UNIVERSIDADE DE SÃO PAULO
CAMPUS DE SÃO CARLOS
INSTITUTO DE CIÊNCIAS MATEMÁTICAS DE SÃO CARLOS
I.C.M.S.C.

A Bayesian approach in the detection of outliers for normal and weibull distri butions

ACHCAR, J.A.

nº 49

Notas do ICMSC - USP

ISSN 0103-2577

A Bayesian approach in the detection of outliers for normal and weibull distri butions

ACHCAR, J.A.

nº 49

São Carlos (SP)
1989

A BAYESIAN APPROACH IN THE DETECTION OF OUTLIERS FOR NORMAL AND WEIBULL DISTRIBUTIONS

JORGE ALBERTO ACHCAR
ICMSC - UNIVERSIDADE DE SÃO PAULO, C.Postal 668

13560 - SÃO CARLOS, SP., BRAZIL

SUMMARY

"The problem of detection of outliers is of great practical interest. In this paper, we present some Bayesian methods for detection of outliers for normal and Weibull distributions combining usual graphical procedures with posterior densities for prediction values and predictive densities for discordant observations".

Key words: Outliers, Normal distribution, Weibull distribution, posterior densities, predictive densities.

1. INTRODUCTION

An usual problem in data analysis is the detection of outliers in a random sample of a population with a specified probability distribution. Usually, the statisticians consider graphical methods to verify the adequability of a proposed model and often they observe one or more discordant observations that are natural candidates to be outliers. In this situation, the statistician uses hypothesis tests or confidence intervals to detect changes in location or scale of the discordant value (see for example, Barnett and Lewis, 1978). In this context, we also observe some Bayesian results in the literature (see for example, Box and Tiao, 1968; Abraham and Box, 1978; Guttman, Dutter and Freeman, 1978). Another Bayesian approach is explored by Pettit and Smith (1985) using predictive densities and influence measures based on Kullback-Leibler distances.

In this paper, we develop some Bayesian methods combining usual graphical methods, posterior densities and predictive densities for two special cases: normal distribution and Weibull distribution with two parameters.

2. DETECTION OF OUTLIERS IN NORMAL DISTRIBUTIONS

Let $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)'$ be a random sample of size n of a normal distribution $N\{\mu,\sigma^2\}$. In the verification of the adequability of the normal distribution the statistician usually considers plots of the ordered observations $x_{(i)}$, $i = 1,2,\ldots,n$ against the normal scores $z_{(i)} = \Phi_Z^{-1}\left(\frac{i}{n+1}\right)$ where Φ_Z denotes the distribution function of a random variable with a standard normal distribution.

Suppose that an observation $x_{(i)}$ is discordant and let $x^*_{(i)} = \mu + \sigma Z_{(i)}$ be the prediction value of $x_{(i)}$. Consider $\underset{\sim}{x}^c_{(i)}$ the vector of all observations x_j in $\underset{\sim}{x}$ such that $j \neq i$. From the data $x^c_{(i)}$ we find a posterior density for μ and σ and with $z_{(i)}$ fixed, we find the marginal posterior density for $x^*_{(i)} = \mu + \sigma Z_{(i)}$. Therefore, we have a criterion to decide if $x_{(i)}$ is an outlier by checking HPD intervals for $x^*_{(i)}$.

In the same way, we could have $m > 1$ discordant observations with the indexes of these observations in a set I. Thus, we could find the posterior density for each $x^*_{(i)}$, $i \in I$ based on the data $x^c_{(I)}$ (all x_j in $\underset{\sim}{x}$ such that $j \notin I$).

2.1. A BAYESIAN ANALYSIS WITH μ AND σ UNKNOWN

A noninformative prior for μ and σ (see Box and Tiao, 1973) is given by:

$$p(\mu,\sigma) \propto \frac{1}{\sigma} \tag{1}$$

where $-\infty < \mu < \infty$ and $\sigma > 0$.

Let $x_{(i)}$ be a discordant value. The joint posterior density for μ and σ based on the data $\underset{\sim}{x}^c_{(i)}$ and considering the prior (1) is given by:

$$p(\mu,\sigma|\underset{\sim}{x}^c_{(i)}) = K\,\sigma^{-n} \exp\{-\frac{1}{2\sigma^2}[\nu s^2_{(i)} + (n-1)(\mu - \bar{x}_{(i)})^2]\} \tag{2}$$

where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu = n-2$,

$$s^2_{(i)} = \nu^{-1} \sum_{u \neq i}^{n} (x_{(u)} - \bar{x}_{(i)})^2, \quad (n-1)\,\bar{x}_{(i)} = \sum_{u \neq i}^{n} x_{(u)}$$

and $K = \sqrt{\frac{n-1}{2\pi}} \left[\frac{1}{2}\Gamma(\frac{\nu}{2})\right]^{-1} (\frac{\nu s^2_{(i)}}{2})^{\nu/2}$ (see Box and Tiao, 1973).

Consider the transformation of variables $\sigma = \sigma$ and $x^*_{(i)} = \mu + \sigma z_{(i)}$ where $z_{(i)}$ is fixed. The Jacobian of this transformation is equal to one. Therefore, the joint posterior density for $x^*_{(i)}$ and σ is given by:

$$p(x^*_{(i)}, \sigma \mid \underset{\sim}{x}^c_{(i)}, z_{(i)}) = K\sigma^{-n} \exp\{-\frac{1}{2\sigma^2}[\nu s^2_{(i)} + (n-1)(x^*_{(i)} - \sigma z_{(i)} - \bar{x}_{(i)})^2\} \quad (3)$$

where $\sigma > 0$ and $-\infty < x^*_{(i)} < \infty$.

To find the marginal posterior density for $x^*_{(i)}$ we use the Laplace's method for integrals (see for example Tierney and Kadane, 1986).

From (3), we determine the marginal posterior density for $x^*_{(i)}$:

$$p(x^*_{(i)} \mid \underset{\sim}{x}^c_{(i)}, z_{(i)}) = K\int_0^\infty e^{-nh(\sigma)} d\sigma \quad (4)$$

where $-nh(\sigma) = -n\ell n\, \sigma - \frac{\nu s^2_{(i)}}{2\sigma^2} - \frac{(n-1)}{2\sigma^2}[x^*_{(i)} - \sigma z_{(i)} - \bar{x}_{(i)}]^2$.

