Homework 1

Class

September 7, 1998

1 Problem 1.8

by Mitchell Craig Warachka
a) x has Weibull distribution

$$\frac{c}{\alpha^c} x^{c-1} exp^{\frac{-x^c}{\alpha^c}} \tag{1}$$

using a change of variables $x = y^{\frac{1}{c}}$ and inserting the Jacobian $\frac{1}{c} y^{\frac{1}{c}-1}$ yields

$$y \ \frac{c}{\alpha^c} \frac{1}{c} y^{\frac{1}{c}} \frac{1}{y} y y^{-\frac{1}{c}} exp^{-\frac{y}{\alpha^c}} = \frac{1}{\alpha^c} exp^{-\frac{y}{\alpha^c}} \tag{2}$$

which can be seen as an exponential (α^c) which is a gamma $(1, \alpha^c)$.
b) Likelihood Equations

$$log f(x_i|\alpha, c) = log(c) - clog(\alpha) + (c-1)log(x_i) - \frac{x_i^c}{\alpha^c} \tag{3}$$

summing over n observations and differentiating with respect to α

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\partial}{\partial \alpha} log f(x_i|\alpha, c) &= \frac{-nc}{\alpha} + \frac{c}{\alpha^{c+1}} \sum_{i=1}^{n} x_i^c = 0 \\
&\Rightarrow \ nc = \frac{c}{\alpha^c} \sum_{i=1}^{n} x_i^c \\
&\Rightarrow \ \alpha^c = \frac{1}{n} \sum_{i=1}^{n} x_i^c \\
&\Rightarrow \ \alpha = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^c \right)^{\frac{1}{c}}
\end{aligned}$$

summing over n observations and differentiating with respect to c while using the formula $\frac{\partial(x^a)}{\partial a} = x^a log(x)$

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\partial}{\partial c} log f(x_i|\alpha, c) &= \frac{n}{c} - nlog(\alpha) + \sum_{i=1}^{n} log(x_i) - \frac{1}{\alpha^c} \sum_{i=1}^{n} x_i^c log(\frac{x_i}{\alpha}) = 0 \\
&\Rightarrow \ nlog(\alpha) - \sum_{i=1}^{n} log(x_i) = \frac{n}{c} - \frac{1}{\alpha^c} \sum_{i=1}^{n} x_i^c log(\frac{x_i}{\alpha})
\end{aligned}$$

no explicit solutions for α and c.
c) Now sample contains censored observations at level y_0. If $x_i \leq y_0$ get x_i or if $x_i > y_0$ get no reading just use the cdf at time y_0 , which is $1 - exp^{-\frac{y_0^c}{\alpha^c}}$. If observation is censored, use $exp^{-\frac{y_0^c}{\alpha^c}}$ in likelihood as probability that x_i made it past y_0. $[P(x_i > y_0) = 1 - P(x_i \leq y_0) = 1 - F(y_0)]$ Define a new random variable δ_i which is equal to 1 if $x_i \leq y_0$ and 0 if $x_i > y_0$. The new log likelihood over n observations becomes

$$
\begin{aligned}
&log\left(\Pi_{i=1}^n f(x_i|\alpha,c)^{\delta_i}(1-F(y_0))^{1-\delta_i}\right)\\
&\sum_{i=1}^n \left(log(f(x_i|\alpha,c)^{\delta_i} + log(1-F(y_0)^{1-\delta_i}\right)\\
&\sum_{i=1}^n \left[\delta_i\left(log(c) - clog(\alpha) + (c-1)log(x_i) - \frac{x_i^c}{\alpha^c}\right) + (1-\delta_i)\left(-\frac{y_0^c}{\alpha^c}\right)\right]
\end{aligned}
$$

The situation for c and α is known from part b to not be explicit. Taking a derivative of the above with respect to α and c, it can be shown that the added complication of censored data does not yield explicit maximum likelihood estimators in terms of the data.

