T R E A T M E N T

O F

O U T L Y I N G

O B S E R V A T I O N S

E. J. C o c k a y n e

1 9 6 2

CONTENTS

1. Introduction to the Problem Page 1

2. Grubbs' Significance Tests 4

2.1 Distribution of the Difference Between the Extreme and Mean 6

2.2 Distribution of S_n^2/S^2 and S_1^2/S^2 8

2.3 Example and Comments 13

3. Another Example of the Significance Approach 14

4. Criteria Based on Rejection Rates 17

5. Anscombe's Method 19

5.1 Simple Sample 22

5.11 Rule 1: No Spurious Observation: Calculation of Premium 25

5.12 Rule 1: One Spurious Observation: Calculation of Protection 34

5.13 Rule 2: n = 3: No Spurious Observation 39

5.2 Complex Patterns of Data 41

5.3 Results 49

APPENDIX: Integration and Computation 50

1. INTRODUCTION TO THE PROBLEM

The subject of statistics is concerned with making decisions or inferences about a given population from a representative sample of the population. Whenever data has been collected for such a purpose, it may be necessary or desirable to subject it to a critical examination to decide whether or not the sample is representative of the population in question so that any conclusions will be valid. The results of such an inspection may lead one to suspect the consistency of the sample, to feel that certain observations have been subject to abnormal errors of some kind and thus that the data, as it stands, does not truly represent the population. When such doubts exist, the following choice of procedures comes to mind.

1. Suppress the doubt. Proceed with inference or estimation techniques. This is clearly undesirable as sample suspicions must lead one to question any results from the sample.

2. Repeat the experiment partially or completely. In practice this would probably be impossible due to economic considerations. Further, unless we have great experience in the particular field, we cannot be sure that further results would be any more consistent than the original. If only the 'doubtful' readings are to be repeated we still have the problem of defining exactly the word 'doubtful'.

3. The intuitive temptation is to reject observations which have unusually large residual magnitudes and to use the remainder as a sample of lower order for estimation, decision-making purposes, etc. As in (2) the problem is, "where do we draw the line between normal and 'unusually large' residual magnitudes?"

The problem in certain cases may be easily solved by the simple application of common sense and(or) experimental knowledge. We here quote Rider [5]: 'In the final analysis it would seem that the question of the rejection or retention of a discordant observation reduces to a question of common sense. Certainly the judgement of an experienced observer should always be allowed considerable influence in reaching a decision.'

Such a solution to the problem would tell us nothing about the effect which rejecting observations has on the subsequent inferences or estimates.

The accompanying table lists the residuals of fifteen observations of the vertical semi-diameter of Venus. This example has almost become classic in any discussion of rejection of observations.

-".30	+.48	+ .63	-.22	+.18
- .44	-.24	- .13	-.05	+.39
+1.01	+.06	-1.40	+.20	+.10

The residual magnitudes 1.01 and 1.40 appear large compared with the remainder and one is tempted (in the interests of obtaining a more accurate estimate of the semi-diameter) to reject one or both of the observations yielding these residuals.

When a visual inspection using common sense and experience fails to expose suspicious elements or yields further doubt, it may be appropriate to perform some sort of analytical inspection using probability arguments and possibly to formulate and apply an analytical rejection rule to one's observations before finally accepting the sample. Rider continues 'This judgement can undoubtedly be aided by the application of one or more tests based on the theory of probability but any test which requires an inordinate amount of calculation seems to be hardly worth while, and the testimony of any criterion which is based on a complicated hypothesis should be accepted with extreme caution.'

This thesis is concerned with such analytical inspections and rejection rules. More specifically, it is intended to compare two 'classical' approaches to the problem, namely significance testing and rejection rate considerations with a recent technique, that of linking one's rejection criterion to the effect which it has on the resulting estimate(s).

Terminology

An observation with an unusually large residual magnitude will be termed an OUTLIER.

An observation which has been affected by some abnormal error and is not to be grouped with the remaining data for decisions or inferences will be called SPURIOUS.

2. GRUBBS' SIGNIFICANCE TESTS

In 1950 Frank E. Grubbs published a paper in which he developed significance tests of certain hypotheses concerning spurious readings. He pointed out that generally an observer will suspect a certain number of readings, some of which will be deemed unusually large and the remainder unusually small. He assumes that the observer will wish to decide whether or not the suspicious elements should be rejected by means of some significance test. Such an assumption has been made often by statisticians seeking means of dealing with outliers. We describe Grubbs' work to illustrate the significance testing approach.

(i) For testing the largest observation in a sample of n from a normal population, Grubbs suggests the statistic:

$$\frac{S_n^2}{S^2} = \frac{\sum_{1}^{n-1}(x_i-\bar{x}_n)^2}{\sum_{1}^{n}(x_i-\bar{x})^2},$$

where the x_i are the order statistics from a sample of n, and

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i \,, \quad \bar{x}_n = \frac{1}{n-1}\sum_{i=1}^{n-1} x_i \,.$$

A similarly defined statistic S_1^2/S^2 can be used for testing the smallest observation.

(ii) For testing whether the two largest observations are too large, the statistic suggested is:

$$\frac{S_{n-1,n}^{\ 2}}{S^2} = \frac{\sum_{i=1}^{n-2}(x_i - \bar{x}_{n-1,n})^2}{\sum_{i=1}^{n}(x_i - x)^2}$$

where $\bar{x}_{n-1,n} = \frac{1}{n-2}\sum_{i=1}^{n-2} x_i$,

and a similarly defined statistic $S_{1,2}^2/S^2$ can be used to test the two smallest readings. Clearly we could generalise the statistics for higher numbers of suspicious elements. Grubbs admitted that the powers of the above tests had not been determined for various models but pointed out the intuitive appeal of his quantities. In his paper some sample distributions of these statistics are derived, some percentage points computed and some examples given of the application of his work to reaching decisions on rejection. We give here a brief summary of his research.

2.1 Distribution of the difference between the extreme and mean: Normal theory

The joint p.d.f of the order statistics is

$$(2.1)\quad dF(x_1, \cdots, x_n) = \frac{n!}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2\right] dx_1 \cdots dx_n$$

where $x_1 < x_2 \cdots < x_n$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n-1}(x_i-\bar{x})^2 + (x_n-\bar{x})^2$$

where $\bar{x}_n = \frac{1}{(n-1)} \sum_{1}^{n-1} x_i$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n-1}(x_i-\bar{x}_n+\bar{x}_n-\bar{x})^2 + (x_n-\bar{x})^2$$

$$= \sum_{1}^{n-1}(x_i-\bar{x}_n)^2 + 2\sum_{i=1}^{n-1}(\bar{x}_n-\bar{x})(x_i-\bar{x}_n) + (n-1)(\bar{x}_n-\bar{x})^2 + (x_n-\bar{x})^2.$$

The crossed term vanishes and

$$(n-1)(\bar{x}_n-\bar{x})^2 = \frac{(n-1)}{n^2(n-1)^2}\left\{n\sum_{1}^{n-1} x_i-(n-1)\sum_{1}^{n} x_i\right\}^2$$

$$= \frac{1}{n-1}\{x_n-\bar{x}\}^2 .$$

Therefore $\sum_1^n (x_i-\bar{x})^2 = (x_n-\bar{x})^2 \frac{n}{n-1} + \sum_1^{n-1} (x_i-\bar{x}_n)^2$.

Thus, repeating this process, we find that

$$\sum_1^n x_i^2-n\bar{x}^2 = \frac{n}{n-1}(x_n-\bar{x})^2 + \frac{n-1}{n-2}(x_{n-1}-\bar{x}_n)^2 + \frac{n-2}{n-3}(x_{n-2}-\bar{x}_{n,n-1})^2$$

$$+ \cdots + \frac{3}{2}(x_3 - \frac{x_1+x_2+x_3}{3})^2 + \frac{2}{1}(x_2 - \frac{x_1+x_2}{2})^2 .$$

We substitute this identity for $\sum_1^n x_i^2$ in the probability element (2.1) and consider the orthogonal transformation

$$\begin{aligned}
\sqrt{2\cdot1}\,\sigma\eta_2 &= -x_1 + x_2 \\
\sqrt{3\cdot2}\,\sigma\eta_3 &= -x_1 - x_2 + 2x_3 \\
&\vdots \\
\sqrt{n(n-1)}\,\sigma\eta_n &= -x_1 - x_2 - x_3 \cdots - x_{n-1} + (n-1)x_n \\
\sqrt{n}\,\sigma\eta_{n+1} &= x_1 + x_2 + x_3 \cdots + x_{n-1} + x_n
\end{aligned} \tag{2.2}$$

The transformation has $|J| = \sigma^n$ and it turns out that after integrating out η_{n+1} the density function of $\eta_2 \cdots \eta_n$ is

$$dF(\eta_2\eta_3\cdots\eta_n) = \frac{n^{\frac{1}{2}}}{(\sqrt{2\pi})^{n-1}} \exp\left\{ -\frac{1}{2}\sum_{i=2}^n \eta_i^2 \right\} d\eta_2 d\eta_3 \cdots d\eta_n; \tag{2.3}$$

subject to $0 \leq \eta_2 \leq \infty$ and $\sqrt{\frac{r}{r-2}}\,\eta_r \geq \eta_{r-1}$.

(See K.R. Nair [4])

We now make the transformation $\sqrt{\frac{r-1}{r}}\,\eta_r = \frac{x_r - \bar{x}'_r}{\sigma} = u_r$

($\bar{x}'_r$ here is mean of $x_1, x_2, \cdots, x_r$) and define

$$F_n(u) = \int_0^u dF(u_n).$$

After integrating out the other variables one obtains

$$F_n(u) = n\sqrt{\frac{n}{n-1}} \int_0^u \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\cdot\frac{nx^2}{n-1}} F_{n-1}\left(\frac{nx}{n-1}\right) dx \ ;$$

thus the functions can be successively generated from the first, which is

$$F_2(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-x^2} dx.$$

$F_2(u)$ is a well-known and tabulated function denoted by erf u. Grubbs computed extensive tables of $F_n(u)$ for $n = 2$ to 25 and for values of u at .05 intervals in $(0, 4.90)$ together with percentage points at 10, 5, 1, .5 percent levels.