The first derivative of $-nh(\sigma)$ is given by:

$$-nh'(\sigma) = -\frac{n}{\sigma} + \frac{\nu s^2_{(i)}}{\sigma^3} + \frac{(n-1)z_{(i)}(x^*_{(i)} - \sigma z_{(i)} - \bar{x}_{(i)})}{\sigma^2} + \frac{(n-1)(x^*_{(i)} - \sigma z_{(i)} - \bar{x}_{(i)})^2}{\sigma^3} \quad (5)$$

From $-nh'(\sigma) = 0$, we find the maximum $\hat{\sigma}$ that satisfies,

$$n = \frac{v s^2_{(i)}}{\hat{\sigma}^2} + \frac{(n-1) Z_{(i)} (x^*_{(i)} - \hat{\sigma} Z_{(i)} - \bar{x}_{(i)})}{\hat{\sigma}} + \\ + \frac{(n-1)\ (x^*_{(i)} - \hat{\sigma}\ Z_{(i)} - \bar{x}_{(i)})^2}{\hat{\sigma}^2} \qquad (6)$$

That is,

$$\hat{\sigma} = \frac{(n-1) Z_{(i)} \cdot (\bar{x}_{(i)} - x^*_{(i)}) + \sqrt{(n-1)^2 Z^2_{(i)} (\bar{x}_{(i)} - x^*_{(i)})^2 + 4nb(x^*_{(i)})}}{2n} \qquad (7)$$

where $b(x^*_{(i)}) = v\ s^2_{(i)} + (n-1)\ (x^*_{(i)} - \bar{x}_{(i)})^2$.

The second derivative of $-nh(\sigma)$ is given by:

$$-nh''(\sigma) = -\frac{1}{\sigma} \{ - \frac{n}{\sigma} + \frac{3}{\sigma} [\frac{v\ s^2_{(i)}}{\sigma^2} + \\ + \frac{(n-1) Z_{(i)}}{\sigma} (x^*_{(i)} - \sigma Z_{(i)} - \bar{x}_{(i)}) + \\ + \frac{(n-1)\ (x^*_{(i)} - \sigma Z_{(i)} - \bar{x}_{(i)})^2}{\sigma^2}] + \\ + \frac{(n-1)\ Z^2_{(i)}}{\sigma} + \frac{(n-1)\ Z_{(i)}}{\sigma^2} (x^*_{(i)} - \sigma Z_{(i)} - \bar{x}_{(i)}) \} \qquad (8)$$

From (6), we find:

$$-nh''(\hat{\sigma}) = -\frac{1}{\hat{\sigma}} \{ \frac{2n}{\hat{\sigma}} + \frac{(n-1)\ Z_{(i)} (x^*_{(i)} - \bar{x}_{(i)})}{\hat{\sigma}^2} \} \qquad (9)$$

Therefore, the marginal posterior density for $x^*_{(i)}$ approximated by the Laplace method is given by:

$$p(x^*_{(i)} \mid \underset{\sim}{x}^c_{(i)}, z_{(i)}) \propto \{\frac{2n}{\hat{\sigma}^2} + \frac{(n-1)\, z_{(i)}(x^*_{(i)} - \bar{x}_{(i)})}{\hat{\sigma}^3}\}^{-1/2} \cdot$$

$$\hat{\sigma}^{-n} \exp\{-\frac{1}{2\hat{\sigma}^2}[v\, s^2_{(i)} + (n-1)(x^*_{(i)} - \hat{\sigma} z_{(i)} - \bar{x}_{(i)})^2]\} \quad (10)$$

where $-\infty < x^*_{(i)} < \infty$, $\quad b(x^*_{(i)}) = v\, s^2_{(i)} +$

$$+ (n-1)\,(x^*_{(i)} - \bar{x}_{(i)})^2 \quad , \quad \text{and}$$

$$\hat{\sigma} = \frac{(n-1) z_{(i)}\,(\bar{x}_{(i)} - x^*_{(i)})}{2n} +$$

$$+ \frac{\sqrt{(n-1)^2 z^2_{(i)} (\bar{x}_{(i)} - x^*_{(i)})^2 + 4nb(x^*_{(i)})}}{2n}$$

2.2. A BAYESIAN ANALYSIS WITH σ KNOWN

With σ known, a noninformative prior density for μ is given by,

$$p(\mu) \propto \text{constant} \quad (11)$$

where $-\infty < \mu < \infty$.

With this locally uniform prior for μ , the posterior density for μ based on the data $\underset{\sim}{x}^c_{(i)}$ is given by:

$$p(\mu \mid \underset{\sim}{x}^c_{(i)}) = \frac{\sqrt{n-1}}{\sqrt{2\pi}\,\sigma} \exp\{-\frac{(n-1)}{2\sigma^2}(\mu - \bar{x}_{(i)})^2\} \quad (12)$$

where $-\infty < \mu < \infty$ (see Box and Tiao, 1973).

Thus, the posterior distribution for μ given $\underset{\sim}{x}^c_{(i)}$ is given by a normal density $N\{\bar{x}_{(i)} ; \frac{\sigma^2}{n-1}\}$.

Consider the transformation $x^*_{(i)} = \mu + \sigma Z_{(i)}$ (prediction value) where $Z_{(i)}$ is fixed. Thus, $\mu = x^*_{(i)} - \sigma Z_{(i)}$ and $d\mu/dx^*_{(i)} = 1$. Therefore, the posterior density for $x^*_{(i)}$ is given by,

$$p(x^*_{(i)} \mid \underset{\sim}{x}^c_{(i)} , Z_{(i)}) =$$

$$= \frac{\sqrt{n-1}}{\sqrt{2\pi}\,\sigma} \exp\{-\frac{(n-1)}{2\sigma^2}(x^*_{(i)} - \sigma Z_{(i)} - \bar{x}_{(i)})^2\} \qquad (13)$$

where $-\infty < x^*_{(i)} < \infty$.