With respect to α

$$
\begin{aligned}
&\sum_{i=1}^n \left[\delta_i\left(\frac{-c}{\alpha} + \frac{c}{\alpha^{c+1}}x_i^c\right) + (1-\delta_i)\left(\frac{c}{\alpha^{c+1}}y_0^c\right)\right] = 0\\
\Rightarrow\quad & -\sum_{i=1}^n \delta_i + \frac{1}{\alpha^c}\sum_{i=1}^n \delta_i x_i^c + \frac{y_0^c}{\alpha^c}\sum_{i=1}^n (1-\delta_i) = 0\\
\Rightarrow\quad & \frac{1}{\alpha^c}\left(\sum_{i=1}^n x_i^c + y_0^c\sum_{i=1}^n(1-\delta_i)\right) = \sum_{i=1}^n \delta_i\\
\Rightarrow\quad & \alpha = \left[\frac{(\sum_{i=1}^n x_i^c + y_0^c\sum_{i=1}^n(1-\delta_i))}{\sum_{i=1}^n \delta_i}\right]^{\frac{1}{c}}
\end{aligned}
$$

With respect to c

$$
\begin{aligned}
&\sum_{i=1}^n \left[\delta_i\left(\frac{1}{c} - log(\alpha) + log(x_i) - \frac{x_i^c}{\alpha^c}log(\frac{x_i}{\alpha})\right) + (1-\delta_i)\left(\frac{-y_0^c}{\alpha^c}log(\frac{y_0}{\alpha})\right)\right] = 0\\
\Rightarrow\quad & \frac{1}{c}\sum_{i=1}^n \delta_i - log(\alpha)\sum_{i=1}^n \delta_i + \sum_{i=1}^n log(x_i)\delta_i - \frac{1}{\alpha^c}\sum_{i=1}^n \delta_i x_i^c log(\frac{x_i}{\alpha}) - \frac{y_0^c}{\alpha^c}log(\frac{y_0}{\alpha})\sum_{i=1}^n(1-\delta_i) = 0\\
\Rightarrow\quad & log(\alpha)\sum_{i=1}^n \delta_i - \sum_{i=1}^n \delta_i log(x_i) = \frac{1}{c}\sum_{i=1}^n \delta_i - \frac{1}{\alpha^c}\sum_{i=1}^n \delta_i x_i^c log(\frac{x_i}{\alpha}) - \frac{y_0^c}{\alpha^c}log(\frac{y_0}{\alpha})\sum_{i=1}^n(1-\delta_i)
\end{aligned}
$$

again no explicit solutions found for either parameter

(1.10)
The function to be maximized in the most general case is

$$\frac{Max}{c, \gamma, \alpha} \Pi_{i=1}^{17} \frac{c}{\alpha^c}(x_i - \gamma)^{c-1} exp^{\frac{-(x_i-\gamma)^c}{\alpha^c}} \left(1 - exp^{-\frac{(216-\gamma)^c}{\alpha^c}}\right) \left(1 - exp^{-\frac{(244-\gamma)^c}{\alpha^c}}\right) \tag{4}$$

where the first 17 x_i are the uncensored observations. Estimates were about .0078 for c and 3.5363 for α with a γ of 100.

2 Problem 1.49

Bárbara Patricia González A famous problem in classical Statistics is the *Behrens-Fisher* problem. It comes from the simple setting of two normal populations with unknown means and variances, because there is no optimal test to compare the means through the test of the hypotesis $H_0 : \theta = \mu$. Consider $X_1, ..., X_n$ a sample from $N(\theta, \sigma^2)$ and $Y_1, ..., Y_m$ a sample from $N(\mu, \tau^2)$ where $\theta, \mu, \tau, \sigma$ are unknown.

(a) Explain why a reasonable test should be based on the pivotal quantity

$$T = \frac{(\theta - \mu) - (\bar{x} - \bar{y})}{\sqrt{s_x^2/n + s_y^2/m}} \tag{5}$$

with $\bar{x} = \sum_i x_i/n, \bar{y} = \sum_i y_i/m, s_x^2 = \sum_i (x_i - \bar{x})^2/n - 1$ and $s_y^2 = \sum_i (\bar{y}_i - y)^2/m - 1$.

(b) Show that the distribution of T depends on σ/τ even when $\theta = \mu$ and *is not* a Student t-distribution.