Note: the distribution of $(x_n - \bar{x}_n)$ can be obtained very easily from the above distribution by the use of the formula:

$$(x_n - \bar{x}_n) = \frac{n}{n-1}(x_n - \bar{x}).$$

2.2. Distribution of S_n^2/S^2 and S_1^2/S^2

From equation (2.3) we have the joint distribution of the η's is given by:

$$dF(\eta_2, \eta_3, \cdots, \eta_n) = \frac{n!}{(\sqrt{2\pi})^{n-1}} \exp\left\{-\frac{1}{2}\sum_{i=2}^{n}\eta_i^2\right\} d\eta_2 d\eta_3 \cdots d\eta_n \ .$$

Grubbs used the polar transformation

$$\begin{aligned}
\eta_2 &= r\sin\theta_n\sin\theta_{n-1}\cdots\sin\theta_4\sin\theta_3 \\
\eta_3 &= r\sin\theta_n\sin\theta_{n-1}\cdots\sin\theta_4\cos\theta_3 \\
\eta_4 &= r\sin\theta_n\sin\theta_{n-1}\cdots\sin\theta_5\cos\theta_4 \\
&\vdots \\
\eta_{n-1} &= r\sin\theta_n\cos\theta_{n-1} \\
\eta_n &= r\cos\theta_n
\end{aligned} \qquad (2.4)$$

Then

$$\sum_{i=2}^{n}\eta_i^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 = r^2$$

and

$$\sum_{i=2}^{n-1}\eta_i^2 = \sum_{i=1}^{n-1}(x_i-\bar{x}_n)^2 = r^2\sin^2\theta_n \ .$$

(2.5) Therefore $S_n^2/S^2 = \sin^2\theta_n$.

The Jacobian of the transformation is

$$r^{n-2}\sin^{n-3}\theta_n\sin^{n-4}\theta_{n-1}\cdots\sin^3\theta_6\sin^2\theta_5\sin\theta_4 .$$

Therefore integrating out r in $(0, \infty)$

$$dF(\theta_n, \theta_{n-1}, \cdots, \theta_4, \theta_3) =$$

$$\frac{n!}{(2\pi)^{(n-1)/2}}\, 2^{(n-3)/2}\,\Gamma\left(\frac{n-1}{2}\right)\sin^{n-3}\theta_n\cdots\sin^2\theta_5\sin\theta_4\, d\theta_n,\cdots,d\theta_3 .$$

The restrictions on η_i's were

$$\eta_2 \geq 0 \qquad \sqrt{\frac{r}{r-2}}\,\eta_r \geq \eta_{r-1} \qquad\qquad r \geq 3 .$$

Therefore the new sample space is given by

$\tan\theta_n \leq \sqrt{\frac{n}{n-2}}\,\sec\theta_{n-1}$ when $n \geq 4$ since

$$\tan\theta_n \cos\theta_{n-1} = \frac{\eta_{n-1}}{\eta_n}$$

and $0 \leq \theta_3 \leq \pi/3$.

Now write $k_n = \dfrac{n!}{(2\pi)^{(n-1)/2}}\; 2^{(n-3)/2}\,\Gamma(\frac{n-1}{2})$.

Then $k_n \int_0^{\pi/3}\int_0^{\ell_3}\cdots\int_0^{\ell_{n-2}}\int_0^{\ell_{n-1}} \sin^{n-3}\theta_n \cdots \sin^2\theta_5 \sin\theta_4 \, d\theta_n \cdots d\theta_3 = 1$ (2.6)

where $\ell_r = \tan^{-1}\sqrt{\frac{r+1}{r-1}}\,\sec\theta_r$.

Grubbs then considered special cases $n = 3, 4, \cdots$ etc. in turn. He defined

$$m_r = \tan^{-1}\sqrt{\frac{r}{r-2}}$$

$$M_r = \tan^{-1}\sqrt{(r-2)r} \qquad\qquad L_r = \sec^{-1}\sqrt{\frac{r-2}{r}}\,\tan\theta_r$$

He reversed the order of integration in (2.6), remembering that the varying ℓ_r are monotonic.

For $n = 3$

$$k_3 \int_0^{\pi/3} d\theta_3 = 1 .$$

Thus $P(\theta_3 \leq \theta) = k_3 \int_0^{\theta} d\theta_3 \qquad 0 \leq \theta \leq M_3 = \tan^{-1}\sqrt{3}$

For $n = 4$

$$P(\theta_4 \leq \theta) = \frac{k_4}{k_3} \int_0^{\theta} \sin\theta_4 d\theta_4 \qquad \text{when } 0 < \theta < m_4$$

$$P(\theta_4 \leq \theta) = \frac{k_4}{k_3} \int_0^{m_4} \sin\theta_4 d\theta_4 + k_4 \int_{m_4}^{\theta} \int_{L_4}^{\pi/3} \sin\theta_4 d\theta_3 d\theta_4$$

$$\text{when } m_4 < \theta < M_4.$$

For a sample of n

$$P(\theta_n \leq \theta) = \frac{n}{\sqrt{\pi}} \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})} \int_0^{\theta} \sin^{n-3}\theta_n d\theta_n \quad \text{when } 0 < \theta < m_n.$$

This may be shown to be equal to

$$P(\theta_n < \theta) = \frac{n}{2} I_{\sin^2\theta}\left(\frac{n-2}{2}, \frac{1}{2}\right) \qquad 0 < \theta < m_n$$

where $I_p(m, n)$ is the incomplete Beta function given by:

$$\frac{\int_0^p x^{m-1}(1-x)^{n-1} dx}{\int_0^1 x^{m-1}(1-x)^{n-1} dx}$$

and

$$P(\theta_n < \theta) = \frac{n}{2} I_{n/(2(n-1))}\left(\frac{n-2}{2}, \frac{1}{2}\right)$$

$$+ k_n \int_{m_n}^{\theta} \int_{L_n}^{M_{n-1}} \int_{L_{n-1}}^{M_{n-2}} \cdots \int_{L_4}^{\pi/3} \sin^{n-3}\theta_n \cdots \sin\theta_4 d\theta_3 \cdots d\theta_n$$

$$\text{when } m_n < \theta < M_n.$$

The required distribution follows using equation (2.5).

The theory developing the distribution of S_1^2/S^2 is similar.

Grubbs computed tables of percentage points of these statistics. He then continued to determine the distributions of his proposed statistics for the simultaneous testing of the two largest or smallest observations.

From the above equations (2.4) it is clear that

$$\sum_{i=2}^{n} \eta_i^2 = r^2$$

$$\sum_{i=2}^{n-2} \eta_i^2 = r^2 \sin^2\theta_n \sin^2\theta_{n-1}.$$

Therefore

$$\frac{S^2_{n,n-1}}{S^2} = \sin^2\theta_n \sin^2\theta_{n-1}.$$

Grubbs then made the following transformation in equation (2.6)

$$\sin\Delta_n = \sin\theta_n \sin\theta_{n-1}$$

$$\Delta_i = \theta_i \qquad 3 \leq i \leq n-1.$$

After a quite complex evaluation of Jacobian and limits we obtain an expression of the form

$$k_n \int \cdots \int f(\Delta_3, \cdots, \Delta_n) d\Delta_n \cdots d\Delta_3 = 1$$

and by reversing integration orders, etc., distributions of $\Delta_3\Delta_4\cdots\Delta_n$ can be found in a way similar to that above. Hence the distributions of $S^2_{n,n-1}/S^2$ and S^2_{12}/S^2. Some percentage points were tabulated by Grubbs.

2.3 Example and Comments

Grubbs applied his theory to the data given in the introduction paragraph. The statistics S_1^2/S^2 is applied to test the least observation, i.e. that having residual -1.40, and it was found that the observation would be rejected at the 5 percent significance level. Considering the remainder as a sample of 14 observations and testing the largest, Grubbs found that this observation would be retained at 5 percent level. He remarks that it would have been of interest to test the largest and smallest observations simultaneously by means of the statistic $S_{1,n}^2/S^2$ (with obvious notation). Although he hinted at a method for determination of this sample distribution, no percentage points were computed.

One would be interested to compare tests of the two largest observations in a sample of n, first by use of the statistic $S_{n-1,n}^2/S^2$, secondly by repeated application of the statistic S_n^2/S^2 to samples of n and $n-1$ respectively.

The use of such statistics $S_{n,n-1}^2/S^2$, S_{12}^2/S^2 (and indeed Grubbs hints at generalisations to simultaneous treatment of r suspicious elements by such statistics as $S_{1,2,\ldots,r}^2/S^2$) seems of doubtful validity since one sufficiently extreme observation can cause several unusually high residuals.

3. ANOTHER EXAMPLE OF THE SIGNIFICANCE APPROACH

In a paper soon to be published, Professor S. S. Wilks, Princeton University, discusses significance tests for multidimensional outliers.

Suppose we have a k-dimensional random variable $(X_1, \cdots, X_k)$ and a sample of order n from it:

$$\begin{matrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \cdot & \cdot & \cdots & \cdot \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{matrix}$$

Suppose the means of the columns are $\bar{x}_{1.}, \bar{x}_{2.}, \cdots, \bar{x}_{k.}$. Then the scatter matrix is (A) where

$$A_{ij} = \sum_{\ell=1}^{n} (x_{i\ell} - \bar{x}_{i.})(x_{j\ell} - \bar{x}_{j.}) \qquad (i, j = 1, \cdots, k)$$

(A) is a $k \times k$ matrix.

Suppose we wish to test whether the n^{th} observation is outlying or not; then we delete this observation and compute the new scatter matrix (A') of the remaining (n-1) observations.

The statistic r is then defined by:

$$r = \frac{\det(A')}{\det(A)}$$

For a homogeneous sample from a normal parent population, Wilks found that this statistic has a Beta-distribution and would reject the n^{th} reading if $r < \lambda$ where λ is some constant determined by the size of the test.

Suppose we now take a special case of Wilks' method, namely $k = 1$, i.e. we are dealing with a one-dimensional random variable. Then we test the significance of an observation x_j by the statistic

$$\frac{S_j^2}{S^2} = \frac{\sum_{i \neq j}^{n} (x_i - \bar{x}_j)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$(3.1) \qquad \frac{S_j^2}{S^2} = \frac{\sum_{i \neq j}^{n} (x_i - \bar{x}_j)^2}{\sum_{i \neq j}^{n} (x_i - \bar{x}_j)^2 + \frac{n}{n-1}(x_j - \bar{x})^2}$$

(See Identity in Paragraph 2, Page 7)

Note there is great similarity between the above and Grubbs' statistic for testing extreme observations. We are here, however, not considering order statistics but simply testing the significance of an observation assumed to be from a certain parent population (here assumed normal).

From (3.1) we have

$$(3.2) \qquad s_j^2/s^2 = \frac{1}{1 + \dfrac{n(x_j-\bar{x})^2}{(n-1)\sum_{i\neq j}^{n}(x_i-\bar{x}_j)^2}}$$

Now $\sum_{i\neq j}^{n}(x_i-\bar{x}_j)^2/\sigma^2$ is the sample variance of (n-1) observations and therefore has a χ^2_{n-2} probability distribution under the hypothesis of no spurious observation.