That is, the posterior density for $x^*_{(i)}$ given $\underset{\sim}{x}^c_{(i)}$ and $Z_{(i)}$ is a normal density $N\{\sigma Z_{(i)} + \bar{x}_{(i)} ; \frac{\sigma^2}{n-1}\}$ where $(n-1)\bar{x}_{(i)} = \sum_{u \neq i}^{n} x_{(u)}$.

3. DETECTION OF OUTLIERS IN WEIBULL DISTRIBUTIONS

Let $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)'$ be a random sample of size n of an Weibull distribution with parameters β and p with density,

$$f(x|\beta,p) = p\beta(\beta x)^{p-1} \exp\{-(\beta x)^p\} \qquad (14)$$

where $x > 0$, $\beta > 0$ and $p > 0$.

The survival function $P\{X > x\}$ where x is a fixed value is given by,

$$P\{X > x\} = 1 - F(x|\beta,p) = \exp\{-(\beta x)^p\} \qquad (15)$$

where $F(x|\beta,p)$ denotes the distribution function of the random variable X with density (14).

From (15), we observe that,

$$\ell n\{-\ell n[1-F(x|\beta,p)]\} = p\,\ell n\beta + p\,\ell n\,x \qquad (16)$$

Thus, to verify the adequability of the Weibull distribution, the statistician considers plots of $\ell n\{-\ell n[1 - F_n(x_{(i)})]\}$ against $\ell n(x_{(i)})$ where $F_n(x_{(i)})$ is the empirical distribution function of the ordered data $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$. Thus, if there is an approximate linear relation, the statistician assumes that the Weibull model is adequate for the data analysis.

Suppose that an observation $x_{(i)}$ is discordant and let $\underset{\sim}{x}^c_{(i)}$ be the vector of all x_j in $\underset{\sim}{x}$ such that $j \neq i$.

From (15) and considering the empirical distribution function $F_n(x_{(i)})$, we have the prediction value $x^*_{(i)}$ given by:

$$x^*_{(i)} = \frac{1}{\beta}\,[-\ell n(1-F_n(x_{(i)}))]^{1/p} \qquad (17)$$

Therefore, we can verify if $x_{(i)}$ is really an outlier if we have a posterior density for $x^*_{(i)}$ given the data set $\underset{\sim}{x}^c_{(i)}$.

3.1. A BAYESIAN ANALYSIS WITH p KNOWN

With the data $\underset{\sim}{x}^c_{(i)}$ and with p known, the likelihood function for β is given by:

$$\ell(\beta|\underset{\sim}{x}^c_{(i)}, p) \propto \beta^{p(n-1)} \exp\{-\beta^p \sum_{u\neq i}^{n} x^p_{(u)}\} \qquad (18)$$

Considering a noninformative prior density for β, given by,

$$p(\beta) \propto \frac{1}{\beta}, \qquad (19)$$

the posterior density for β given $\underset{\sim}{x}^c_{(i)}$ is:

$$p(\beta|\underset{\sim}{x}^{c}_{(i)}\ ,\ p)\ \propto\ \beta^{p(n-1)-1}\ \exp\{-\beta^{p}\ \sum_{u\neq i}^{n} x^{p}_{(u)}\} \tag{20}$$

where $\beta > 0$.

With the transformation (17), we have

$$\beta = \{-\ell n\ [1 - F_n(x_{(i)})]\}^{1/p}/x^{*}_{(i)} \quad \text{and}$$

$$|d\beta/dx^{*}_{(i)}| = \{-\ell n[1 - F_n(x_{(i)})]\}^{1/p}/x^{*2}_{(i)}\ .$$

Thus, the posterior density for $x^{*}_{(i)}$ is:

$$p(x^{*}_{(i)}|\underset{\sim}{x}^{c}_{(i)}\ ,\ p) = \frac{p\{(\sum_{u\neq i}^{n} x^{p}_{(u)})(-\ell n[1-F_n(x_{(i)})])\}^{n-1}}{\Gamma(n-1)\ \ x^{*p(n-1)+1}_{(i)}}\ .$$

$$\exp\{-\ \frac{(\sum_{u\neq i}^{n} x^{p}_{(u)})(-\ell n[1 - F_n(x_{(i)})])}{x^{*p}_{(i)}}\ \} \tag{21}$$

where $x^{*}_{(i)} > 0$.

The mode of this posterior density is given by:

$$\hat{x}^{*}_{(i)} = \{\frac{p(\sum_{u\neq i}^{n} x^{p}_{(u)})(-\ell n\ [1 - F_n(x_{(i)})])}{p(n-1)+1}\}^{1/p} \tag{22}$$

3.2. USE OF PREDICTIVE DENSITIES

Considering p known, let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics for the observed data. Suppose that $X_{(i)}$ is a discordant observation. The predictive density for the order statistics $X_{(i)}$ given $\underset{\sim}{x}^{c}_{(i)}$ (all observations x_j in $\underset{\sim}{x}$ with $j\neq i$) is given by:

$$p(x_{(i)}|\underset{\sim}{x}^{c}_{(i)}) = \int p(x_{(i)}|\beta)\ p(\beta|\underset{\sim}{x}^{c}_{(i)}\ ,\ p)d\beta \qquad (23)$$

where $x_{(i)} > 0$, $p(x_{(i)}|\beta)$ is the marginal distribution for the order statistics $X_{(i)}$ and $p(\beta|\underset{\sim}{x}^{c}_{(i)}\ ,\ p)$ is the posterior density for β given $x^{c}_{(i)}$ (see for example, Aitchison and Dunsmore, 1975).

With the noninformative prior (19), we have the posterior density (20) for β,

$$p(\beta|\underset{\sim}{x}^{c}_{(i)},\ p) = \frac{p(\sum_{u\neq i}^{n} x^{p}_{(u)})^{n-1}\ \beta^{p(n-1)-1}}{\Gamma(n-1)}\ .$$

$$\exp\{-\beta^{p} \sum_{u\neq i}^{n} x^{p}_{(u)}\} \qquad (24)$$

where $\beta > 0$.

