INFERENCES FOR INCOMPLETELY SPECIFIED MODELS IN THE THEORY OF OUTLIERS

BY

R. KHOT AND R.C. JAIN

School of Studies in Statistics, Vikram University, Ujjain

(Received : August, 1978)

SUMMARY

The problem of an outlier by applying the theory of incompletely specified model in the sense of Bancraft (1964) has been discussed. The bias and mean square error of the Preliminary test estimator for testing the outlying observation have been studied. The mean square error of this estimator is compared with that of the usual unbiased estimator in the two cases according to whether the largest or the smallest observation is an outlier.

INTRODUCTION

There could be two approaches to the problem of outlying observations depending upon the interest of the scientist which may be either in testing whether a particular observation is an outlier or alternatively in obtaining a more accurate estimate of a population parameter by retaining or discarding the anamolous observation after this test. In the former case, the test for an outlier would be an end itself but in the latter case if would constitute a preliminary step for estimation of a population parameter subsequent to the outlier test. In like manner, in outlier test would also constitute a preliminary step for testing a hypothesis about a population parameter. In both these cases, the test for an outlying observation can be termed as a preliminary test. In such cases then the power and size of the subsequent test will also be important. We have discussed in this paper the problem of an outlier by applying the theory of incompletely specified model in the sense of Bancraft (1964). Further, in this problem, we study the bias and mean square error of the 'Preliminary test estimator' obtained with the help of McKay's test (1935) for testing the outlying observation. The mean square error

(MSE) of this estimator is then compared with that of the usual unbiased estimator in the two cases according to whether the largest or the smallest observation is an outlier.

STATEMENT OF THE PROBLEM

Let $X_1, X_2, \ldots, X_n$ ($n>3$) be an ordered sample in which X_n or X_1 is an outlying observation for which we assume that either of the observations $X_1, X_2 \ldots, X_{n-1}$ or $X_2, X_3 \ldots, X_n$ constitute a random sample of size $n-1$ from $N(\mu_1, \sigma^2)$ and that X_n or X_1 is a random sample of one from $N(\mu_2, \sigma^2)$. If $\mu_1 \neq \mu_2$, then X_n or X_1 belongs to a universe different from that generating the other $n-1$ observations and as such X_n or X_1 will be termed an outlier. Out problem is to estimate μ_1. For this, we first test the hypothesis $H_o: \mu_1=\mu_2$ against $\mu_1 \neq \mu_2$ with the help of modified form of McKay's test (1935) which when used this way can be referred to as a preliminary test in sense of Bancraft (1964).

For large samples, the possibility of having more than one outlier needs to be considered. If an observation is an outlier, consider the remaining observations as a sample of size $n-1$ and according to Ans Combe (1960) the procedure discussed above can again be applied; and so on. The estimate of μ_1 will be the mean of the retained observations.

RULE OF PROCEDURE

For an ordered sample $X_1, X_2, \ldots, X_n$ of size n from a normal population with known variance σ^2, let

$$\bar{X}_{n-1}=\frac{1}{n-1}\sum_{i=1}^{n-1} x_i,\ \bar{X}'_{n-1}=\frac{1}{n-1}\sum_{i=2}^{n} Xi$$

and

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n} Xi=\frac{1}{n}[(n-1)\bar{X}_{n-1}+X_n]$$

$$=\frac{1}{n}[(n-1)\bar{X}^1{}_{n-1}+X_1].$$

Depending then on whether X_n or X_1 is an outlier, we define a random variable Z as $Z=X_n-\bar{X}_{n-1}$ if X_n is suspected to be an outlier or as $Z=\bar{X}'_{n-1}-X_1$ if X_1 is suspected to be an outlier.