$\frac{\sqrt{n}(x_j-\bar{x})}{\sigma\sqrt{n-1}}$ is an N(0, 1) random variable under the null hypothesis. Therefore $\frac{n(x_j-\bar{x})^2}{\sigma^2(n-1)}$ has a χ^2_1 distribution. Also the above quadratic forms may be expressed, one as the square of a linear combination of the sample $(x_1, \cdots, x_n)$ and the other as a sum of squares of linear combinations of the x_i (see Paragraph 2.1). Further, all these linear combinations are independent -- proved for the case of sample size three in paragraph 5, page 28. Thus the above forms are independent and

$$\frac{n(x_j-\bar{x})^2}{\sigma^2(n-1)} \Bigg/ \frac{\sum_{i\neq j}^{n}(x_i-\bar{x}_j)^2}{\sigma^2(n-2)}$$

has an $F_{1,n-2}$ distribution.

From (3.2) our rejection criterion reads:

$$\text{Reject if } \quad r = \frac{1}{1 + \lambda F_{1,n-2}} < k_1$$

{k_1 constant depending on size of test, λ constant depending on n}

Since r is a decreasing function of F the rule may be formulated: Reject if the observed value of

$$F_{1,n-2} > k_2$$

{k_2 constant determined by size}

The above treatment appears to sidestep the distribution difficulties encountered by Grubbs.

4. CRITERIA BASED ON REJECTION RATES

Many statisticians in the past who have been concerned with outlying observations and rejection criteria, have taken rejection rates or the proportion of observations rejected in the long run as their major consideration. We illustrate such methods with a rule devised by E. J. Stone as early as 1868.

Stone's hypothesis was that for a given observer and group of observations there will exist some number m which will be such that on the average one observation in every m will be subject to some gross error and ought to be discarded. The number m was originally called 'Modulus of Carelessness.' Stone formulated his rejection rule as follows:

Calculate k such that the probability that a reading shall deviate from the mean by $k\sigma$ is $1/m$.

Then reject any observations such that the absolute value of its deviation is greater than $k\sigma$. In this way, argued Stone, we shall be rejecting at precisely 'the correct rate' to eliminate errors. The rule is fairly simple to work with using normal theory and Stone tabulated some corresponding values of k and m.

Rider states that the law can be interpreted as giving the rejection criterion when we are to reject if the probability of the corresponding error is less than $1/m$. This statement eliminates any question of moduli of carelessness.

This method is possibly the simplest using rejection rates. There have been several refinements and other attempts based on rejection rates. One slight change in Stone's criterion due to Edgeworth may be described as follows:

Suppose $\Pr\{|\text{deviation}| > k\sigma\} = \psi(k)$.

Then Stone would urge:

Reject all observations such that $|d| > k\sigma$

where $\psi(k) = 1/m$.

Edgeworth suggested the alternative:

Reject all observations such that $|d| > k\sigma$ where

$$\{1 - \psi(k)\}^n = 1 - 1/m.$$

The rejection rate technique seems to have a certain amount of intuitive appeal in certain cases where m is reasonably

accurately known. One can foresee trouble if the estimate of m is poor. As mentioned in Paragraphs 2 and 5, another source of trouble is that one sufficiently extreme observation can distort several other residuals to the extent where one would suspect the parent observations.

5 ANSCOMBE'S METHOD

The last major point of view to be exposed in this thesis is that of Professor F. J. Anscombe, whose paper is dealt with in detail below. The above treatments of rejection rules are, in his opinion, somewhat misguided, the principal notion in his work being illustrated as follows:

An experimental scientist who is setting out to estimate certain parameters, applies a routine rejection rule primarily to safeguard the accuracy of his experiment. Thus, rejection rules should not be thought of as significance tests as is the case in Grubbs' work, Paragraph 2, and Wilks' research, Paragraph 3, and rejection rates are of no more than incidental interest. If the experiment had been conducted simply to investigate the quantity of spurious observations to be expected in any given sample or just how wild the spurious readings are, then clearly significance tests and rejection rates are relevant. From this reasoning, the principal factor influencing the choice of a rejection rule in any particular experiment where parameter estimate(s) are desired, should be the effect

which the chosen rule has on the accuracy of the resulting estimate(s).

Anscombe suggests the analogy that a rejection rule can be likened to a householder's fire-insurance policy. The questions to be answered before selecting such a policy are:

1. What is the premium payable?
2. What is the protection given?
3. What danger is there of a fire?

The last item corresponds in the analogy to significance testing for spurious observations and computation of rejection rates as mentioned above. Anscombe points out that "a householder satisfied that fires DO occur does not bother much about (3) providing that the premium is moderate and the protection good."

To complete the analogy it remains to define 'premium' and 'protection'. Anscombe uses the estimate variance as his measure of these, and states that any other definition of expected loss could be employed. The premium payable is defined as "The percentage increase in the variance of estimation errors due to using the rejection rule when in fact all observations come from a homogeneous source." The protection given is "The reduction in variance (or mean squared error) due to the rule, when spurious readings are present."

Notation

In the following theory the notation used is that shown below:

Observations: $y_1 y_2 \cdots y_n$

Residuals: $z_1 z_2 \cdots z_n$

Sample Size: n

Degrees of Freedom of Residuals: ν

Estimates: These are denoted by the symbol '$\wedge$' e.g. an estimate of parameter θ is denoted '$\hat{\theta}$'

P.d.f. of Normal Random Variable: $\phi(y) = e^{-u^2/2}/\sqrt{2\pi}$

Cumulative Distribution Function of Standard Normal Random Variable: $Pr(Y \leq y) = \Phi(y)$

Sample Mean: $\bar{y} = \sum_1^n y_i / n$

We now present a more detailed survey of parts of Professor Anscombe's work, the principle of which has been discussed above. The introduction of the ideas of premium and protection facilitates investigations and comparisons of any suggested rejection rules. The paper makes the following assumptions:

(i) The factor(s) causing a spurious observation will not affect any other observation, i.e. the observations are independent. Further, the degree of the spuriousness does not depend on the value that should have been observed.

(ii) Computation costs can be ignored. If we were not prepared to concede this point, the premium would have

to include extra computational costs resulting from use of the rejection rule.

(iii) The rejection rule is impartial. We have no prior information concerning the parameters to be estimated. This assumption will be illustrated when we consider complex patterns of data below.

The theory is applied to normal distributions throughout and the rejection criterion is based on the magnitude of residuals in each case. Firstly we consider possible spurious observations in a simple sample from which one parameter (population mean) is to be estimated; the variance is assumed known. The case $n = 3$ is developed in detail. The techniques are then applied to complex patterns of data, i.e. we investigate the effect of rejection rules on a sample drawn for the estimation of several parameters in a complex system (variances are again assumed known) together with some problems which influence the procedure. Anscombe, in his paper, makes some mention of the method when σ^2 is unknown. We do not consider this case here.

5.1 Simple Sample

We are given a sample of size n (≥ 3) which is thought to be from $N(\mu, \sigma^2)$ where σ^2 is known and it is desired to estimate μ. Possibly one or more of the y_i are spurious coming from a different source and should be rejected. In order to do this we formulate the following rejection rules:

Rule 0

For given C, reject all y_i such that $|z_i| > C\sigma$. Estimate μ from the mean of the retained observations.

This rule appears reasonable at first sight but when one considers that one sufficiently wild observation can cause a number of residuals to exceed C in absolute value, it is clear that such a rule could (with high probability) reject several 'innocent' observations. Thus we formulate Rule 1, which can reject only one observation, that having the greatest $|z_i|$.

Suppose M is the suffix of the observation having the greatest $|z_i|$

$$\text{i.e.} \quad |z_M| = \max_i |z_i| \qquad i = 1, \ldots, n.$$

We suppose that observations are recorded sufficiently accurately that no two $|z_i|$'s are equal.

Rule 1

For given C, reject y_M if $|z_M| > C\sigma$, otherwise no rejections. Estimate μ from the mean of the remaining observations.

Thus we have:

$$\hat{\mu} = \bar{y} \qquad \text{if} \quad |z_M| < C\sigma$$

$$= \bar{y} - \frac{z_M}{n-1} \qquad \text{if} \quad |z_M| > C\sigma.$$

In this simple case residual degrees of freedom is equal to $(n-1)$. Hence $\nu = n - 1$ and

$$\hat{\mu} = \bar{y} - \frac{z_M}{v} \qquad \text{if} \quad |z_M| > C\sigma.$$

We can extend Rule 1 to reject more than one observation and at the same time avoid the difficulties of Rule 0 by the following formulation:

Rule 2

Apply Rule 1. If an observation is rejected, consider the remaining observations as a sample of size (n-1) and apply Rule 1 again and so on. Estimate μ by the mean of the retained observations.

Note that if $n = 2$, $|z_1| = |z_2|$. Hence if an attempt is made to apply Rule 1 to a sample of only two observations either there are no rejections or both observations are rejected. Thus Rule 2 applied to n observations ($n \geq 3$) leads to the rejection of 0 or 1 or 2 or ... or (n-2) or all n observations.

It is of course possible to change the value of C in successive applications of Rule 1. This is simply a matter of choice depending on what protection one requires and how much premium one is willing to pay.

For example, in a special case to be considered below we take

$$C_r = t_\alpha \sqrt{\frac{r-1}{r}}$$

where t_α is independent of r.

We now develop the necessary theory to determine the premium and protection for the application of Rule 1.

5.11 Rule 1. No Spurious Observations -- Calculation of Premium

We calculate the proportional increase in variance of $\hat{\mu}$ due to the unnecessary application of the rejection rule. The joint distribution of the residuals is independent of the distribution of $\bar{y}$. Therefore $\bar{y}$ and z_M are independent random variables.

$$z_i = y_i - \bar{y} = -\frac{1}{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} y_j + \left(\frac{n-1}{n}\right) y_i .$$

The y_i $(i = 1, \ldots, n)$ are independent, hence, using the theorem on sums of independent normal random variables, z_i is a normal random variable.

$$\text{The mean of } z_i = -\frac{1}{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} E(y_j) + \frac{n-1}{n} E(y_i)$$

$$= -\frac{(n-1)\mu}{n} + \frac{(n-1)\mu}{n} = 0,$$

$$\text{and } \operatorname{var}(z_i) = E(z_i^2)$$

$$= \left(\frac{1}{n}\right)^2 \sum_{\substack{j=1 \\ j \neq i}}^{n} E(y_j^2) + \left(\frac{n-1}{n}\right)^2 E(y_i^2)$$

$$= \frac{\sigma^2}{n^2} \{(n-1)^2 + (n-1)\} = \frac{v\sigma^2}{n} .$$

Therefore $\left(\frac{n}{v}\right)^{1/2} \frac{z_i}{\sigma}$ is an $N(0, 1)$ random variable.

We now define a random variable T as the following function of z_M.

$$
\begin{aligned}
&T = 0 && \text{if} \quad |z_M| < C\sigma \\
&T = -\left(\frac{n}{\nu}\right)^{1/2} \frac{z_M}{\sigma} && \text{if} \quad |z_M| > C\sigma.
\end{aligned} \tag{5.1}
$$

Since there are no spurious observations z_M has distribution as given above.

$$\text{Therefore} \quad E(T) = 0. \tag{5.2}$$

Then, since under Rule 1 we reject 0 or 1 observations, we have:

$$\hat{\mu} = \bar{y} + \frac{\sigma T}{(n\nu)^{1/2}}$$

and $\bar{y}$ and T are independent.