The marginal density for the order statistics $X_{(i)}$ is given by,

$$p(x_{(i)}|\beta) = \frac{n!}{(i-1)!(n-i)!}\ \{F(x_{(i)}|\beta)\}^{i-1}\ .$$

$$\{1- F(x_{(i)}|\beta)\}^{n-i}\ f(x_{(i)}|\beta) \qquad (25)$$

Thus, with the Weibull distribution with shape parameter p known, we have (from (14) and (15)):

$$p(x_{(i)}|\beta) = \frac{n!\ p\ \beta^{p}\ x^{p-1}_{(i)}}{(i-1)!(n-i)!}\ \exp\{-(n-i+1)\beta^{p}\ x^{p}_{(i)}\}$$

$$\sum_{k=0}^{i-1} \binom{i-1}{k} (-1)^{k} \exp\{-K\ \beta^{p}\ x^{p}_{(i)}\} \qquad (26)$$

where $x_{(i)} > 0$.

Thus, the predictive density (23) is given by:

$$p(x_{(i)} \mid \underset{\sim}{x}^{c}_{(i)}) = \frac{n!\; p^{2}\; x_{(i)}^{p-1}\; (\sum_{u\neq i}^{n} x_{(u)}^{p})^{n-1}}{(i-1)!\;(n-i)!\;\Gamma(n-1)} \cdot$$

$$\sum_{k=0}^{i-1} \binom{i-1}{k} (-1)^{k} \int_{0}^{\infty} \beta^{pn-1} \exp\{-[(n-i+1+k)\, x_{(i)}^{p} +$$

$$+ \sum_{u\neq i}^{n} x_{(u)}^{p}]\, \beta^{p}\}\, d\beta$$

That is,

$$p(x_{(i)} \mid \underset{\sim}{x}^{c}_{(i)}) = \frac{n!\; p(n-1)\; x_{(i)}^{p-1}\; (\sum_{u\neq i}^{n} x_{(u)}^{p})^{n-1}}{(i-1)!\;(n-i)!} \cdot$$

$$\sum_{k=0}^{i-1} \binom{i-1}{k} \cdot \cdot \frac{(-1)^{k}}{((n-i+1+k)x_{(i)}^{p} + \sum_{u\neq i}^{n} x_{(u)}^{p})^{n}} \qquad (27)$$

where $x_{(i)} > 0$.

With the predictive density (27), we can verify if a discordant observation $x_{(i)OBS}$ is an outlier by calculating $P\{X_{(i)} > x_{(i)OBS}\} = 1 - P\{X_{(i)} \leq x_{(i)OBS}\}$, where,

$$P\{X_{(i)} \leq x_{(i)OBS}\} = \frac{n!\; p(n-1)\; (\sum_{u\neq i}^{n} x_{(u)}^{p})^{n-1}}{(i-1)!\;(n-i)!} \cdot$$

$$\cdot \sum_{k=0}^{i-1} \binom{i-1}{k} (-1)^{k} \int_{0}^{x_{(i)OBS}} \frac{x_{(i)}^{p-1}\; dx_{(i)}}{((n-i+1+k)\, x_{(i)}^{p} + \sum_{u\neq i}^{n} x_{(u)}^{p})^{n}}$$

That is,

$$P\{X_{(i)} \leq x_{(i)OBS}\} = \frac{n!}{(i-1)!\,(n-i)!} \sum_{k=0}^{i-1} \binom{i-1}{k} \frac{(-1)^k}{(n-i+1+k)}$$

$$\cdot \left\{1 - \left[\frac{\sum_{u \neq i}^{n} x_{(u)}^{p}}{(n-i+1+k)\, x_{(i)OBS}^{p} + \sum_{u \neq i}^{n} x_{(u)}^{p}}\right]^{n-1}\right\} \qquad (28)$$

4. EXAMPLES

4.1. AN EXAMPLE WITH THE NORMAL DISTRIBUTION

In table 1 we have the residuals of a fitted linear regression. To check normality, the statistician considers the plot of the ordered residuals against the normal scores. We observe in figure 1 that the ordered residual $r_{(17)} = 3.70795$ is clearly a discordant value of the assumption of normality. Associated to $r_{(17)}$ we have the normal score $z_{(17)} = 1.7990$.

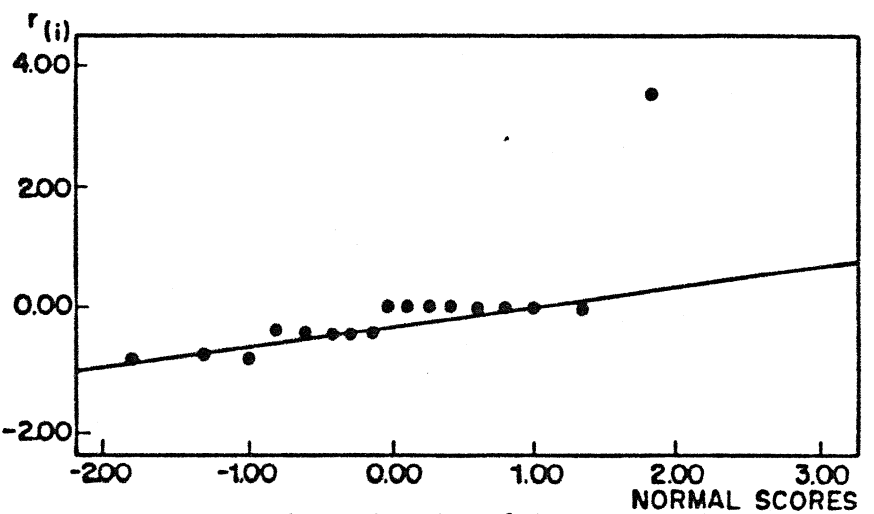

Figure 1 - *Normal Scores*

i	$r_{(i)}$	$z_{(i)}$
1	-0.90712	-1.79899
2	-0.72471	-1.31509
3	-0.70499	-1.02550
4	-0.42345	-0.80379
5	-0.39549	-0.61643
6	-0.25781	-0.44898
7	-0.23049	-0.29359
8	-0-20659	-0.14518
9	-0.19918	0.00000
10	-0.17080	0.14518
11	-0.11557	0.29359
12	0.00417	0.44898
13	0.05868	0.61643
14	0.09826	0.80379
15	0.14332	1.02550
16	0.14563	1.31509
17	3.70795	1.79899

