{1}$$

in which (as mentioned above) $-\infty < \mu < \infty$ governs the center of the distribution and $\sigma > 0$ controls the scale (or spread, or dispersion); in this problem the vector of unknown quantities is $\boldsymbol{\theta} = (\mu, \sigma)$. The great French mathematician, statistician, physicist and astronomer Pierre Simon de Laplace (1749–1827) first studied this distribution when trying to model measurement errors arising from the use of telescopes to locate objects in the solar system, and the great economist John Maynard Keynes (1883–1946) also studied its properties. Equation (1) has two standard names: the *Laplace* distribution (after its founder) and the *double exponential* distribution (we'll see why this name is appropriate below). Here I'll call (1) the Laplace(μ, σ) distribution.

(a) Use `R` to plot (on the same graph) the Laplace distributions for $-10 \le y_i \le 10$ with the following parameter values: $(\mu, \sigma) = \{(0,1), (0,2), (0,4), (1,1), (1,3)\}$. What aspect of the distribution does μ capture? How about σ? Explain briefly.

(b) By breaking the distribution up into two pieces — $(-\infty < y_i < 0)$ and $(0 < y_i < \infty)$ — show that if $(Y_i \,|\, \mu\, \sigma\, \mathcal{B}) \sim$ Laplace(μ, σ) then $E(Y_i \,|\, \mu\, \sigma\, \mathcal{B}) = \mu$ and $V(Y_i \,|\, \mu\, \sigma\, \mathcal{B}) = 2\,\sigma^2$. Do these findings agree with your exploration in (a)? Explain briefly.

(c) With $\boldsymbol{y} = (y_1, \ldots, y_n)$, show that the likelihood and log-likelihood functions in the sampling model $(Y_i \,|\, \mu\, \sigma\, \mathcal{B}) \overset{\text{IID}}{\sim}$ Laplace(μ, σ) are

$$\ell(\mu\, \sigma \,|\, \boldsymbol{y}\, \mathcal{B}) = c\, \sigma^{-n} \exp\left(-\frac{1}{\sigma} \sum_{i=1}^{n} |y_i - \mu| \right) \tag{2}$$

and

$$\ell\ell(\mu\, \sigma \,|\, \boldsymbol{y}\, \mathcal{B}) = c - n \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n} |y_i - \mu| \,, \tag{3}$$

respectively. Briefly explain what's immediately difficult about this log likelihood function when trying to identify $\hat{\mu}_{MLE}$ (especially when compared with a sampling model in which the measurement errors are Gaussian). Show that, whatever $\hat{\mu}_{MLE}$ is,

$$\hat{\sigma}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \mu| \,. \tag{4}$$

(d) To get some idea of what's going on with $\hat{\mu}_{MLE}$, consider a cut-down version of model (1) in which we temporarily pretend that σ is known: $(Y_i \,|\, \mu\, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Laplace}(\mu, \sigma)$ for $(i = 1, \ldots, n)$ with known $\sigma > 0$. Suppressing the irrelevant constant c, the likelihood and log-likelihood functions for μ in this new cut-down model are then

$$\ell(\mu \,|\, \boldsymbol{y}\, \mathcal{B}) = \exp\left(-\frac{1}{\sigma}\sum_{i=1}^{n} |y_i - \mu|\right) \tag{5}$$

and

$$\ell\ell(\mu \,|\, \boldsymbol{y}\, \mathcal{B}) = -\frac{1}{\sigma}\sum_{i=1}^{n} |y_i - \mu|\,, \tag{6}$$

respectively. Now consider the following two artificial data sets: $\boldsymbol{y}_1 = (1, 2, 9)$ and $\boldsymbol{y}_2 = (1, 2, 5, 9)$. It turns out that $\hat{\sigma}_{MLE}$ is about 2.7 for the first data set and 2.8 for the second one; to get useful pictures let's use the rounded value of $\sigma = 3$.

(i) Plot the likelihood and log-likelihood functions with $\boldsymbol{y}_1$ and study them. Will our usual way of finding $\hat{\mu}_{MLE}$ work with this data set? Will our usual Fisher information calculation be helpful in coming up with $\widehat{SE}(\hat{\mu}_{MLE})$? Explain briefly.

(ii) Repeat (i) with $\boldsymbol{y}_2$. What's even more unusual about the likelihood and log-likelihood functions with this data set? Explain briefly.

The MLE for μ in data set 1 turns out to be 2, and $\hat{\mu}_{MLE}$ in the second data set can be taken to be any number between 2 and 5, for example the midpoint 3.5. Notice that in both of these artificial data sets the MLE of μ is the *median* of the data set; this turns out to be true in general, although it's not particularly pleasant to prove it.

(iii) Regarding the likelihood function (when thinking in a Bayesian way) as an un-normalized density for μ, do you recognize it as something that could serve as the basis of a conjugate prior having a familiar mathematical form? Explain briefly.

(iv) How might you go about summarizing a posterior distribution in a setting in which no standard-family conjugate prior exists? Discuss with your TA how you might use *random samples from the posterior* to summarize it — this (Bayesian computation when no standard closed-form expression for the posterior exists) will be our subject next week.