Therefore $E(\hat{\mu}) = E(\bar{y}) + \frac{\sigma E(T)}{(n\nu)^{1/2}} = E(\bar{y}) = \mu$

so that $\hat{\mu}$ is an unbiased estimate of μ.

$$
\begin{aligned}
\operatorname{var}(\hat{\mu}) &= \operatorname{var}(\bar{y}) + \operatorname{var}\left\{\frac{\sigma T}{(n\nu)^{1/2}}\right\} \\
&= \frac{\sigma^2}{n}\left[1 + \frac{1}{\nu}\{E(T^2) - (E(T))^2\}\right] \\
&= \frac{\sigma^2}{n}\left\{1 + \frac{E(T^2)}{\nu}\right\} \qquad \text{using (5.2).}
\end{aligned}
$$

The premium payable, p, is the proportional increase in $\operatorname{var}(\hat{\mu})$.

Therefore $$p = \frac{\sigma^2/n\{1 + \frac{E(T^2)}{\nu}\} - \sigma^2/n}{\sigma^2/n}$$

(5.3) Therefore $$p = \frac{E(T^2)}{\nu} .$$

Suppose $\Pr\{|z_M| < z\} = F(z)$
and the probability density function of T is $g(T)$ in $-\infty < T < \infty$. Then

$$\frac{E(T^2)}{\nu} = \frac{1}{\nu} \int_{-\infty}^{\infty} T^2 g(T)dT = \frac{1}{\nu} \int_C^{\infty} \frac{nz^2}{\nu\sigma^2} d\{F(z)\}$$

(5.4)
$$= \frac{n}{\nu^2\sigma^2} \int_C^{\infty} z^2 \frac{dF}{dz} dz .$$

Now a rejection occurs if $|z_M| > C$.
Therefore $\Pr\{\text{Rejection}\} = \Pr\{|z_M| > C\}$
$= 1 - \Pr\{|z_M| < C\}$
$= 1 - F(C)$.

Thus we have that, in the long run, the proportion of observations rejected, i.e. the rejection rate

(5.5) $$= \frac{1}{n} \{1 - F(C)\} .$$

<u>Special Case</u> $n = 3$

We consider $\sigma^2 = 1$. This assumption makes no difference to proportional variance increases. We have observations y_1, y_2, y_3 from $N(\mu, 1)$ and corresponding residuals z_1, z_2, z_3 satisfying $\sum_{i=1}^{3} z_i = 0$.

Define $x_1 = (z_2 - z_1)\ 1/\sqrt{2}$

$x_2 = (2z_3 - z_2 - z_1)\ 1/\sqrt{6}$.

This implies that x_1 and x_2 are N(0, 1) variables and are independent. To show this we notice that

$$x_1 = \frac{1}{\sqrt{2}} \{y_2 - \bar{y} - (y_1 - \bar{y})\} = \frac{1}{\sqrt{2}} (y_2 - y_1).$$

Similarly $x_2 = \frac{1}{\sqrt{6}} \{2y_3 - y_2 - y_1\}$.

Therefore, using theory of linear combinations of independent normal random variables, we have that x_1 and x_2 are N(0, 1) variables. Since the means of x_1 and x_2 are zero,

$$\begin{aligned} \text{cov}(x_1, x_2) &= E(x_1 x_2) \\ &= E\{(1/\sqrt{12})(y_2 - y_1)(2y_3 - y_2 - y_1)\} . \end{aligned}$$

Since the y_i's are independent and identically normally distributed,

$$\text{Cov}(y_i, y_j) = 0 \qquad (i \neq j)$$

$$\text{Cov}(x_1, x_2) = \frac{1}{\sqrt{12}}\{\text{var}(y_1^2) - \text{var}(y_2^2)\} = 0.$$

Since for normal variables $\text{cov}(x_i, x_j) = 0$ implies x_i and x_j are independent, we have that x_1 and x_2 are independent as hypothesised.

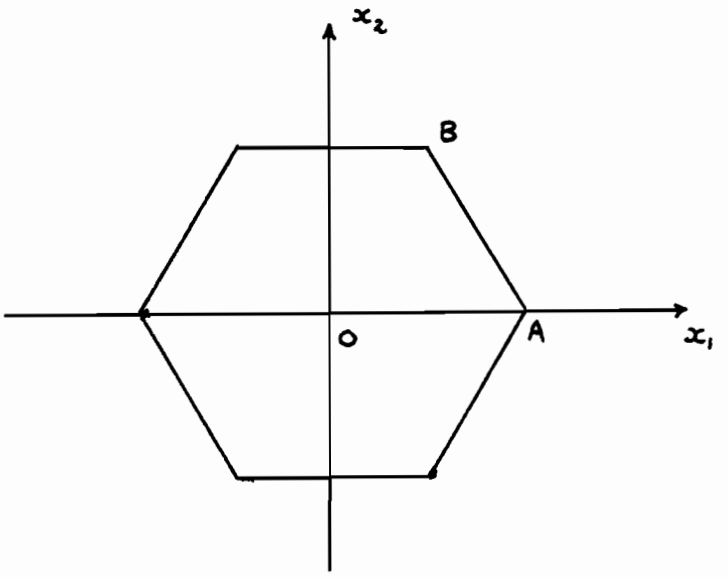


Figure 5.1

Since $\sum_{i=1}^{3} z_i = 0$ we derive

$$z_1 = -\frac{x_1}{\sqrt{2}} - \frac{x_2}{\sqrt{6}} \; ; \; z_2 = \frac{x_1}{\sqrt{2}} - \frac{x_2}{\sqrt{6}} \; ; \; z_3 = \frac{2x_2}{\sqrt{6}} \; .$$

Therefore $|z_1| < z$ implies $-z < \frac{x_1}{\sqrt{2}} + \frac{x_2}{\sqrt{6}} < z$

$|z_2| < z$ implies $-z < \frac{x_1}{\sqrt{2}} - \frac{x_2}{\sqrt{6}} < z$

$|z_3| < z$ implies $-z < \frac{2x_2}{\sqrt{6}} < z \; .$

All these conditions are satisfied if no rejection occurs when $C = z$ and geometrically the conditions imply that (x_1, x_2) lies inside a regular hexagon in the x_1, x_2 plane, centre the origin (See Fig. 5.1). The vertices and midpoints of sides are distant $z\sqrt{2}$ and $z\sqrt{3}/2$ from the centre respectively.

The probability that an observed (x_1, x_2) lies in an area dx_1dx_2 is the probability element:

$$\left(\frac{1}{\sqrt{2\pi}}\right)^2 e^{-\frac{1}{2}x_1^2} e^{-\frac{1}{2}x_2^2} dx_1 dx_2.$$

Transforming to polar co-ordinates:

$$x_1 = r\cos\theta$$

$$x_2 = r\sin\theta$$

the new probability element becomes

$$\frac{1}{2\pi} e^{-\frac{1}{2}r^2} r dr d\theta = \frac{d\theta}{2\pi} e^{-\frac{1}{2}r^2} d(\tfrac{1}{2}r^2).$$

The probability that (x_1, x_2) lies outside the hexagon is equal to the probability that $|z_m| > z$, which is $1 - F(z)$ (see Page 27). To find this probability we integrate the probability element over the exterior, E, of the hexagon. From the symmetry of the figure it is clear that we can split this integral into six equal parts and obtain:

$$1 - F(z) = 6 \int_0^{\pi/3} \frac{d\theta}{2\pi} \int_R^\infty e^{-\frac{1}{2}r^2} d(\tfrac{1}{2}r^2)$$

where R is the length of the segment OP (see figure 5.2), O the centre of the hexagon and P an interior point of the side AB.

Therefore $1 - F(z) = \frac{3}{\pi} \int_0^{\pi/3} e^{-\frac{1}{2}R^2} d\theta$

Applying the sine rule to ΔAOP in figure 5.2 we see that

$$\frac{R}{\sin \pi/3} = \frac{z\sqrt{2}}{\sin(2\pi/3 - \theta)}$$

or $$R = z\sqrt{\frac{3}{2}} \sec(\frac{\pi}{6} - \theta)$$

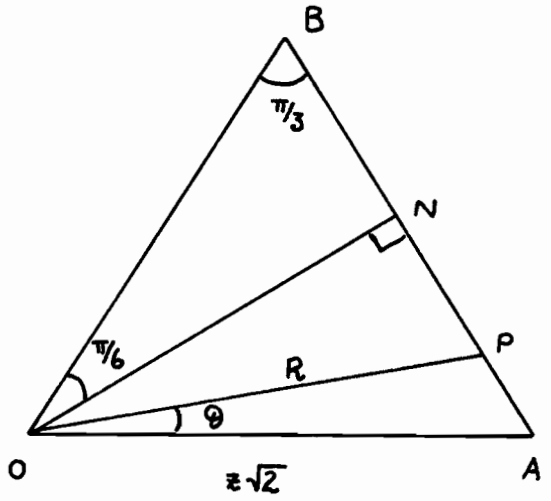


Figure 5.2

Therefore $1 - F(z) = \frac{3}{\pi} \int_0^{\pi/3} e^{-\frac{3}{4}z^2 \sec^2(\frac{\pi}{6} - \theta)} d\theta.$

Now make the transformation $\phi = \frac{\pi}{6} - \theta$ (ϕ measured from ON).