TABLE 1 - *Residuals and Normal Scores*

Let $\underset{\sim}{r}^{c}_{(17)}$ be the vector of all residuals less the largest one $r_{(17)} = 3.70795$. Thus, $\bar{r}_{(17)} = -0.24288$ and $s^2_{(17)} = 0.102415$ (see notation in (2)). With $z_{(17)} = 1.8$, $n = 17$, $v = n-2 = 15$, we have (from (10)) the marginal posterior density for the prediction value $r^*_{(17)}$ approximated by Laplace's method given by:

$$p(r^*_{(17)} | \underset{\sim}{r}^{c}_{(17)}, z_{(17)}) \propto \left(\frac{34}{\hat{\sigma}^2} + \frac{28.8\,(r^*_{(17)} + 0.2429)}{\hat{\sigma}^3}\right)^{-1/2}$$

$$\hat{\sigma}^{-17} \exp\left(-\frac{1}{2\hat{\sigma}^2}\left[1.5361 + 16(r^*_{(17)} - 1.8\,\hat{\sigma} + 0.2429)^2\right]\right) \qquad (29)$$

where $-\infty < r^*_{(17)} < \infty$,

$b(r^*_{(17)}) = 1.5361 + 16\,(r^*_{(17)} + 0.24288)^2$,

and,

$$\hat{\sigma} = \frac{28.8\,(-0.24288 - r^*_{(17)})}{34} + \frac{\sqrt{829.44\,(-0.24288 - r^*_{(17)})^2 + 68\,b(r^*_{(17)})}}{34}$$

In figure 2 we have the plot of the posterior density (29). Clearly, $r_{(17)} = 3.70795$ is not inside any HPD interval for $r^*_{(17)}$, that is, $r_{(17)}$ is an outlier.

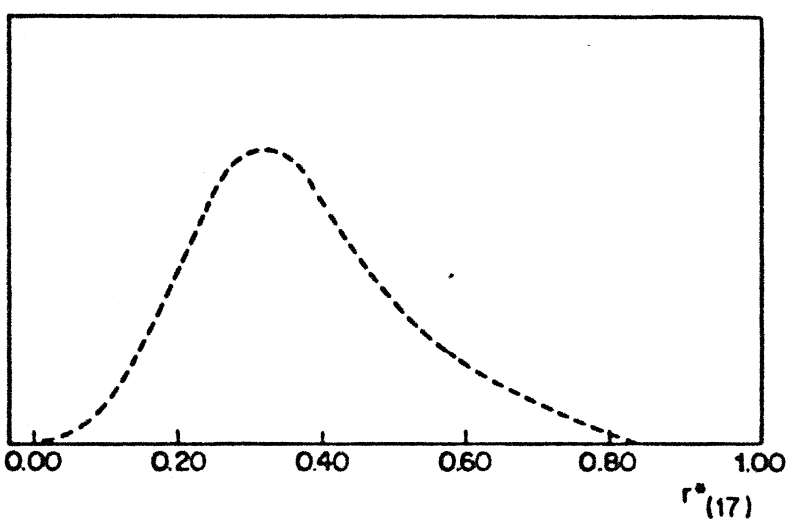

Figure 2 - *Posterior Density* $p(r^*_{(17)} | \underset{\sim}{r}^c_{(17)}, z_{(17)})$

(σ and μ Unknown)

Assuming $\sigma^2 = 0.10$ known, we have in figure 3 the plot of the posterior density for $r^*_{(17)}$ given $\underset{\sim}{r}^c_{(17)}$ (from (13)).

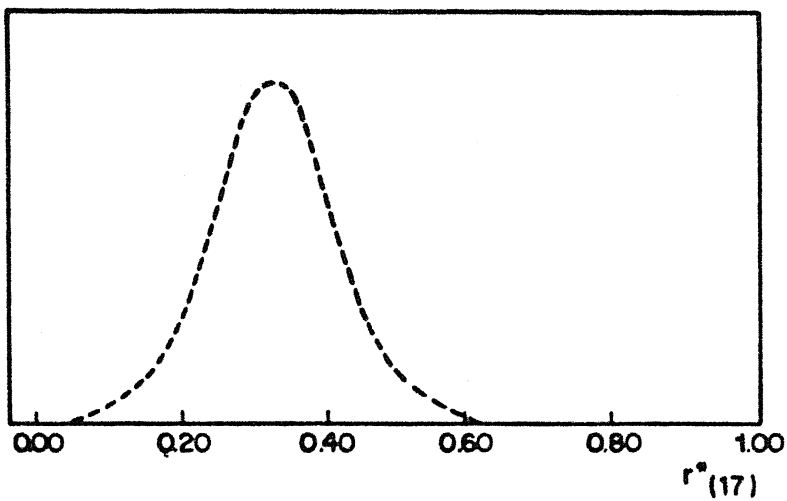

Figure 3 - *Posterior Density* $p(r^*_{(17)} | \underset{\sim}{r}^c_{(17)}, z_{(17)})$ *(With* $\sigma^2 = 0.10$ *Known)*

Since $r^*_{(17)} | \underset{\sim}{r}^c_{(17)}, z_{(17)} \sim N(\sigma z_{(17)} + \bar{r}_{(17)} ; \frac{\sigma^2}{n-1})$, a 95% HPD interval for $r^*_{(17)}$ (with $\sigma = 0.32$) is given by (0.17632; 0.48992), that is, since $r_{(17)} = 3.70795$ is not included in this interval, we conclude that $r_{(17)} = 3.70795$ is an outlier.