For a pre-assigned significance level α and a critical value of the statistic Z corresponding to this significance level if $|Z| \geqslant \xi_\alpha$, then X_n or X_1 will be considered an outlier and in that case $\bar{X}_{n-1}$ or

$\bar{X}'_{n-1}$ is then used as an estimate of μ_1. Alternatively if $|Z| < \xi_\alpha$, then X_n or X_1 will not be considered an outlier and in that case $\bar{X}$ is then used as an estimate of μ_1.

The estimation procedure based on an incompletely specified model calls for determining the bias and the mean square error of the preliminary estimator $\bar{X}^*$ where

$$\bar{X}^* = \begin{cases} \bar{X}_{n-1} \text{ if } X_n \text{ is an outlier} \\ \bar{X}'_{n-1} \text{ if } X_1 \text{ is an outlier} \\ \bar{X} \text{ if neither } X_n \text{ nor } X_1 \text{ is an outlier.} \end{cases}$$

The bias and MSE of $\bar{X}^*$ are now derived for the two cases depending on whether X_n or X_1 is an outlier.

Case I. (When X_n is an outlier)

Let $\triangle = \mu_2 - \mu_1$, $\sigma_Z^2 = \dfrac{n\,\sigma^2}{n-1}$ and $\sigma_{\bar{X}}^2 = \dfrac{\sigma^2}{n}$.

Since. $Z = X_n - X_{n-1}$, the assumption about the distributions of X_n and $\bar{X}_{n-1}$ imply that Z is distributed as $N(\triangle, \sigma_Z^2$ and

$$\bar{X} \text{ as } N\left[\frac{(n-1)\,\mu_1 + \mu_2}{n}, \sigma_{\bar{X}}^2\right]$$

and further that Z and $\bar{X}$ are independent.

The expected value of $\bar{X}^*$ is now given by

$$E(\bar{X})^* = E[\bar{X}_{n-1}, |Z| \geqslant \xi_\alpha]\, P[|Z| \geqslant \xi_\alpha] + E[\bar{X}, |Z| < \xi_\alpha]\, P[|Z| < \xi_\alpha]. \quad \ldots(1)$$

Since Z and $\bar{X}$ are independently distributed, their joint density can be writen as

$$f(Z, \bar{X}) = \frac{1}{\sigma_Z \sqrt{2\pi}}\, e^{-1/2\left[\frac{Z-\triangle}{\sigma_Z}\right]^2},$$

$$\frac{1}{\sigma_{\bar{X}} \sqrt{2\pi}}\, e^{-1/2\,\sigma_{\bar{X}}^2\left[\bar{X} - \frac{(n-1)\,\mu_1+\mu_2}{n}\right]^2} \quad \ldots(2)$$

The expressions for the second component of (1) can easily be written as

$$E[\bar{X}, |Z| < \xi_\alpha] = \frac{(n-1)\mu_1 + \mu_2}{nP[|Z| < \xi_\alpha]}[\Phi(\delta+\xi_\alpha) - \Phi(\delta-\xi_\alpha)]$$

where

$$\delta = \frac{\triangle}{\sigma_Z} = \frac{\triangle\sqrt{n-1}}{\sigma\sqrt{n}}$$

In order to evaluate the first term of (1) we consider the distribution of Z and $\bar{X}_{n-1}$ which is known to be a bivariate normal with $\rho=-n^{-\frac{1}{2}}$: Evaluating the integrals, we obtain

$$E[\bar{X}_{n-1}, |Z| \geqslant \xi_\alpha]=\frac{1}{P[|Z| \geqslant \xi_\alpha]}[\mu_1-\mu_1\{\Phi(\delta+\xi_\alpha)-\Phi(\delta-\xi_\alpha)\}$$

$$-\frac{\rho\sigma}{\sqrt{n-1}}\{\Phi(\delta+\xi_\alpha)-\Phi(\delta-\xi_\alpha)\}] \quad \ldots(3)$$

combining (2) and (3) and subtracting 1, the bias of $\bar{X}^*$ expressed as a fraction of σ is given by

$$\frac{\text{Bias}}{\sigma}=\frac{\delta}{\sqrt{n(n-1)}}[\Phi(\delta+\xi_\alpha)-\Phi(\delta-\xi_\alpha)]$$

$$-\frac{\rho}{\sqrt{n-1}}+[\Phi(\delta+\xi_\alpha)-\Phi(\delta-\xi_\alpha)] \quad \ldots(4)$$