Then $1 - F(z) = \frac{3}{\pi} \int_{-\pi/6}^{\pi/6} e^{-\frac{3}{4}z^2 \sec^2 \phi} d\phi.$

If $\tan\phi = t$, then $d\phi = dt/(1 + t^2)$ and

$$(5.6) \qquad 1 - F(z) = \frac{3}{\pi} \int_{-1/\sqrt{3}}^{1/\sqrt{3}} \frac{e^{-\frac{3}{4}z^2(1+t^2)}}{(1 + t^2)} dt \ .$$

The rejection rate using (5.5) is

$$\frac{1}{3}\{1 - F(C)\}$$

$$(5.7) \quad = \quad \frac{1}{\pi} \int_{-1/\sqrt{3}}^{1/\sqrt{3}} \frac{e^{-\frac{3}{4}C^2(1+t^2)}}{(1+t^2)} \, dt \quad .$$

From (5.4) the premium payable

$$P = \frac{n}{\nu^2} \int_C^\infty z^2 \frac{dF(z)}{dz} \, dz$$

Differentiating (5.6) under the integral sign with respect to z and multiplying by z^2 we obtain:

$$z^2 \frac{dF(z)}{dz} = \frac{3}{4\pi} \int_{-1/\sqrt{3}}^{1/\sqrt{3}} z^2 e^{-\frac{3}{4}z^2(1+t^2)} \, 2.3.z \, dt.$$

Then $$\int_C^\infty z^2 \frac{dF(z)}{dz} \, dz = \int_C^\infty \frac{3dz}{4\pi} \int_{-1/\sqrt{3}}^{1/\sqrt{3}} z^2 e^{-\frac{3}{4}z^2(1+t^2)} \, 2.3.z \, dt.$$

Inverting the order of integration and making the substitution $\lambda = \frac{3}{4}(1+t^2)$, and the transformation $u = \lambda z^2$ we obtain

$$\int_C^\infty z^2 \frac{dF(z)}{dz} \, dz = \frac{3}{\pi} \int_{-1/\sqrt{3}}^{1/\sqrt{3}} \frac{dt}{(1+t^2)} \int_{\lambda C^2}^{-\infty} - \frac{ue^u}{\lambda} \, du$$

$$= \frac{3}{\pi} \int_{-1/\sqrt{3}}^{1/\sqrt{3}} e^{\lambda C^2} (C^2 - \frac{1}{\lambda}) \frac{dt}{(1+t^2)}$$

$$= \frac{3}{\pi} \int_{-1/\sqrt{3}}^{1/\sqrt{3}} e^{-\frac{3}{4}C^2(1+t^2)} \left\{ \frac{4}{3(1+t^2)} + C^2 \right\} \frac{dt}{(1+t^2)} .$$

Therefore, from (5.4)

$$p = \frac{3}{4} \int_{C}^{\infty} z^2 \frac{dF(z)}{dz} \, dz$$

Therefore, (5.8) $$p = \frac{3}{\pi} \int_{-1/\sqrt{3}}^{1/\sqrt{3}} e^{-\frac{3}{4}C^2(1+t^2)} \left\{ \frac{1}{(1+t^2)} + \frac{3C^2}{4} \right\} \frac{dt}{(1+t^2)} .$$

Using an IBM 650 Digital Computer, the values of C were computed which lead to premiums of 5, 4, 2, 1, and 1/2 percent respectively. The values obtained are tabulated in section 5.3. Very good agreement was found between values computed by the author and those published in Professor Anscombe's paper.

The latter also computed the premium using the empiric formula given below for which there appears to be no rigorous justification.

(5.9) $$p = \frac{n}{\nu}\{2t_\alpha \phi(t_\alpha) + \alpha\}$$

where $\alpha = 2\phi\{-(\frac{n}{\nu})^{1/2}C\}$, $t_\alpha = (\frac{n}{\nu})^{1/2}C$.

The values of C obtained above were inserted in this approximate formulation and the values of p computed and tabulated (section 5.3).

5.12 Rule 1. One Spurious Observation -- Calculation of Protection

In this case it is clear that, should the rejection rule fail to reject the spurious observation we shall have a biased estimate for μ. The protection criterion is the value of $E(\hat{\mu} - \mu)^2$, i.e. the protection is defined as the proportional decrease in $E(\hat{\mu} - \mu)^2$ due to the application of the rejection rule.

It is convenient to consider the spurious observation (say y_n) as a member of the population $N(\mu + a\sigma, \sigma^2)$ while the y_i, $i = 1, \ldots, (n-1)$ are independent observations from $N(\mu, \sigma^2)$.

Now $$\bar{y} = \frac{1}{n} \sum_{i=1}^{n-1} y_i + \frac{y_n}{n},$$

and y_i and y_n are independent -- assumption (i), Page 21 and of course $\bar{y}$ is a normal variable.

The mean of $\bar{y} = E(\bar{y})$

$$= \frac{1}{n} \sum_{i=1}^{n-1} E(y_i) + \frac{1}{n} E(y_n)$$

$$= \frac{(n-1)\mu}{n} + \frac{(\mu + a\sigma)}{n}$$

$$= \mu + \frac{a\sigma}{n},$$

and the variance of $\bar{y}$ is σ^2/n as in the case of no spurious observations dealt with in paragraph (5.11). $\bar{y}$ and the residuals are independent.

Define a new random variable S as follows:

$$(5.10)\qquad \begin{aligned} S &= a\sigma/n && \text{if } |z_M| < C\sigma \\ S &= a\sigma/n - z_M/n-1 && \text{if } |z_M| > C\sigma. \end{aligned}$$

It follows that S and $\bar{y}$ are independent. Then, under Rule 1,

$$\hat{\mu} = (\bar{y} - \frac{a\sigma}{n}) + S.$$

Therefore $E(\hat{\mu}-\mu)^2 = E(\bar{y} - \frac{a\sigma}{n} - \mu)^2 + E(S^2) + 2E\{S(y-\frac{a\sigma}{n}-\mu)\}$.

The last term is zero since $\bar{y}$ and S are independent and

$E(\bar{y} - \frac{a\sigma}{n} - \mu) = 0$. Therefore, since

$$E(\bar{y} - \frac{a\sigma}{n} - \mu)^2 = \frac{\sigma^2}{n}$$

$$(5.11)\qquad E(\hat{\mu}-\mu)^2 = \frac{\sigma^2}{n} + E(S^2).$$

If we had ignored the existence of the spurious observation and estimated μ with $\bar{y}$ the variance of the estimate would have been

$$(5.12)\qquad \begin{aligned} E(\hat{\mu}-\mu)^2 = E(\bar{y}-\mu)^2 &= E\{(\bar{y} - \frac{a\sigma}{n} - \mu) + \frac{a\sigma}{n}\}^2 \\ &= \frac{\sigma^2}{n}\{1 + \frac{a^2}{n}\}. \end{aligned}$$

From (5.11) and (5.12) we calculate the protection. The rejection rate is given by the same formula as in the case of no spurious observations, i.e. equation (5.5).

Special Case n = 3

We consider $\sigma^2 = 1$, noting that this assumption has no effect on the proportional variance decrease due to applying the rejection rule. As in the case of no spurious observation we define

$$x_1 = \frac{1}{\sqrt{2}}(z_2 - z_1)$$

$$x_2 = \frac{1}{\sqrt{6}}(2z_3 - z_2 - z_1).$$

In this case y_1, y_2 are from $N(\mu, 1)$ and y_3 is from $N(\mu + a, 1)$, all the observations being independent. Then, by a proof similar to the previous one, (page 28) we have x_1 is an $N(0, 1)$ variable but x_2 is distributed as $N(\sqrt{\frac{2}{3}}a, 1)$.

$$\mathrm{Cov}(x_1, x_2) = E\{(x_1)(x_2 - \sqrt{\tfrac{2}{3}}a)\}$$

$$= E[\{\tfrac{1}{\sqrt{2}}(y_2 - y_1)\}\{(2y_3 - y_2 - y_1)/\sqrt{6} - 2a/\sqrt{6}\}]$$

$$= \tfrac{1}{\sqrt{12}}E\{(y_2 - y_1)(2y_3 - y_2 - y_1 - 2a)\}$$

$$= \tfrac{1}{\sqrt{12}}E\{(y_2 - y_1)(2y_3 - y_2 - y_1)\} - \tfrac{2a}{\sqrt{12}}E(y_2 - y_1)$$

$$= \tfrac{1}{\sqrt{12}}\{2E(y_3)(y_2 - y_1) + E(y_1^2 - y_2^2)\}$$

$$= \tfrac{2}{\sqrt{12}}E(y_3)E(y_2 - y_1) = 0.$$

For normal variables $\text{cov}(x_1, x_2) = 0$ implies that $x_1 x_2$ are independent. The joint p.d.f of x_1, x_2 is therefore:

$$\frac{1}{2\pi} e^{-\frac{1}{2}\{x_1^2 + (x_2 - \sqrt{\frac{2}{3}}a)^2\}}. \tag{5.13}$$

Let $|z_M| = W$. Then (page 35) S is a function of W, say $h(W)$. Suppose we denote the $\Pr\{|z_M| < z\} = F(z)$ but note that in this case $F(z)$ depends on a.

$$E(S^2) = E[\{h(W)\}^2]$$

$$= \int_0^\infty \{h(w)\}^2 dF(w)$$

$$= \frac{a^2}{n^2} F(C) + \int_C^\infty \{\frac{a}{n} - \frac{\rho(w)}{n-1}\}^2 dF(w)$$

where $\rho(w)$ has the property that $\rho^2(w) = w^2$.

Thus

$$E(S^2) = \frac{a^2}{n^2} F(C) + \frac{a^2}{n^2}(1 - F(C))$$

$$- \frac{2a}{n(n-1)} \int_C^\infty \rho(w) dF(w) + \frac{1}{(n-1)^2} \int_C^\infty w^2 dF(w)$$

and also when $w > C > 0$, $\rho(w) = w$.

Therefore $E(S^2) = \frac{a^2}{n^2} + \frac{1}{n\nu} \int_C^\infty (\frac{nw^2}{\nu} - 2aw) dF(w)$.

Using this equation and (5.11) and (5.12) we see that the protection, R, is given by

$$R = \frac{n}{\nu(n+a^2)} \int_C^\infty (2aw - \frac{nw^2}{\nu}) dF(w)$$

and when $n = 3$, $\nu = 2$

$$R = \frac{3}{2(3+a^2)} \int_C^\infty (2aw - \frac{3w^2}{2}) dF(w)$$

or returning to our original variable (quite valid in $w > 0$)

$$(5.14) \qquad R = \frac{3}{2(3+a^2)} \int_C^\infty (2az - \frac{3z^2}{2}) dF(z) \ .$$

As a check for this formula we see that if $a = 0$ we get the equation (5.4) for the premium with a change of sign. This is to be expected since under the hypothesis of no spurious observation, protection and premium are numerically equal. It remains to calculate $F(z)$. As in the case of no spurious observation, the probability that (x_1, x_2) lies inside the hexagon of Fig. 5.1, page 29, is equal to $F(z)$. Thus to determine $F(z)$ we integrate the probability element (5.13) over the interior of the hexagon. Using an IBM 650 Digital Computer, the author, using values of C which gave 5, 4, 2, 1, 1/2 percent premiums in the null case, computed R the protection for values of a equal to 1/2, 1, 2, 3. The values are tabulated in 5.3. Integration and computational notes are given in the appendix.

5.13 Rule 2. n = 3: No Spurious Observation

The calculation of premium and protection under Rule 2 are more difficult. For general applications Anscombe suggested the use of Monte Carlo techniques. We simply outline the geometric solution when $n = 3$ and there are no spurious observations (σ^2 is assumed to be unity).

