Introductory Examples

Rice, Section 10.2, page 378, gives a data set of the melting point of beeswax. This data is in the file Chapter 10, beeswax.txt. Figure 1 gives a relative frequency histogram of the data. The histogram has the general shape of a normal distribution.

Consider the family of probability density functions (pdf)'s

$$\mathcal{F} = \left\{ f : f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ where } \mu \in R, \sigma^2 > 0 \right\}$$

This is a family of distributions, with parameter space $\Theta = \{(\mu,\sigma^2) : \mu \in R, \sigma^2 > 0\}$. Thus for an appropriate choice of some $f \in \mathcal{F}$, or equivalently an appropriate choice of μ, σ^2 in the parameter space, one can obtain a good description or fit of the data with this particular distribution.

Figure 1 is produced using the R program in the file bees.r

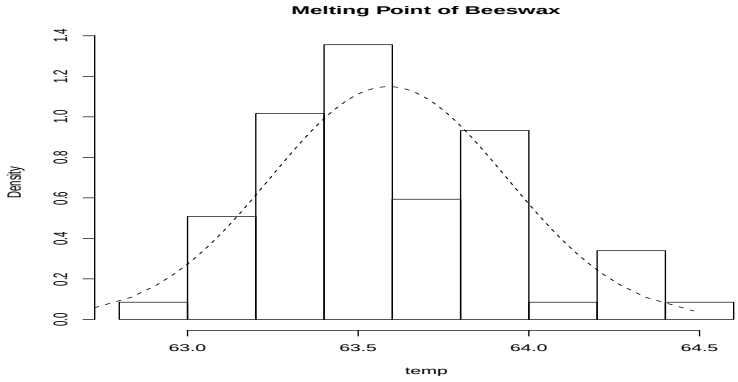

Figure 1: Beeswax Melting Point with Normal Overlay.

Another example of parametric modeling is the modeling of Illinois rainfall data. This data is described in Rice, page 414, problem 42. In this handout the rainfall from the 5 years is combined into one data set. Figure 2 gives a histogram of this data. Overlaid on the histogram is a fitted Gamma distribution.

Here the family of distributions is

$$\mathcal{F} = \left\{ f : f(x;\lambda,\alpha) = \left[\begin{array}{ll} \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{array} \right. , \text{ where } \lambda > 0, \alpha > 0 \right\}$$

The parameter space is $\Theta = \{(\lambda,\alpha) : \lambda > 0, \alpha > 0\}$. Using $\lambda = 1.90$ and $\alpha = .44$ gives a good fitting Gamma distribution to the histogram.

Natural questions are which values of the parameters are reasonable and does any member of the parametric family fit the data well.

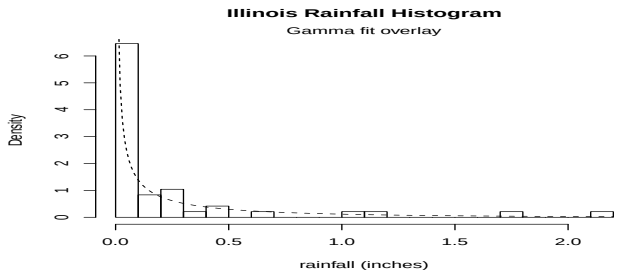

Figure 2: Illinois Rainfall with Gamma Density Overlay

A statistic is a function of the observable data. Often the observable data is $X_1, \ldots, X_n$ where these are an observed random sample from an experiment.

As an example consider the following calculated from the beeswax data, in particular the melting point.

```
n  =  59
mean = 63.58881
s2  =  0.1205624
mu.3  =  257146.4
median.x  =  63.53
sum.log =  244.9930
```

Notice also that the choice of the value of the parameter estimates in the above examples is also a statistic. Often a statistic is a single real valued object, but sometimes it is natural to think of these as a vector valued random variable. For example consider the random variable $(\hat{\lambda}, \hat{\alpha})$ calculated from the rainfall data.

Using the Michelson's alpha particle data in section 8.2 we can also obtain the following data, and hence statistic. We obtain from the sample of size $n = 1207$ the value $\hat{\lambda} = 0.8354598$. Thus for each 10 second interval the count will be, according to the model, an observation from a Poisson distribution with parameter $\theta = 10 * \lambda$. We can estimate θ by $\hat{\theta} = 8.35$. The table below gives the estimated counts, based on the fitted model.

```
$note [1] "Alpha particle counts in 10 second intervals"
$lam.hat [1] 0.83545
$chi.sq [1] 8.975796
$table
           counts   Prob   expected    Chisq
     0 - 2    18   0.0104   12.5692   2.3465
        3     28   0.0229   27.6043   0.0057
        4     56   0.0478   57.6557   0.0475
        5    105   0.0798   96.3381   0.7788
        6    126   0.1111  134.1443   0.4945
        7    146   0.1326  160.1031   1.2423
        8    164   0.1385  167.1997   0.0612
        9    161   0.1286  155.2096   0.2160
       10    123   0.1074  129.6713   0.3432
       11    101   0.0816   98.4865   0.0641
       12     74   0.0568   68.5680   0.4303
       13     53   0.0365   44.0660   1.8113
       14     23   0.0218   26.2967   0.4133
       15     15   0.0121   14.6465   0.0085
       16      9   0.0063    7.6479   0.2391
       17+     5   0.0056    6.7931   0.4733
```

The observed χ^2 value is 8.98. The R output for this analysis gives 8.975796, but how many digits are worth keeping? The construction and meaning of the Pearson's χ^2 test is discussed later in the course as an application of generalized likelihood ratio tests. So what does this observed number mean? Later we find this gives supporting evidence to the assumption that a Poisson model is a reasonable and *good* model for this experiment. Specifically the data is consistent with the assumption that the data is generated by a Poisson distribution.