Taking out $r_{(17)} = 3.70795$ (an outlier), we can find 95% HPD intervals for each prediction value $r^*_{(i)}$, $i = 1,2,\ldots,16$ to check for other outliers. Let us assume $\sigma^2 = 0.10$ known. In table 2, we have the normal scores for the 16 remainder ordered residuals and in figure 4 we have the plot of the residuals against the normal scores. We observe an approximate straight line indicating the normality of the residuals, but for some

points (specially the residual $r_{(3)} = -0.70499$) we should check carefully against outliers.

i	$r_{(i)}$	$z_{(i)}$
1	-0.90712	-1.77150
2	-0.72471	-1.28113
3	-0.70499	-0.98637
4	-0.42345	-0.75968
5	-0.39549	-0.56716
6	-0.25781	-0.39413
7	-0.23049	-0.23247
8	-0.20659	-0.07686
9	-0.19918	0.07686
10	-0.17080	0.23247
11	-0.11557	0.39413
12	0.00417	0.56716
13	0.05868	0.75968
14	0.09826	0.98637
15	0.14332	1.28113
16	0.14563	1.77150

TABLE 2 - *Normal Scores for the Ordered Residuals ($r_{(17)} = 3.70795$ not Included)*

In table 3, we have HPD intervals for each prediction value $r^*_{(i)}$, $i = 1,2, \ldots, 16$ assuming $\sigma = 0.32$ known. We observe that the residual $r_{(3)} = -0.70499$ is not included in the 95% HPD interval for the prediction value $r^*_{(3)}$ that is, we can consider $r_{(3)}$ as an outlier.

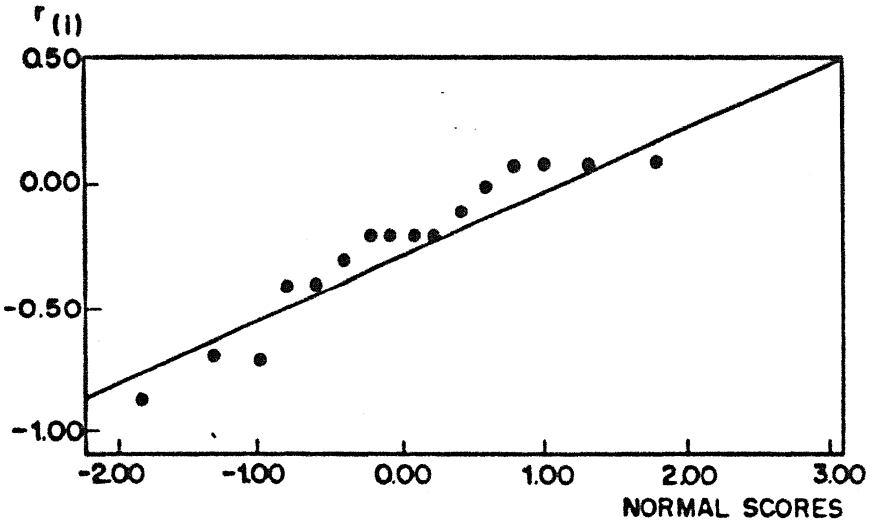

Figure 4 - *Normal Scores*

($r_{(17)}$ = 3.70795 Not Included)

i	$r_{(i)}$	$z_{(i)}$	$\bar{r}_{(i)}$	95% HPD INTERVAL FOR $r^*_{(i)}$
1	-0.90712	-1.77150	-0.19860	(-0.92742; -0.60354)
2	-0.72471	-1.28113	-0.21076	(-0.78266; -0.45878)
3	-0.70499(*)	-0.98637	-0.21207	(-0.68965; -0.36577)
4	-0.42345	-0.75968	-0.23085	(-0.63589; -0.31201)
5	-0.39549	-0.56716	-0.23271	(-0.57614; -0.25226)
6	-0.25781	-0.39413	-0.24189	(-0.52995; -0.20607)
7	-0.23049	-0.23247	-0.24371	(-0.48004; -0.15616)
8	-0.20659	-0.07686	-0.24530	(-0.43183; -0.10795)
9	-0.19918	0.07686	-0.24580	(-0.38314; -0.05926)
10	-0.17080	0.23247	-0.24769	(-0.33524; -0.01136)
11	-0.11557	0.39413	-0.25137	(-0.28719; 0.03669)
12	0.00417	0.56716	-0.25933	(-0.23980; 0.08408)
13	0.05868	0.75968	-0.26299	(-0.18183; 0.14205)
14	0.09826	0.98637	-0.26563	(-0.11194; 0.21194)
15	0.14332	1.28113	-0.26863	(-0.02061; 0.30327)
16	0.14563	1.77150	-0.26879	(0.13615; 0.46003)

TABLE 3 - *95% HPD Intervals for $r^*_{(i)}$, i = 1,2,..., 16 with σ = 0.32 Known (Residual $r_{(17)}$ = 3.70795 not Included)*

4.2. AN EXAMPLE WITH THE WEIBULL DISTRIBUTION

In table 4, we have the life times of 15 units submitted to a laboratory experiment.

4.83	5.17	6.75	7.98	8.16
8.44	8.54	8.58	8.69	8.97
9.91	11.91	11.93	13.16	13.87

TABLE 4 - *Life Times of 15 Units (Hours/100)*

In figure 5, we have the plot of $\ell n\{-\ell n[1 - F_n(x_{(i)})]\}$ versus $\ell n(x_{(i)})$ where $F_n(x_{(i)})$ is the empirical distribution function. From figure 5, we conclude that there is an approximate straight line indicating the adequability of the Weibull distribution, but some points are natural candidates to be outliers.