Suppose $M = 3$ so that $|z_3| > |z_1|, |z_2|$

$$\hat{\mu} = \bar{y} \qquad \text{if} \quad |z_3| < C_3$$
$$= \bar{y} - \frac{z_3}{2} \quad \text{if} \quad |z_3| > C_3 \quad \text{and} \quad |z_1 - z_2| < 2C_2.$$

There is no estimate if $|z_3| > C_3$ and $|z_1 - z_2| > 2C_2$. Here the condition $|z_1 - z_2| \lesseqgtr 2C_2$ is equivalent to $|z_i| \lesseqgtr C_2$ $(i = 1, 2)$ and is thus the rejection criterion at the second application of Rule 1. We recall that applying Rule 2 here we reject 0, 1 or all 3 observations

when $$x_1 = (z_2 - z_1)\frac{1}{\sqrt{2}}$$

$$x_2 = (2z_3 - z_2 - z_1)\frac{1}{\sqrt{6}}$$

$|z_3| > C_3$ implies (x_1, x_2) lies outside the hexagon in Fig. 5.1. $|z_1 - z_2| > 2C_2$ implies (x_1, x_2) lies inside one of six similar sectors of angle $\pi/3$ and vertices distant $2C_2\sqrt{2}$ from the origin as shown in Fig. 5.3 (Page 40). The sectors lie inside or outside the hexagon depending on the sign of $C_3 - 2C_2$. Since no observation is spurious, the joint density function of (x_1, x_2) is a spherical

normal, centre the origin and the probability of rejecting the whole sample is

$$\frac{1}{2\pi}\int_B d\Theta e^{-\frac{1}{2}r^2}\, d(\tfrac{1}{2}r^2)$$

where B is the area outside the hexagon and inside the sectors. After a determination of the integral of the probability density over A, the area exterior to both the hexagon and the sectors, we can calculate the variance of $\hat{\mu}$ subject to the condition that not all observations are rejected and hence obtain a measure of the premium in a way similar to that used for Rule 1.

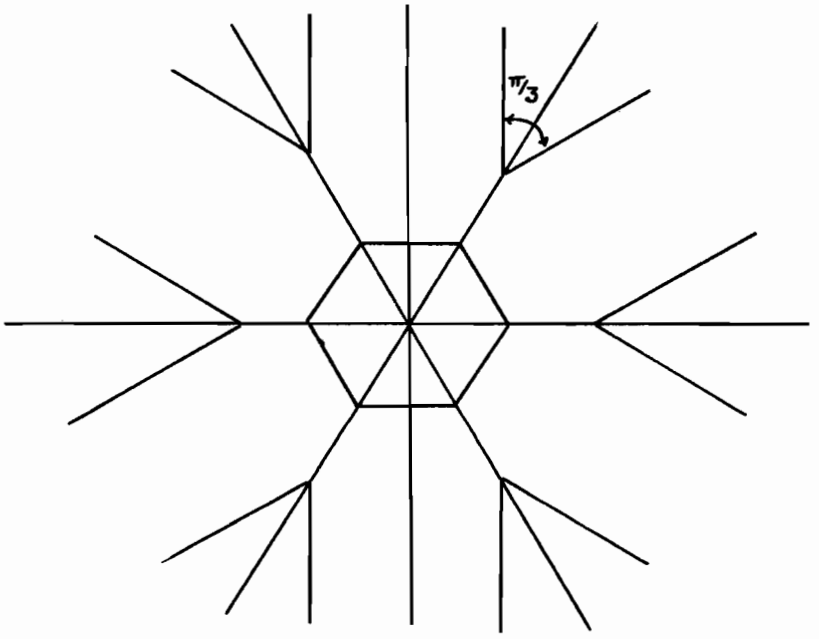


Figure 5.3

5.2 Complex Patterns of Data

We now generalise the above, somewhat, and consider the problem of rejection rules to complex systems of data. Such considerations clearly have wide application in the design of experiments. Hence we assume that the observations (if none are spurious) constitute samples from independent normal distributions of common (known) variance σ^2 and means which are given linear functions of certain unknown parameters; it being the purpose of the experiment to estimate these parameters by the method of least squares -- which method we note is equivalent to maximum likelihood estimation under normal distribution theory. If we assume that this estimation has been effected, we can compute residuals.

The first point made by Anscombe is that, in general, the distribution and in particular the variance of a residual z_i depends on i. This leads to severe complications in any attempted outlier analysis and for the sake of simplicity we consider below only such designs which have residuals with equal variances, i.e. $\sigma^2 v/n$. All ordinary factorial designs with equal replications, balanced incomplete block designs and Latin Squares have this property.

Another important factor which influences the theory is correlation between the various residuals. In some designs there exist pairs of residuals having correlation coefficients equal to ± 1. As an illustration of this,

Anscombe considers the 3 x 3 Latin Square design in which the residuals are equal in sets of three. Twenty-seven pairs have correlation coefficient equal to -1/2 and the remaining nine have $\rho = 1$. Suppose we are of the opinion that one residual value is excessively large. Then this suspected 'maverick' appears three times in the list of residuals and we are faced with the problem of deciding which one of the three observations, having this residual value, to reject from the sample, remembering that estimates of row, column and treatment effects will differ greatly according to which observation we chose to reject. One of the main assumptions, paragraph (five) was that our rejection rules were to be impartial. If we adhere rigidly to this supposition, we cannot reject one observation from the three in preference to the other two. If all three were rejected, we would have too few observations left for the estimation and thus would probably 'reject' the whole experiment. In practice an experimenter might well use some small piece of prior information in order to break such a deadlock. If, therefore, we wish to apply a rejection rule based on residuals to the observations from a certain experiment, the latter should be designed in such a way that there are no pairs of residuals having correlations ± 1. This may be extended as illustrated by the following statement: The probability of rejecting an 'innocent' observation rather than the true spurious reading is high if the correlation

between their residuals is 'high'. Clearly a more thorough investigation would be necessary to define the word 'high' in this statement. Professor Anscombe gives details of residual correlations for particular designs. We now give some more notation and matrix theory necessary for the discussion of the rejection rules.

Represent the observations $\{y_i\}$ by the $n \times 1$ column vector y.

There are $n - \nu$ parameters (unknown) which we denote by a vector Θ $[(n-\nu) \times 1]$

Let the coefficient matrix be the $n \times (n-\nu)$ dimensioned matrix A such that if no observation is spurious

$$E(y) = A\Theta$$

and y has a spherical normal distribution. Since the observations are independent A has rank $n - \nu$.

We define the matrix V by the equation

$$A^T A = V^{-1} \tag{5.15}$$

Suppose after least squares estimation of Θ, the residuals z are given by the matrix equation

$$z = Qy. \tag{5.16}$$

Then

$$Q = I_n - AVA^T.$$

To see this consider the sum of squares of residuals

i.e. $(y - A\theta)^T(y - A\theta)$.

Therefore the equation giving least squares estimate $\hat{\theta}$ is

$(y - A\hat{\theta})^T A = 0$ (Equivalent to $n - \nu$ linear equations)

Therefore $y^T A = (A\hat{\theta})^T A$

$\hat{\theta}^T A^T A = y^T A$

$\hat{\theta}^T = y^T AV$ [using (5.15)]

$\hat{\theta} = VA^T y$ [using fact that V is symmetric]

Then $z = y - A\hat{\theta}$

$= y - AVA^T y = (I_n - AVA^T)y$, hence Q.

Note Q is $n \times n$ and symmetric.

Further

$QQ = (I-AVA^T)(I-AVA^T) = I+AVA^TAVA^T-2AVA^T = I+AVA^T-2AVA^T$ from (5.15)

$= Q$. Therefore Q is idempotent.

The y_i are independent but the z_i satisfy $n - \nu$ linear relations which gives us that the rank of Q is ν.

The variance-covariance matrix of the residuals (using the fact that residual means are zero) is:

$$E(zz^T) = E\{Qy(Qy)^T\} = E(Qyy^TQ) = QE(yy^T)Q$$
$$= Q\sigma^2 IQ = \sigma^2 QQ = Q\sigma^2$$

(since Q idempotent.)

If all residuals have same variance $\nu\sigma^2/n$ it is clear that each element in the principal diagonal of Q is equal to ν/n.

Considering, then, only patterns such that when no observations are spurious, residual variances are equal and residual correlation coefficients are not 'close' to $\pm$ 1, we proceed to formulate rejection rules. We use the statements of Rule 0 and Rule 1, as on page 23 , but the last sentence of each is amended to read:

'Estimate the unknown parameters from the retained observations by the method of least squares'.

It will generally be necessary to effect this process in two stages. Firstly, estimates of missing observations or those held to be spurious are computed, then the actual parameter estimation is effected.

Suppose y_j has to be pre-estimated. If we use any arbitrary value of this observation in the least squares parameter estimation let the corresponding residuals be $\{z_i\}$. Now replace y_j by $y_j - \eta$ where it is the purpose to estimate η.

Substituting in (5.16) the residuals become $\{z_i'\}$ where

(5.17) $$z_i' = z_i - \eta q_{ij} \qquad \text{(The } q_{ij} \text{ are elements of } Q\text{)}$$

Minimising the sum of squares of these new residuals, we obtain an estimate of η given by

$$\hat{\eta} = \frac{\sum_i z_i q_{ij}}{\sum_i q_{ij}^2} .$$

Note next that since Q is symmetric and idempotent

$$\sum_i q^2_{ij} = q_{jj} = \nu/n \qquad \text{(using our assumptions)}$$

and $$\sum_i z_i q_{ij} = \sum_i q_{ij} \sum_k q_{ik} y_k$$

$$= \sum_k \{y_k \sum_i q_{ij} q_{ik}\}$$

$$= \sum_k y_k q_{jk} = z_j .$$

(5.18) Therefore $\hat{\eta} = (\frac{n}{\nu}) z_j$

Therefore providing that no other observation is spurious, by using the original data with $y_j - (\frac{n}{\nu}) z_j$ substituted for y_j, We shall obtain the correct least squares estimates of the unknown parameters.

Using (5.17) and (5.18) the new residuals are

$$z_i' = z_i - \frac{n}{\nu} z_j q_{ij} .$$

From the var-cov. matrix $Q\sigma^2$ we have $\text{cov}(z_i z_j) = q_{ij}\sigma^2$ Therefore the correlation coefficient (ρ_{ij}) of z_i and z_j is given by

$$\frac{\operatorname{cov} z_i z_j}{\{\operatorname{var} z_i \operatorname{var} z_j\}^{1/2}} = \frac{q_{ij}\sigma^2}{\{\frac{\nu\sigma^2}{n}\cdot\frac{\nu\sigma^2}{n}\}^{1/2}} = \frac{n}{\nu} q_{ij}.$$

Thus we can write

$$z_i' = z_i - z_j p_{ij} \quad \text{(note that if } i = j,\ z_i' = 0)$$

$$\operatorname{var}(z_i') = (1 - \rho_{ij}^2)\frac{\nu\sigma^2}{n} \tag{5.19}$$

Thus we have a method of adaption of the Simple Sample rejection Rule 1 to more complex patterns. We cannot use the above theory to apply Rule 1 to the new residuals and thus extend Rule 2 to complex patterns because (5.19) shows that the equivariance residual assumption is no longer valid. We here quote Professor Anscombe, "Provided the correlations are small, it may seem reasonable -- it is certainly simplest -- to take no account of the changes in variance in formulating the rule which would run:

Apply Rule 1. If an observation is rejected compute the revised residuals and apply Rule 1 again and so on. Finally compute the least squares estimates of the unknown parameters from the retained observations."

