Predictive Distributions

Suppose one would like to predict future observations. How should one proceed? Here we will see how such predictions are typically constructed from a Bayesian viewpoint.

As an example, let's consider blood pressure data. Recall that we had the data below on the diastolic blood pressure of Americans:

76, 71, 82, 63, 76, 64, 64, 74, 70, 64, 75, 81, 75, 78, 66, 62, 79, 82, 78, 62, 72, 83, 79, 41, 80, 77, 67.

From this data, we find $\overline{x} = 71.89$, and $s^2 = 85.18$, so that $s = \sqrt{85.18} = 9.22$

So, one obvious way to proceed, and what frequentists might tend to do, would be to predict future observations from this distribution, as a best guess:

$$\textit{next observation} \sim N(71.98, 85.15)$$

This takes our best estimate of the mean and our best estimate of the SD (or variance), and uses these values to describe the population and to make future predictions.

Some questions:

- Is this really the best we can do?
- What important factor might be missing from this method?
- Why is it philosophically difficult for frequentists to correct for this missing factor?
- In what way can Bayesian analysis improve on this situation?

Predictive Distributions

We realize that the exact mean and variance to plug into the predictive formula is uncertain. If we are uncertain about these values, using single point estimates will underestimate the full uncertainty inherent in making these predictions, resulting in the spread of the distribution of predictions being too narrow. Rather than knowing these values exactly, we know them only up to the posterior distribution. So, if we average over the posterior distribution, we can restore the missing uncertainty. The distribution created by averaging future predictions over the posterior densities of all unknown parameters is called the "**predictive density**" in Bayesian analysis.

Note that there is, philosophically speaking, no frequentist equivalent to this concept, since there are no distributions over unknown parameter values to average over in that paradigm.

In practice, what we need to do is to calculate the following predictive density:

$$\text{Predictive distribution}(z) = \int f(z|\theta) f(\theta|x)\ d\theta$$

In the above, x is the observed data and θ is the vector of unknown parameters. Thus $f(z|\theta)$ is the prediction we would make for future data if θ were known exactly, but since it is not, we average this quantity over $f(\theta|x)$, the posterior density of θ after observing the data x.

Note that the above formula is completely general, in that f can be any density or even an extremely complex model, and θ can represent one, two, or even thousands of unknown parameters. In practice, this can be a difficult integration problem (although it is actually not very difficult for Normal densities). So, most of the time we *simulate* values from the predictive density, rather than getting the exact analytical solution.

To do this, we follow the formula, and proceed in three steps:

1. First, draw a sample from the posterior density of θ, i.e., make a random draw from $f(\theta|x)$.

2. Then, pretend, for a moment, that this is the exactly correct value, and draw an random value from $f(z|\theta)$.

3. Repeat steps 1 and 2 many times, typically 1000 or more. Since a different value of θ is used each time, we automatically restore the uncertainty missing when we plug in just a single value, such as $\overline{x}$.

Let's use R to apply this method to our blood pressure example. Somewhat unrealistically for now (restriction to be removed within the next week), suppose we know the variance is equal to 9 exactly. We showed last class that the posterior density for μ, the unknown mean is $\mu \sim N(71.69, 2.68)$. So this is the distribution we must average our predictions over.

Further, once the mean μ is "known", each prediction is made by drawing $z \sim N(\mu, SD = 9)$. [If the SD were also not known, we would also draw randomly from its posterior distribution, really drawing from the joint density of μ and the SD.]

In R, this can be implemented as follows:

```
mu <- rnorm(10000, mean = 71.69, sd = sqrt(2.68))
z <- rnorm(10000, mean = mu, sd = 9)
```

Once we have z, we can summarize it through quantiles, means, variances, graphs, etc.

```
> mean(z)
[1] 71.8398
> sqrt(var(z))
[1] 9.197573
> quantile(z, c(0.025, 0.975))
    2.5%    97.5%
54.04826 90.22260
```

Note how this is a density that is more spread out than if we ignore the uncertainty in μ:

```
z <- rnorm(10000, mean = 71.69, sd = 9)
```

```
> mean(z)
[1] 71.69397
> sqrt(var(z))
[1] 8.981591
> quantile(z, c(0.025, 0.975))
    2.5%    97.5%
53.93716 89.28465
```

It is only slightly too narrow here, since we had only a single unknown parameter that we are ignoring, and because this parameter was reasonably well estimated. The differences can be much larger in other situations.

Similarly for other distributions, such as the binomial (see assignment), and for more complex situations (for example, this is very useful in dealing with missing data through multiple imputation).