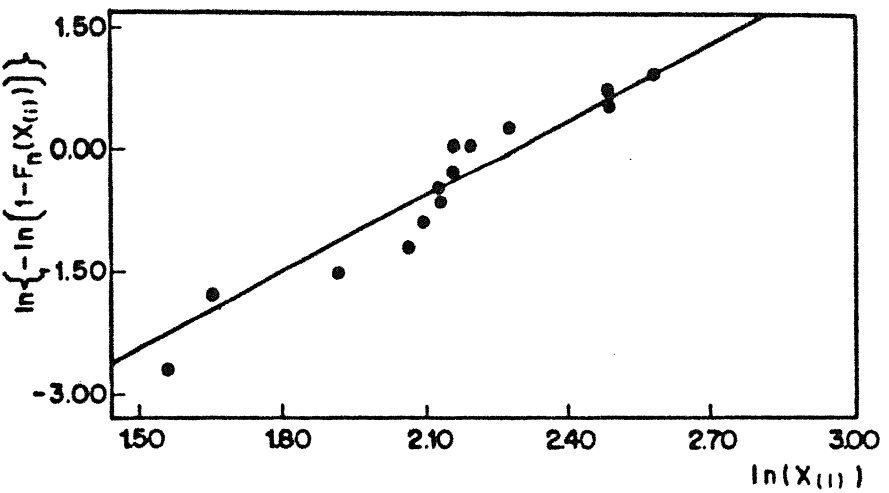

Figure 5 - *Plot of* $\ell n\{-\ell n[1 - F_n(x_{(i)})]\}$ *Versus* $\ell n(x_{(i)})$

Considering p = 2 known (from the least squares estimators of p and β in (16)), and considering observations $x_{(1)}$, $x_{(3)}$ and $x_{(4)}$ as discordant values (see figure 5), the posterior densities for $x^*_{(1)}$, $x^*_{(3)}$ and $x^*_{(4)}$ (given in (21)) are given by:

$$p(x^*_{(1)} | \underset{\sim}{x}^c_{(1)}, p = 2) = \frac{(9.0705)(10^{17})}{x^{*29}_{(1)}} \exp\left(-\frac{91.3645}{x^{*2}_{(1)}}\right)$$

$$p(x^*_{(3)} | \underset{\sim}{x}^c_{(3)}, p = 2) = \frac{(9.8090)(10^{24})}{x^{*29}_{(3)}} \exp\left(-\frac{290.5396}{x^{*2}_{(3)}}\right)$$

$$p(x^*_{(4)} | \underset{\sim}{x}^c_{(4)},\ p = 2) = \frac{(8.0972)(10^{26})}{x^{*29}_{(4)}} \exp\left(-\frac{398.2110}{x^{*2}_{(4)}}\right)$$

where $x^*_{(1)},\ x^*_{(3)},\ x^*_{(4)} > 0$.

In figures 6, 7 and 8 we have the plots of these posterior densities. We can conclude that $x_{(1)} = 4.83$, $x_{(3)} = 6.75$ and $x_{(4)} = 7.98$ are candidates to be outliers since they are not included in the usual HPD intervals for $x^*_{(i)}$, $i = 1,\ 3$ and 4, respectively.

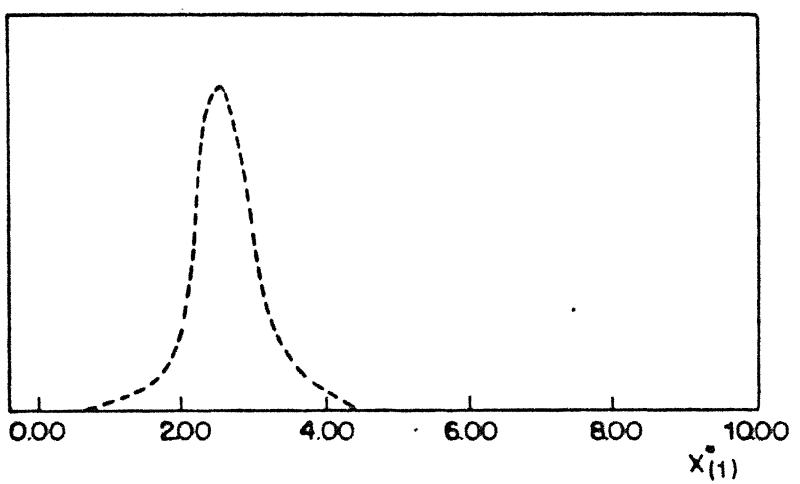

Figure 6 - *Posterior Density* $p(x^*_{(1)} | \underset{\sim}{x}^c_{(1)},\ p = 2)$