It remains to find a criterion by means of which the premium and protection for the rejection rule can be defined. In order to do this the parameters are split into two groups, those 'of interest' and the remainder. It is clear that the experiment may have been performed solely

to estimate a certain set of parameters from the model while we are not particularly 'interested' in the remainder. Our critical statistic will be the determinant of the variance-covariance matrix of the parameters of interest. We note that an orthogonal transformation exists which will diagonalize this variance-covariance matrix, i.e. the covariances of parameter estimates can be made zero; the product of variances of estimates after the transformation being equal to the determinant of the original matrix, and hence the use of this statistic. Thus a possible (one can suggest many others) definition of the premium charged by the rejection rule is the proportional magnification of the determinant of the variance-covariance matrix of the parameters of interest. The corresponding protection given by the Rule is the proportional decrease of the value of this determinant when spurious readings are present.

5.3 Results: Professor Anscombe's Theory

n = 3. Rule 1.

C	Premium	Approx. Premium	R-Rate*	1/R-Rate	Protection When a =			
					.5	1	2	3
2.39038	.05	.05336	.00319	313.5	-.04135	-.02874	-.0009827	+.0004662
2.46002	.04	.04242	.00243	411.0	-.03390	-.02463	-.001206	+.0004095
2.66184	.02	.02088	.00107	938.8	-.01826	-.01505	-.001455	+.0002675
2.84623	.01	.01032	.000475	2107	-.009977	-.009015	-.001342	+.0001673
3.01727	.005	.00512	.000214	4667	-.005202	-.005312	-.001098	+.00009915

* R-Rate indicates Rejection Rate

The signs in the protection tabulation show that in the case considered no saving in variance is to be obtained by the application of the defined rejection rules when a = .5, 1, or 2. A slight saving is seen to exist when a = 3. This saving would be increased if higher values of a were considered.

APPENDIX: INTEGRATION AND COMPUTATION: PROFESSOR ANSCOMBE'S PAPER

The computation of the premium from equation (5.8), page 33 , was straightforward numeric integration using Simpson's Rule. The results, i.e. values of C which give rise to 5, 4, 2, 1, 1/2 percent premiums respectively, are tabulated (page 49). For comparison, the approximate premiums using these values of C calculated from Anscombe's empiric formula equation (5.9), page 33, are also listed.

To compute the protections for varying values of a using these values of C was more difficult and a brief account of the method used is given here. As stated on page 38 ,

$$F(z, a) = \iint_I \frac{1}{2\pi} e^{-\frac{1}{2}\{x_1^2+(x_2-\sqrt{\frac{2}{3}}a)^2\}} dx_1 dx_2.$$

Where I is the interior of the hexagon, figure 5.1, we write F(z, a) to make it clear that in the case where spurious observations occur, F(z) depends also on a. Then

$$F(z, a) = 2\iint_{I_1} \frac{1}{2\pi} e^{-\frac{1}{2}\{x_1^2+(x_2-\sqrt{\frac{2}{3}}a)^2\}} dx_1 dx_2$$

$$+ 2\iint_{I_2} \frac{1}{2\pi} e^{-\frac{1}{2}\{x_1^2+(x_2-\sqrt{\frac{2}{3}}a)^2\}} dx_1 dx_2,$$

where I_1 and I_2 are the regions bounded by the interior of the hexagon and the first and fourth quadrants respectively. See figure 5.4.

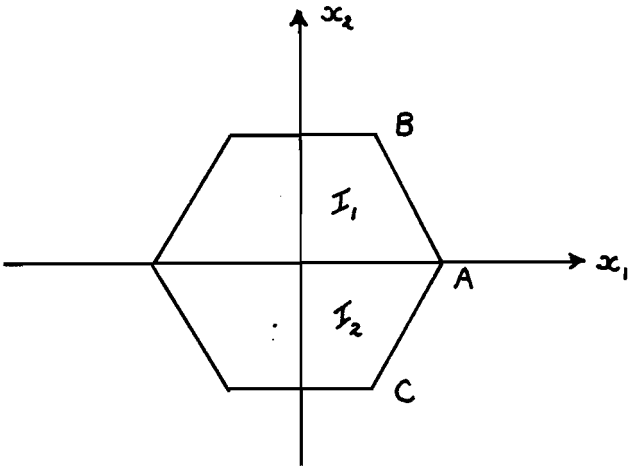


Figure 5.4

We make the transformations

$$\eta = \sqrt{\frac{2}{3}}a \; ; \quad b = z\sqrt{\frac{3}{2}} \; .$$

Then the equations of the lines AB, AC are respectively,

$$x_1 = \frac{1}{\sqrt{3}}(2b - x_2)$$

$$x_2 = \frac{1}{\sqrt{3}}(2b + x_2)$$

Therefore

$$F(b, \eta) = \frac{1}{\pi}\int_0^b e^{-\frac{1}{2}(x_2-\eta)^2} dx_2 \int_0^{(2b-x_2)/\sqrt{3}} e^{-\frac{1}{2}x_1^2} dx_1$$

$$+ \frac{1}{\pi}\int_{-b}^0 e^{-\frac{1}{2}(x_2-\eta)^2} \int_0^{(2b+x_2)/\sqrt{3}} e^{-\frac{1}{2}x_1^2} dx_1 .$$

Therefore

$$F(b,\eta) = \frac{1}{\sqrt{2\pi}} \int_0^b e^{-\frac{1}{2}(x_2-\eta)^2} \operatorname{erf}\left(\frac{2b-x_2}{\sqrt{6}}\right) dx_2$$

$$+ \frac{1}{\sqrt{2\pi}} \int_{-b}^0 e^{-\frac{1}{2}(x_2-\eta)^2} \operatorname{erf}\left(\frac{2b+x_2}{\sqrt{6}}\right) dx_2.$$

In second integral substitute $x_2 = -t$. Then

$$F(b,\eta) = \frac{1}{\sqrt{2\pi}} \int_0^b e^{-\frac{1}{2}(x_2-\eta)^2} \operatorname{erf}\left(\frac{2b-x_2}{\sqrt{6}}\right) dx_2$$

$$+ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(-t-\eta)^2} \operatorname{erf}\left(\frac{2b-t}{\sqrt{6}}\right)(-dt).$$

Changing the variable back to x_2 and combining the integrals we have

$$F(b,\eta) = \frac{1}{\sqrt{2\pi}} \int_0^b \operatorname{erf}\left(\frac{2b-x_2}{\sqrt{6}}\right) \left\{e^{-\frac{1}{2}(x_2-\eta)^2} + e^{-\frac{1}{2}(-x_2-\eta)^2}\right\} dx_2.$$

$$= \frac{1}{\sqrt{2\pi}} \int_0^b \operatorname{erf}\left(\frac{2b-x_2}{\sqrt{6}}\right) e^{-\frac{1}{2}x_2^2} e^{-\frac{1}{2}\eta^2} \left(e^{x_2\eta} + e^{-x_2\eta}\right) dx_2$$

$$= \sqrt{\frac{2}{\pi}} \int_0^b \operatorname{erf}\left(\frac{2b-x_2}{\sqrt{6}}\right) e^{-\frac{1}{2}x_2^2} e^{-\frac{1}{2}\eta^2} \cosh x_2\eta \, dx_2.$$

Transforming this back to terms of a and z

$$(A1)\quad F(z, a) = \sqrt{\frac{2}{\pi}}\, e^{-\frac{a^2}{3}} \int_0^{z\sqrt{\frac{3}{2}}} \cosh\left(\sqrt{\frac{2}{3}} a x_2\right) e^{-\frac{1}{2}x_2^2} \operatorname{erf}\left(z - \frac{x_2}{\sqrt{6}}\right) dx_2.$$

The protection, R, from equation (5.14), page 38, is given by

$$(A2)\quad R = \frac{3}{2(3+a^2)} \int_C^{\infty} (2az - \frac{3z^2}{2}) \frac{dF(z,a)}{dz}\, dz.$$

From (A1) differentiating under the integral sign

$$\frac{\partial F(z,a)}{\partial z} = \sqrt{\frac{2}{\pi}}\, e^{-\frac{a^2}{3}} \left\{ \sqrt{\frac{3}{2}}\, e^{-3\frac{z^2}{4}} \cosh az \operatorname{erf} \frac{z}{2} + \frac{2}{\sqrt{\pi}} \int_0^{z\sqrt{\frac{3}{2}}} \cosh \sqrt{\frac{2}{3}} a x_2 e^{-\frac{1}{2}x_2^2 - (z - x_2/\sqrt{6})^2} dx_2 \right\}.$$

This may be reduced to

$$(A3)\quad \frac{\partial F(z,a)}{\partial z} = \frac{2\sqrt{3}}{\pi}\, e^{-\frac{a^2}{3} - 3\frac{z^2}{4}} \cosh az \int_0^{z/2} e^{-p^2} dp + \frac{4\sqrt{3}}{\pi}\, e^{-\frac{a^2}{3} - 3\frac{z^2}{4}} \cosh \frac{az}{2} \int_0^{z/2} \cosh ap\, e^{-p^2} dp.$$

Thus from (A2) and (A3) we see that to calculate the protection we have to compute an integral of the form:

$$(A4)\quad R = \int_C^{\infty} f(z)dz \left\{ g_1(z) \int_0^{z/2} v_1(p)dp + g_2(z) \int_0^{z/2} v_2(p)dp \right\}$$

where the functions $f(z)$, $g_1(z)$, $g_2(z)$, $v_1(p)$, $v_2(p)$ also depend on a. If

$$g_1(z) \int_0^{z/2} v_1(p)dp + g_2(z) \int_0^{z/2} v_2(p)dp = k(z).$$

We then have an integral of the form

$$(A5)\qquad R = \int_C^{\infty} f(z)k(z)dz.$$

We first read into the computer a pair of values (a, C). It was decided to make the approximation

$$\int_C^{\infty} f(z)k(z)dz = \int_C^{6C} f(z)k(z)dz$$

noting that the integrand converges to zero rapidly as z gets large. We shall obtain an estimate of the error

$$\int_{6C}^{\infty} f(z)k(z)dz.$$

Simpson's Rule for numeric integration generalised over n equal intervals h reads

$$\int_0^{nh} q(x)dx = \frac{h}{3}\{q(0) + 4q(h) + 2q(2h) + \cdots + 4q((n-1)h) + q(nh)\}$$

where n is even. Since we are concerned with a double integral we shall expect a double sum in the numeric evaluation.

$$\int_0^{z/2} v_i(p)dp = \frac{h}{3}\{v_i(0) + 4v_i(h) + \cdots + 4v_i(\tfrac{z}{2}-h) + v_i(\tfrac{z}{2})\}. \quad i = 1, 2 \tag{A6}$$