The mean square error of $\bar{X}^*$ is given by

$$MSE(\bar{X}^*)=\text{Var}(\bar{X}^*)+(\text{Bias})^2.$$

In order to calculate Var $(\bar{X}^*)$, we have to obtain the expressions for $E\,(\bar{X}^{*2})$ and the same is given by

$$E(\overset{*}{\bar{X}}{}^2)=I\,[(\bar{X}^2_{n-1}, |Z| \geqslant \xi_\alpha].\ P[|Z| \geqslant \xi_\alpha]+E[\overset{*}{\bar{X}}{}^2, |Z| < \xi_\alpha].\ P[|Z| < \xi_\alpha]$$

Proceeding as above, $E(\overset{*}{\bar{X}}{}^2)$ can be evaluated and the final expression for MSE $(\bar{X}^*)$ expressed as fraction of σ^2 is given by

$$\frac{\text{MSE}(\bar{X}^*)}{\sigma^2}=\frac{1}{n-1}+\frac{1}{n(n-1)}[\delta^2\{\Phi(\delta+\xi_\alpha)-\Phi(\delta-\xi_\alpha)\}-\{\Phi(\delta+\xi_\alpha)-\Phi$$
$$(\delta-\xi_\alpha)-(\xi+\xi_\alpha)\Phi(\delta+\xi_\alpha)+(\delta-\xi_\alpha)\Phi(\delta-\xi_\alpha)\}] \quad \ldots(5)$$

Case II. (When X_1 is an outlier)

The expected value of $\bar{X}^*$ is given by

$$E(\bar{X}^*)=[E\bar{X}'_{n-1}, |Z| \geqslant \xi_\alpha]\,P[|Z| \geqslant \xi_\alpha]+E[\bar{X}, |Z| < \xi_\alpha].P[|Z| < \xi_\alpha].$$

Proceeding as in case I, we obtained the expression for $E(\bar{X}^*)$ and the bias expressed as a fraction of σ is given by

$$\frac{\text{Bias}}{\sigma}=\frac{\delta}{\sqrt{n(n-1)}}[\Phi(\delta+\xi_\alpha)-\Phi(\delta-\xi_\alpha)]$$

$$-\frac{\rho}{\sqrt{n-1}}[\Phi(\delta+\xi_\alpha)-\Phi(\delta-\xi_\alpha)] \quad \ldots(6)$$

where $\rho=\frac{1}{\sqrt{n}}$ and $\triangle$ and δ are same as defined in Case I. The MSE$(\bar{X}^*)$ expressed as a fraction of σ^2 is given by

$$\frac{\text{MSE}(\bar{X}^*)}{\sigma^2}=\frac{1}{n-1}+\frac{1}{n(n-1)}[\delta^2\{\Phi(\delta+\xi_\alpha)-\Phi(\delta-\xi_\alpha)\}-\{\Phi(\delta+\xi_\alpha)$$
$$-\Phi(\delta-\xi_\alpha)-(\delta+\xi_\alpha)\Phi(\delta+\xi_\alpha)+(\delta-\xi_\alpha)\Phi(\delta-\xi_\alpha)\}] \quad \ldots(7)$$

which is same as (5) of Case I.

