Use of the 27% Rule in Experimental Design

by

George P. McCabe, Jr.*

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #499

July 1977

Revised July 1978

*George P. McCabe, Jr. is Associate Professor of Statistics and Head of Statistical Consulting at Purdue University, West Lafayette, Indiana 47907. Part of this research was completed while the author was on a sabbatical leave at the Department of Statistics, Princeton University, Princeton, NJ 08540.

USE OF THE 27% RULE IN EXPERIMENTAL DESIGN

George P. McCabe, Jr.

Purdue University

ABSTRACT

For a sample of N iid bivariate normal random variables, upper and lower groups are defined to be all observations with values above the $(1-\alpha)$-th percentile and below the α-th percentile, respectively, on the first variable. A t-type statistic is used to compare the means of the two groups on the second variable. An expression is given for the limiting value, as N approaches infinity, of this t-statistic. For any given correlation in the bivariate normal distribution, the value of α which maximizes the limiting value of the t-statistic can be found. These optimal values of α range from 0.27 to 0.14, with the higher values corresponding to small correlations. For the moderate correlations frequently encountered in practice, $\alpha=0.25$ is a reasonable choice.

1. INTRODUCTION

In research problems, particularly, although not limited to the social sciences, where the collection of data is expensive and or tedious, the following type of sample selection is sometimes used.

A collection of subjects, hopefully a sample in some sense, are measured on a variable, say X. Individuals scoring in the lower third of the sample are designated low-X and individuals scoring in the upper third of the sample are designated high-X. No further measurements are made on the middle group. For the low-X and high-X subjects information on the X-variable score is discarded and only the designations low-X and high-X are kept. Other variables are measured on these selected subjects in an experimental design with low-X and high-X being treated as a two-level factor.

Although this procedure seems reasonable, the consequences of such selection are by no means obvious. Since the choice of thirds in breaking down the data seems somewhat arbitrary, other rules, such as using the lower and upper fourths may be more efficient under given circumstances.

In this paper the most elementary version of this problem is studied. Suppose the experimental design consists of measuring a variable Y on all subjects in the two selected groups. The usual two-sample t-statistic is then used to compare the performance of the high-X and low-X groups on the variable Y. If there is some sort of monotonic relationship between X and Y, then, given sufficient observations, one would expect the t-statistic to give reasonable interpretable results.

In this context, the t-statistic is used as an indicator of correlation between X and Y. To further simplify this special case, it is assumed that X and Y follow a joint normal distribution.

2. BACKGROUND

In 1939, Truman Kelley [1] proposed and discussed a 27% rule in the following context. Suppose that a new item is to be studied for possible inclusion in a test. The item responses are classified as correct or incorrect. A normally distributed criterion which is correlated with a true score can be measured on an initial group. The true score is assumed to be related to the ability to respond correctly to a valid item. Kelley states "twenty-seven percent should

be selected at each extreme to yield upper and lower groups which are most indubitably different with respect to the trait in question." He shows that the difference in tail means times the square root of the number of tail observations is maximized by using a twenty-seven percent rule. Although the proofs are not rigorous by current standards, the results are reasonable in the context given.

In 1946, Frederick Mosteller [2] studied methods for constucting estimates of parameters from counts of data falling into certain categories. He was interested in methods which could be used with punched cards and a counting sorter. Suppose we have data from a bivariate normal distribution with known means and variances and an estimate of the correlation is desired. The maximum likelihood estimator based on the numbers of observations falling in the four corners of the plane determined by the lines $x=\mu_X+k\sigma_X$ and $y=\mu_Y$ is to be used. Mosteller showed that the asymptotic variance of the estimated correlation is minimized by choosing $k=0.6121$ when the true correlation (ρ) is zero. This rule corresponds to dividing the X sample at the (true) twenty-seventh (27.02) percentile.

In 1964, Ross and Weitzman [4] futher studied the variance of the maximum likelihood estimator when $\rho\neq 0$. They found that the variance is minimzed using about 27%, for ρ up to about 0.6. As ρ increases, the cut-off point increases, reaching 35% at $\rho=.90$. They also note that the variance curve as a function of the cutoff point is rather flat in the region from 15% to 50% for all values of ρ.

3. NOTATION AND BASIC RESULTS

Let (X_i,Y_i), $i=1,\ldots,N$ be iid bivariate normal random variables. In what follows, it can be assumed without loss of generality that the means are zero and the variances are one. Let $X_{(1)}\leq X_{(2)}\leq\ldots\leq X_{(N)}$ denote the order statistics of the X sample. For any $n \leq N/2$, let

$$X_{ij}=X_{(j)} \qquad \text{for } j=1,\ldots,n;$$

and

$$X_{2j}=X_{(n-j+1)} \qquad \text{for } j=1,\ldots,n.$$

The Y observation paired with X_{ij} will be denoted by Y_{ij} for $i=1,2$ and $j=1,\ldots,n$.