For integrations over p we used $h = C/100$. We sum up values of the function $f(z)k(z)$ between $z = C$ and $z = 6C$ at intervals of h' weighted with appropriate Simpson's Rule weights, i.e.,

$$\sum f(z)k(z) = f(C)k(C) + 4f(C+h')k(C+h') + \cdots + 4f(6C-h')k(6C-h') + f(6C)k(6C).$$

Note that the $k(z)$'s contain integrals over p which depend on z. In order to generate the successive $k(C)$, $k(C+h')$ etc., it is not necessary to repeat the whole process as given by (A6). If $h' = 4h$ it is possible to add three terms only to the sum in (A6) to obtain the new integral:

$$\int_0^{z/2} v_i(p)dp = \frac{h}{3}\{v_i(0) + 4v_i(h) + \cdots + 4v_i(\tfrac{z}{2}-h) + v_i(\tfrac{z}{2})\} \quad i = 1, 2 \tag{A6}$$

and $$\int_0^{(z+4h)/2} v_i(p)dp = \int_0^{(z/2)+2h} v_i(p)dp$$

$$= \frac{h}{3}\{v_i(0) + 4v_i(h) + \cdots + 4v_i(\tfrac{z}{2}-h)$$

$$+ 2v_i(\tfrac{z}{2}) + 4v_i(\tfrac{z}{2}+h) + v_i(\tfrac{z}{2}+2h)\} \quad i = 1, 2$$

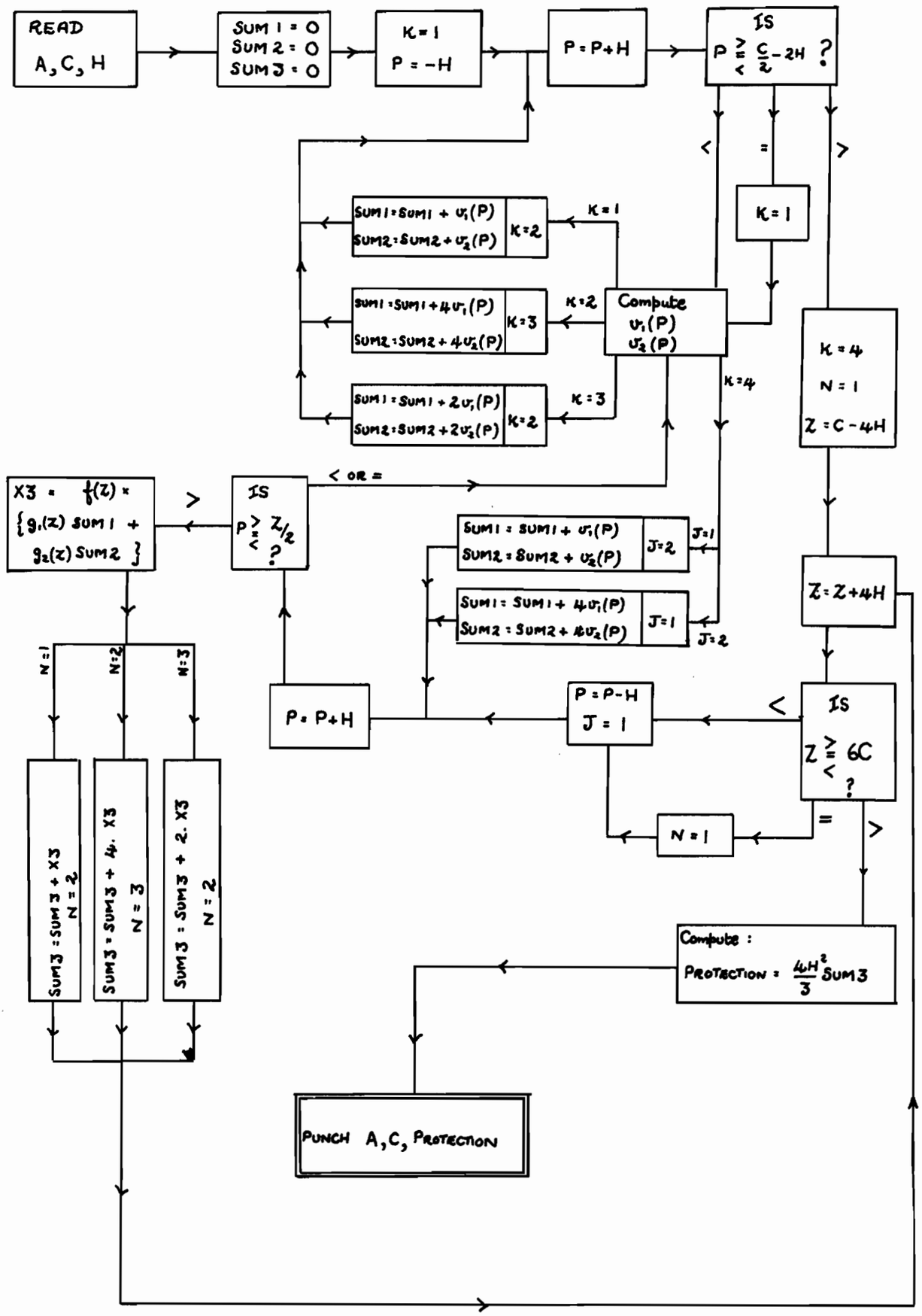
It is apparent that the sum in the new integral is formed by adding on the terms

$$v_i(\tfrac{z}{2}), \quad v_i(\tfrac{z}{2} + h), \quad v_i(\tfrac{z}{2} + 2h)$$

weighted with 1, 4, 1 respectively. Similarly we can obtain the integrals of the $v_i(p)$ $i = 1, 2$ in $(0, \frac{z}{2} + rh)$ from those over $(0, \frac{z}{2} + (r-2)h)$ by the addition of only three suitably weighted terms. We make use of this fact in the computation. A flow chart for the computation using the expression as in (A4) follows.

FLOW CHART FOR PROTECTION COMPUTATION.

$$\int_C^{6C} f(z) \left\{ \int_0^{z/2} \left(g_1(z)\, v_1(p) + g_2(z)\, v_2(p) \right) dp \right\} dz$$



We now obtain an estimate of the error incurred in the calculation of protection by the above computation because we used the approximation

$$\int_{C}^{6C} f(z)k(z)dz = \int_{C}^{\infty} f(z)k(z)dz$$

i.e., we require an estimate of

$$\int_{6C}^{\infty} f(z)k(z)dz \ .$$

Considering (A3) and (A4), let

$$E = \int_{6C}^{\infty} f(z)dz \left\{ g_1(z) \int_{0}^{z/2} v_1(p)dp + g_2(z) \int_{0}^{z/2} v_2(p)dp \right\} .$$

Now $v_1(p)$, $v_2(p)$ are of the order of e^{-p^2}. Since we consider values of $z > 6C$ in the integration over z, we assume that integrals over p from 0 to $z/2$ may be all approximated by integrals over $(0, \infty)$ so that approximately

$$E = \int_{6C}^{\infty} f(z)dz \left\{ \int_{0}^{\infty} (g_1(z)v_1(p) + g_2(z)v_2(p))dp \right\}.$$

The inner integral can then be simply evaluated and is equal to

$$d(\cosh az + 2e^{\frac{a^2}{4}} \cosh \frac{az}{2})$$

where d is a constant. Thus our estimate becomes (see (A2) and (A3))

$$E = D\int_{6C}^{\infty}(2az - \frac{3z^2}{2})e^{-3\frac{z^2}{4}}(\cosh az + 2e^{\frac{a^2}{4}}\cosh\frac{az}{2})dz,$$

where D is another constant.

This may now be split into four integrals by expanding the integrand. Each of these may be evaluated fairly simply. The four integrals are of the following form:

$$I_1 = C_1\int_{6C}^{\infty} ze^{-3\frac{z^2}{4}}\cosh az\, dz$$

$$I_2 = C_2\int_{6C}^{\infty} ze^{-3\frac{z^2}{4}}\cosh\frac{az}{2}\, dz$$

$$I_3 = C_3\int_{6C}^{\infty} z^2e^{-3\frac{z^2}{4}}\cosh az\, dz$$

$$I_4 = C_4\int_{6C}^{\infty} z^2e^{-3\frac{z^2}{4}}\cosh\frac{az}{2}\, dz$$

It is only necessary to evaluate I_1 and I_3 as I_2 and I_4 follow from these respectively by replacing a by a/2.

The final estimate obtained after neglecting a set of terms which were of the order of $e^{-27C^2}\cosh 6aC$ (in the worst case this expression is approximately equal to e^{-100}) was:

$$E = \frac{3(2a^2+1)}{4(a^2+3)} \operatorname{erf} v_1 + \frac{(3-2a^2)}{4(a^2+3)} \operatorname{erf} v_2$$

$$+ \frac{(5a^2+6)}{4(a^2+3)} \operatorname{erf} v_3 + \frac{3(2-a^2)}{4(a^2+3)} \operatorname{erf} v_4 - \frac{3}{2}$$

where $v_1 = \sqrt{3}(3C + a/3)$, $v_2 = \sqrt{3}(3C - a/3)$

$v_3 = \sqrt{3}(3C + a/6)$, $v_4 = \sqrt{3}(3C - a/6)$.

This error is very small, in fact so small that it makes no significant difference to the computed values of the protection.

If the above error estimate were repeated, integrating with respect to z over $(C, 4C)$ instead of $(C, 6C)$ the error will still be negligible and thus it was concluded that it is sufficient to integrate over $(C, 4C)$ and eliminate the error (found from the fact that protection is numerically equal to premium already computed under the null hypothesis $a = 0$) by decreasing H in the double summation using Simpson's Rule. We used $H = C/300$ in the revised program and obtained accuracy approximately 99.5 percent. The maximum error recorded was .56 percent.

Acknowledgement

The author wishes to express his gratitude to Professor I. Guttman under whose guidance this thesis was written. His enthusiasm for this topic proved very stimulating and his experience in outliers, invaluable.

Thanks are also extended to Dr. G. Bach, Mathematical Head of the McGill University Computing Centre for help and advice in the computation of the constants in Professor Anscombe's work.

Lastly, a word of appreciation is expressed to all the staff of the McGill University Computing Centre for their assistance and cooperation at all times.

Bibliography

1. Anscombe, F. J. 1960 "Rejection of Outliers." Technometrics Vol. 2. No. 2., pp. 123-147.

2. Grubbs, F. E. 1950 "Sample criteria for testing outlying observations." Annals of Mathematical Statistics 21, pp. 27-58.

3. Lieblien, J. 1952 "Properties of certain statistics involving the closest pair in a sample of three observations." Journal of Research of National Bureau of Standards 48, pp. 255-68.

4. Nair, K. R. 1948 "Distribution of extreme deviate from sample mean and studentised form". Biometrika, vol. 35, pp. 118-144.

5. Rider, P. R. 1933 "Criteria for rejection of observations." Washington University Studies. New Series. Science and Technology, No. 8.