3 Solution

Since $X_i \sim N(\theta, \sigma^2)$ and $Y_i \sim N(\mu, \tau^2)$ then we know that $\bar{X} \sim N(\theta, \sigma^2/n)$ and $\bar{Y} \sim N(\mu, \tau^2/m)$. If we let $Z = \bar{X} - \bar{Y}$ then $Z \sim N(\theta - \mu, \sigma^2/n + \tau^2/m)$, so a reasonable test would try to use the pivotal quantity T. We also know that

$$T_x = \frac{\theta - \bar{x}}{s_x/\sqrt{n}} \sim t_{n-1} \text{ and } T_y = \frac{\mu - \bar{y}}{s_y/\sqrt{m}} \sim t_{m-1}. \tag{6}$$

If we then define $tan\xi = \frac{s_x/\sqrt{n}}{s_y/\sqrt{m}}$ (ξ can be taken in the first quadrant), it is easy to see that

$$T = \frac{(\theta - \mu) - (\bar{x} - \bar{y})}{\sqrt{s_x^2/n + s_y^2/m}} = \frac{(\theta - \bar{x}) - (\mu - \bar{y})}{\sqrt{s_x^2/n + s_y^2/m}} =$$

$$= \frac{(\theta - \bar{x}) s_x/\sqrt{n}}{s_x/\sqrt{n}\sqrt{s_x^2/n + s_y^2/m}} - \frac{(\mu - \bar{y}) s_y/\sqrt{m}}{s_y/\sqrt{m}\sqrt{s_x^2/n + s_y^2/m}} = T_x sin\xi - T_y cos\xi. \tag{7}$$

Since ξ is known and the distributions of T_x and T_y are known, it follows that the distribution of T can be evaluated. This distribution is tabulated and is called Behrens' distribution.

Behrens' distribution turns out to have a rather difficult form, so that the density at any one point can only be found by a complicated integral.

The Behrens' distribution depends on the ratio σ/τ. It is esiest to see this when $m = n$. We first suppose that it is in fact the case that $\sigma/\tau = 1$. Then the sampling distribution for the case where $\sigma = \tau$ is known must still hold

$$\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}} = \sqrt{\frac{\sum_i(\bar{x}_i - x)^2}{(n-1)n} + \frac{\sum_i(\bar{y}_i - y)^2}{(m-1)m}} = \sqrt{\frac{2(\sum_i(\bar{x}_i - x)^2 + \sum_i(\bar{y}_i - y)^2)}{2(n-1)n}} = s\sqrt{\frac{2}{n}} \tag{8}$$

with $s = \frac{\sum_i(\bar{x}_i - x)^2 + \sum_i(\bar{y}_i - y)^2}{n+m-2}$. It follows that $T \sim t_{m+n-2}$. On the other hand if $\tau = 0$, then necessarily $s_y^2 = 0$, and hence $\sqrt{s_x^2/n + s_y^2/m} = \sqrt{s_x^2/n}$. Since it must also be the case that $\bar{y} = \mu$, the distribution of T is given by

$$T = \frac{\theta - \bar{x}}{s_x\sqrt{n}} \sim t_{n-1} \tag{9}$$

For intermediate values of σ/τ the distribution of T is intermediate between these forms, but is not, in general, a Student t-distribution, as we saw before. (Lee, P. 1989)

Despite this, a different approach is to consider T approximately $\sim t_{n+m-2-\Lambda}$, with Λ equal to the integer nearest

$$\frac{((m-1)s_x^2/n - (n-1)s_y^2/m)^2}{(m-1)s_x^4 + (n-1)s_y^4}. \tag{10}$$

It results that $0 \leq \Lambda \leq max(n-1, m-1)$.

The interpretation of Λ is as follows: if $\sigma \gg \tau$ and $n = m$, then $s_x^2 \gg s_y^2$ and Λ is approximately $m-1$, so the degrees of freedom of t depend on the precision we have estimating σ. If σ and τ are similar, and so are m and n then $\Lambda = 0$. Finally, if n and m are very different, and for example $n \gg m$ then Λ will be big and the degrees of freedom are reduced. (Peña, D. 1995)

4 References

Lee, P. (1989) *Bayesian Statistics: An Introduction.* **Oxford University Press, Oxford.**

Peña, D. (1995) *Estadística. Modelos y métodos.* **Alianza Universidad Textos. Madrid.**