Sometimes it is not always clear which statistical model to use. We may have positive random variables with a large mean. We may wish to use the Gamma family to model these, or perhaps a normal family.

For a r.v. X that is positive (non-negative) with probability 1 we have $P(X \geq 0) = 1$ ($P(X > 0) = 1$). However if $X \sim N(\mu, \sigma^2)$ then $P(X \geq 0) < 1 \neq 1$. However if μ large, relative to appropriate σ^2, then $P(X < 0) \approx 0$ and so the normal may still be *reasonable*.

Specifically if $X \sim \text{Gamma}(\alpha, \lambda)$ we have the following approximation. Recall X has moment generating function M. We will want to study this for large α, so that we will study this limit as $\alpha \to \infty$. Denote the dependence of the mgf on α as M_α. Therefore

Now consider the centred random variable

$$Y_\alpha = \frac{1}{\sqrt{\alpha}} \left(X - \frac{\alpha}{\lambda} \right) .$$

Notice we are dividing by $\sqrt{\alpha}$ instead of standardizing the $\sqrt{\text{Var}(X)} = \sqrt{\alpha}/\sqrt{\lambda}$. Let M_α be the mgf of Y.

$$\begin{aligned} M_\alpha(t) &= \text{E}\left(e^{tY}\right) \\ &= \text{E}\left(e^{\frac{t}{\sqrt{\alpha}}(X - \frac{\alpha}{\lambda})}\right) \\ &= \text{E}\left(e^{\frac{t}{\sqrt{\alpha}}X}\right) e^{-\frac{t\sqrt{\alpha}}{\lambda}} \\ &= M(\frac{t}{\sqrt{\alpha}}) e^{-\frac{t\sqrt{\alpha}}{\lambda}} \end{aligned}$$

Therefore

$$\log\left(M_\alpha(t)\right) = \alpha \left\{ \log(1) - \log(1 - \frac{t}{\lambda\sqrt{\alpha}}) - \frac{t}{\lambda\sqrt{\alpha}} \right\}$$

The student should verify that

$$\lim_{\alpha \to \infty} \log\left(M_\alpha(t)\right) = \frac{1}{2\lambda^2} t^2 .$$

Therefore by the Continuity Theorem, Y_α converges in distribution to $N(0, \frac{1}{\lambda^2})$ as $\alpha \to \infty$. Thus a Gamma distribution with large shape parameter α is very close to a normal distribution. Thus in this setting if a Gamma distribution with large shape parameter is a good fit, then so will a normal distribution be a good fit. This has a practical implication that sometimes there may be different models that may fit data equally well. Thus there may be more than one good answer, at least in terms of model selection or model choice.

As an example we can look at the plot of the beeswax hydrocarbon data. Figure 3 gives the histogram of this data along with a normal and gamma fitted model. The two fitted densities are nearly identical on this plot.

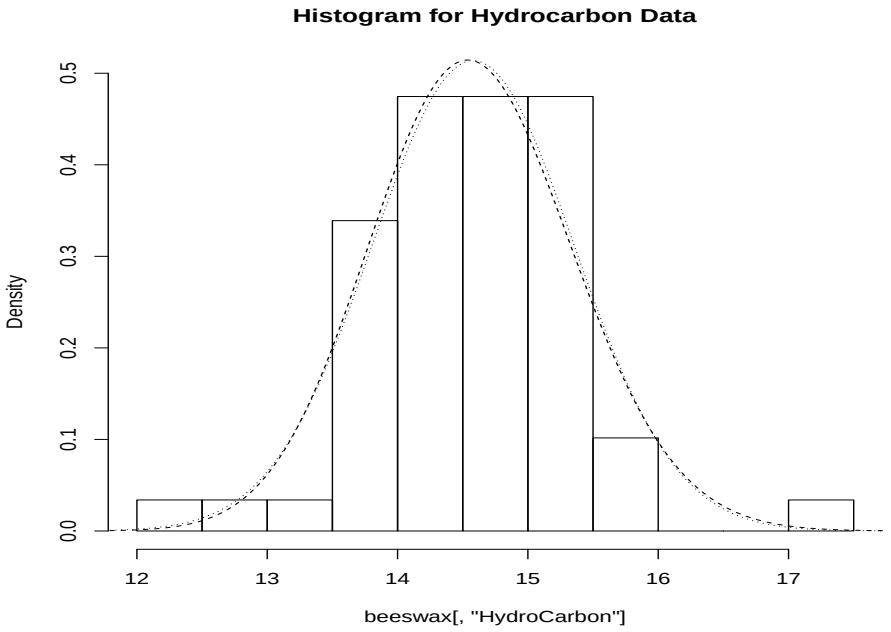

Figure 3: Beeswax Hydrocarbon data with Normal and Gamma Overlay.

When are statistical inferences valid? How do compare estimators or determine properties of an estimator?

With improved computing power can one make use of it for statistical estimation and statistical inference? A recent method to take advantage of this the so called *bootstrap* method, both in a parametric and non-parametric form. What are these?