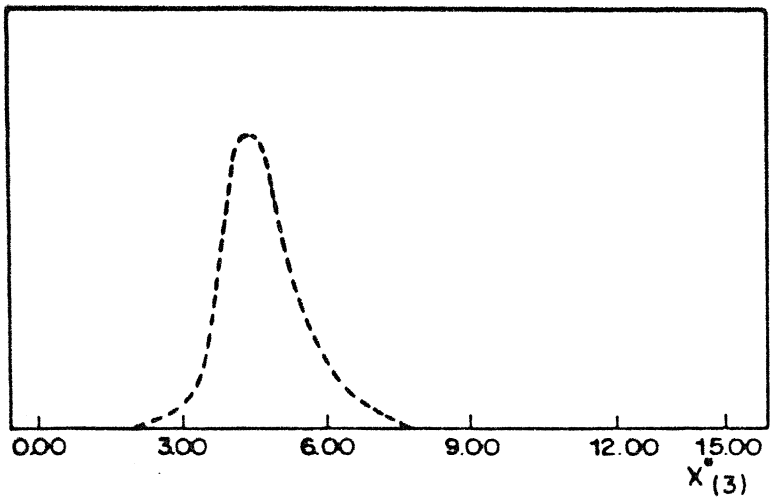

Figure 7 - Posterior Density $p(x^*_{(3)} | \underset{\sim}{x}^c_{(3)}, p = 2)$

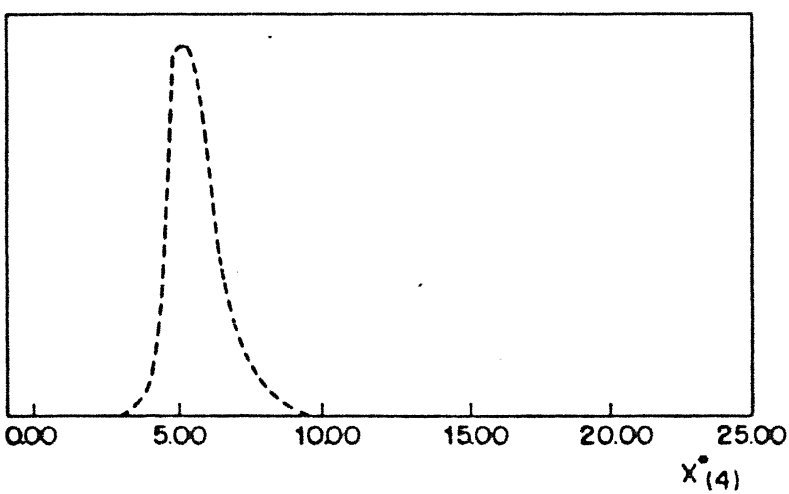

Figure 8 - Posterior Density $p(x^*_{(4)} | \underset{\sim}{x}^c_{(4)}, p = 2)$

From (21), we can verify that $2\left(\sum_{u\neq i}^{n} x_{(u)}^{p}\right)\left[-\ell n(1 - F_n(x_{(i)}))\right]/x_{(i)}^{*p} \sim \chi^2_{2(n-1)}$. Therefore, a $100(1-\alpha)\%$ HPD interval for $x^*_{(i)}$ is given by,

$$\left(\left(2\left(\sum_{u\neq i}^{n} x_{(u)}^{p}\right)\left[-\ell n(1 - F_n(x_{(i)}))\right]/\chi^2_{2(n-1)}\ (\alpha/2)\right)^{1/p}\ ;\right.$$

$$\left.\left(2\left(\sum_{u\neq i}^{n} x_{(u)}^{p}\right)\left[-\ell n(1 - F_n(x_{(i)}))\right]/\chi^2_{2(n-1)}\ (1-\alpha/2)\right)^{1/p}\right),$$

where $\chi^2_{2(n-1)}\ (1-\alpha/2)$ is the $100(\alpha/2)$ percentile of a $\chi^2_{2(n-1)}$ distribution.

In table 5, we have 98% HPD intervals for $x^*_{(i)}$, $i = 1,2,\ldots,15$ considering $p = 2$ known. We conclude that $x_{(1)}$, $x_{(3)}$ and $x_{(4)}$ are outliers. One weakness of this approach is that we cannot check if the last ordered observation $x_{(n)}$ is an outlier since $F_n(x_{(n)}) = 1$. The method can be improved if we define $F_n(x_{(i)}) = (\text{number of observations} \leq x_{(i)})/n$ for $i = 1,2,\ldots,n-1$ and $F_n(x_{(i)}) = (n-1)/n$. Therefore, we have $F_n(x_{(n)}) < 1$.

Another possibility is to use the predictive density $p(x_{(i)}|x^c_{(i)})$ given in (27). In the last column of table 5, we have the values of $P(X_{(i)} > x_{(i)OBS})$ (from (28)) for $i = 1,2,\ldots, 15$. We also conclude that $x_{(1)}$, $x_{(3)}$ and $x_{(4)}$ are outliers since $P(X_{(i)} > x_{(i)OBS})$ are very small (< 0.07) for $i = 1$, 3 and 4.

i	$x_{(i)}$	$F_n(x_{(i)})$	$\sum_{u \neq i}^{n} x_u^2$	98% HPD INTERVAL FOR $x^*_{(i)}$	$P(X_{(i)} > x_{(i)OBS})$
1	4.83(*)	1/15	1324.26	(1.9450; 3.6655)	0.0375
2	5.17	2/15	1320.86	(2.7976; 5.2722)	0.1143
3	6.75(*)	3/15	1302.03	(3.4685; 6.5365)	0.0645
4	7.98(*)	4/15	1283.91	(4.0607; 7.6525)	0.0445
5	8.16	5/15	1281.00	(4.6376; 8.7397)	0.0718
6	8.44	6/15	1276.36	(5.1959; 9.7919)	0.1983
7	8.54	7/15	1274.66	(5.7601;10.8551)	0.2532
8	8.58	8/15	1273.97	(6.3407;11.9493)	0.3918
9	8.69	9/15	1272.07	(6.9473;13.0923)	0.5288
10	8.97	10/15	1267.13	(7.5923;14.3080)	0.6301
11	9.91	11/15	1249.38	(8.2692;15.5836)	0.5862
12	11.91	12/15	1205.74	(8.9641;16.8931)	0.3913
13	11.93	13/15	1205.26	(10.0278;18.8979)	0.5806
14	13.16	14/15	1174.40	(11.4756;21.6263)	0.5924
15	13.87	1	1155.21	———	0.7593

TABLE 5 - *98% HPD Intervals for $x^*_{(i)}$, $i = 1,2,\ldots,15$ with $p = 2$ and $P(X_{(i)} > x_{(i)OBS})$ Using the Predictive Density $p(x_{(i)}|\underset{\sim}{x}^{c}_{(i)})$.*

REFERENCES

ABRAHAM,B.; BOX, G.E.P. (1978). Linear Models and spurious observations. Applied Statistics, 27, 120-130.

AITCHISON,J.; DUNSMORE,I.R. (1975). Statistical Prediction Analysis, Cambridge University Press.

BARNETT,V.; LEWIS,T. (1978). Outliers in Statistical Data, Chichester: Wiley.

BOX,G.E.P.; TIAO,G.C. (1968). A Bayesian approach to some outlier problems, Biometrika, 55, 119-129.

BOX,G.E.P.; TIAO,G.C. (1973). Bayesian Inference in Statistical Analysis, Massachusetts: Addison-Wesley.

GUTTMAN,T.; DUTTER,R.; FREEMAN,P.R. (1978). Care and handling of univariate outliers in the general linear model to detect spurosity - a Bayesinan Approach, Technometrics, 20, 187-193.

PETTIT,L.I.; SMITH, A.F.M. (1985). Outliers and Influential Observations in Linear Models, Bayesian Statistics 2, (Bernardo J.M., et al eds.), 473-494, Elsevier Science Publishers B.V. (North-Holland).

TIERNEY,L.; KADANE,J.B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. Journal of the American Statistical Association, 81, 82-86.