Discussion of Results

Here we discuss the results of bias and mean square error of the preliminary test estimator under case 1 and 2. From (4), (5), (6), and (7) it is observed that the bias and mean square error of the estimator under discussion are the functions of 3 parameters namely $\frac{\triangle}{\sigma}$, n and α and out of which n is fixed in advance and determined by the experiment. Therefore the behaviour of bias and mean square error of the estimator is being studied for different values of α and $\frac{\triangle}{\sigma}$. The study has been made for $\alpha=.05$, .10 and .25 and $\frac{\triangle}{\sigma}\geqslant 0$ since Bias $\left(\frac{\triangle}{\sigma}\right)=-$Bias $\left(-\frac{\triangle}{\sigma}\right)$ and MSE $\left(\frac{\triangle}{\sigma}\right)=$MSE $\left(-\frac{\triangle}{\sigma}\right)$ in both the cases.

Tables 1 to 3 give the bias of the preliminary test estimator under case 1. The bias is zero for $\frac{\triangle}{\sigma}=0$ and at $n=24$ it tends to zero for $\alpha=.05$, .10 and .25. It also decreases with the increase in the level of significance for fixed values of $\frac{\triangle}{\sigma}$. For $\alpha=.05$ and .10, the bias is maximum at $\frac{\triangle}{\sigma}=1.50$ but for $\alpha=.25$ the maximum occurs at $\frac{\triangle}{\sigma}=1.25$.

Since the preliminary test estimator is in general biased and the estimator $\bar{X}_{n-1}(\bar{X}'_{n-1})$ is always unbiased, it seems more appropriate to talk of the relative efficiency which is defined as :

$$\text{R. E.}=\frac{\text{MSE (Unbiased Estimator)}}{\text{MSE (Preliminary Test Estimator)}}\,100\%$$

Tables 4 to 6 give the different values of relative efficiency. From the tables, we observe that for $\alpha=.05$, .10 and .25, the preliminary test estimator is more efficient over the unbiased estimator for $\frac{\triangle}{\sigma}<.75$ and becomes less efficient for $\frac{\triangle}{\sigma}>0.75$. The relative efficiency decreases with the increase in the level of significance from $\alpha=.05$ to .25.

TABLE 1

$\frac{\text{Bias}\,(\overset{*}{\bar{X}})}{\sigma}$ $(\alpha=0.05)$ Case-1

$\frac{\Delta}{\sigma}$	n				
	6	10	14	20	24
0.00	.000	.000	.000	.000	.000
0.25	.029	.017	.013	.008	.007
0.50	.057	.034	.024	.017	.014
0.75	.081	.048	.035	.024	.020
1.00	.101	.059	.036	.029	.024
1.25	.109	.067	.047	.032	.027
1.50	.119	.069	.048	.034	.028
1.75	.118	.069	.047	.032	.027
2.00	.111	.063	.042	.029	.025
3.00	.055	.025	.019	.012	.008

TABLE 2

$\frac{\text{Bias}\,(\bar{X}^*)}{\sigma}$ $(\alpha=0.10)$ Case-1

$\frac{\Delta}{\sigma}$	n				
	6	10	14	20	24
0.00	.000	.000	.000	.000	.000
0.25	.023	.014	.010	.007	.006
0.50	.044	.026	.018	.013	.011
0.75	.062	.036	.026	.018	.015
1.00	.074	043	.030	.021	.018
1.25	.081	.047	.033	.022	.019
1.50	.082	.046	.033	.023	.018
1.75	.078	.044	.031	.021	.017
2.00	.071	.039	.027	.018	.015
3.00	.021	.014	.009	.001	−.009

TABLE 3

$\frac{\text{Bias}\,(\overset{*}{\bar{X}})}{\sigma}$ (α=0 25) Case-1

$\frac{\Delta}{\sigma}$	n				
	6	*10*	*14*	*20*	*24*
0.00	.000	.000	.000	.000	.000
0.25	.011	.006	.005	.003	.003
0.50	.021	.012	.009	.006	.005
0.75	.028	.018	.012	.008	.007
1.00	.033	.019	.014	.009	.008
1.25	.035	.021	.014	.010	.008
1.50	.033	.019	.013	.009	.007
1.75	.030	.016	.012	.008	.006
2.00	.025	.013	.009	.006	.005
3.00	.007	.003	.002	.001	−.002

TABLE 4

Relative Efficiency of $\overset{*}{\bar{X}}$ to the unbiased Estimator, α=0.05.

$\frac{\Delta}{\sigma}$	n				
	6	*10*	*14*	*20*	*24*
0.00	114	108	105	104	103
0.25	112	107	105	103	103
0.50	108	105	103	102	102
0.75	104	102	101	101	101
1.00	098	088	098	099	099
1.25	092	095	096	097	097
1.50	088	092	093	095	096
1.75	084	089	092	094	095
2.00	082	088	091	093	094
3.00	083	090	093	095	096

TABLE 5

Relative Efficiency of X* to the unbiased Estimator $\alpha=0.10$.

$\frac{\Delta}{\sigma}$	*n*				
	6	10	14	20	24
0.00	110	106	104	103	102
0.25	109	105	103	102	101
0.50	102	103	102	101	101
0.75	102	101	100	100	100
1.00	098	098	099	098	099
1.25	094	095	096	097	097
1.50	090	093	095	096	097
1.75	088	092	094	0 5	096
2,00	086	083	094	095	096
3.00	090	094	096	097	100

TABLE 6

Relative Efficiency of X* to the unbiased Estimator, $\alpha=0.25$

$\frac{\Delta}{\sigma}$	*n*				
	6	10	14	20	24
0.00	105	103	102	101	101
0.25	104	102	102	101	101
0.50	102	101	100	100	100
0.75	101	100	100	100	100
1.00	098	099	099	099	099
1.25	096	097	098	098	099
1.50	095	096	097	098	098
1.75	094	096	097	098	098
2.00	094	096	097	098	098
3.00	097	098	099	099	100

In the end, an indiscriminate use of the unbiased estimator (assuming that the extreme observation is outlier) is not good because it is often less efficient than the preliminary test estimator. Similarly an indiscriminate use of the estimator $\bar{X}$ (assuming that no observation is outlier) is also not good because it gives large bias and is less efficient. From the above study we conclude that if we have a prior information that $\frac{\Delta}{\sigma}$ is small (say close to 0.75) the use of $\alpha \leqslant .05$ is recommended for the preliminary test and the estimator will be more efficient over the unbiased estimator. If we have a prior information about $\frac{\Delta}{\sigma} > 1.0$ then the use of unbiased estimator which is more efficient over the preliminary test estimator is recommended.

ILLUSTRATION

We illustrate the above procedure by using the following data due to K. Pearson (1931) which pertain to the capacities (in cubic centimeters) of 17 male Marior skulls :

1230, 1318, 1380, 1420, 1630, 1378, 1348, 1380, 1470, 1445, 1360, 1410, 1540, 1260, 1364, 1410 and 1548.

To test whether the highest observation 1630 is an outlier, on assuming normality and applying the modified form of McKay's test with $\sigma = 97.83$, we conclude that the largest observation 1630 is an outlier at 5% level of significance and therefore cannot be retained for the estimation of the capacities of Male Marior skulls.

Since in this example the estimate of $\frac{\Delta}{\sigma}$ is 2.37 for which the bias of the preliminary test estimator is small but this estimator is also less efficient and therefore cannot be preferred over the unbiased estimator and hence the observation 1630 is anomolous.

ACKNOWLEDGEMENT

The authors express their sincere thanks to the Referees for valuable suggestions.

REFERENCES

Bancraft, T.A. (1964) : Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. *Biometrika* 20, 427-442.

McKay, A.T. (1935) : The distribution of the difference between the extreme observation and sample mean in samples of size n from a normal Universe. *Biometrika* 27, 466-471.

Pearson, K. (1931) : *Tables for Statisticians and Biometricians.* Part 2, Cambridge University Press.

Anscombe, F.J. (1960) : Rejection of Outliers. *Technometrics.* 21, 123-47.