In this section, we let α be a fixed number ($\alpha\varepsilon(0,.5]$) representing the fraction of the N observations in the upper (i=2) and lower (i=1) groups. Thus, we set $n=[\alpha N]$ where $[\cdot]$ denotes the greatest integer function. Let

$$\bar{Y}_i(\alpha)=n^{-1}\ \Sigma_{j=1}^{n}\ Y_{ij}, \quad i=1,2; \tag{3.1}$$

and

$$s_i^2(\alpha)=n^{-1}\ \Sigma_{j=1}^{n}(Y_{ij}-\bar{Y}_i)^2, \quad i=1,2. \tag{3.2}$$

In Section 4, the choice of α is investigated. To this end, we now study the asymptotic behavior (as $N\to\infty$) of the expected values of $\bar{Y}_i$ and s_i^2 for fixed values of α.

Let $\phi(x)$ and $\Phi(x)$ denote the standard normal density and distribution function, respectively. For $x>0$, the Mills ratio is defined by

$$M(x)=(1-\Phi(x))/\phi(x).$$

Also, let a be defined by

$$\alpha=1-\Phi(a).$$

Theorem.

$$\text{(a)} \quad -\lim_{n\to\infty} E\bar{Y}_1(\alpha)=\lim_{n\to\infty} E\bar{Y}_2(\alpha) = \mu(\alpha,\rho) \tag{3.3}$$

and

$$\text{(b)} \quad \lim_{n\to\infty} Es_1^2(\alpha) = \lim_{n\to\infty} Es_2^2(\alpha) = \sigma^2(\alpha,\rho) \tag{3.4}$$

where

$$\mu(\alpha,\rho)=\rho M^{-1}(a),$$

and

$$\sigma^2(\alpha,\rho)=1+\rho^2\ [aM^{-1}(a)-M^{-2}(a)]$$

Proof. First, note that

$$Y_{ij}=\rho X_{ij}+\varepsilon_{ij} \tag{3.5}$$

where the X_{ij} and the ε_{ij} are independent and the ε_{ij} are iid normal with mean zero and variance $(1-\rho^2)$. Therefore,

$$\begin{aligned} E\bar{Y}_2(\alpha) &= En^{-1}\Sigma_{j=1}^{n} Y_{2j} \\ &= \rho En^{-1}\Sigma_{j=1}^{n} X_{2j} \end{aligned} \tag{3.6}$$

Reverting to the order statistic notation, we have

$$n^{-1}\Sigma_{j=1}^{n} X_{2j} = (N/n)N^{-1}\Sigma_{j=1}^{N} J(\tfrac{j}{N+1})X_{(j)} \tag{3.7}$$

where

$$J(Z) = \begin{cases} 0 & \text{if } Z \leq 1-\alpha. \\ 1 & \text{if } Z > 1-\alpha. \end{cases}$$

Results pertaining to quantities such as those in (3.7) can be found in the literature on linear combinations of order statistics. Applying Stigler's [5] Theorem 3, we obtain

$$\lim_{N\to\infty} E\, N^{-1}\, \Sigma_{j=1}^{N}\, J(\tfrac{j}{N+1})\, X_{(j)} = \int_0^1 J(u)\, \Phi^{-1}(u)du \tag{3.8}$$

Here $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal distribution function. The conditions required for the validity of (3.8) are the existence of a first moment for X_i and that $J(\cdot)$ is bounded and continuous almost everywhere. Substituting the above choice for $J(\cdot)$ and letting

$$v=\alpha^{-1}(u-1+\alpha)$$

gives

$$\begin{aligned} \int_0^1 J(u)\Phi^{-1}(u)du &= \int_{1-\alpha}^1 \Phi^{-1}(u)du \\ &= \alpha\int_0^1 \Phi^{-1}(\alpha v+1-\alpha)dv. \end{aligned} \tag{3.9}$$

Now, let $G(x)$ denote the distribution function of a random variable from the upper α tail of a standard normal distribution, i.e.

$$G(x) = \begin{cases} 0 & \text{if } x \leq a \\ \alpha^{-1}(\Phi(x)-1+\alpha) & \text{if } x>a. \end{cases}$$

Since

$$G^{-1}(v)=\Phi^{-1}(\alpha v+1-\alpha)$$

it follows that the last expression in (3.9) is equal to

$$\alpha\int_0^1 G^{-1}(v)dv,$$

which is simply α times the expression that would be obtained from Stigler's Theorem if the underlying distribution had been $G(x)$ and the function $J(\cdot)$ was identically one. Therefore,

$$\begin{aligned}\alpha\int_0^1 G^{-1}(v)dv &= \alpha\int_\alpha^\infty x dG(x)\\ &= \int_x^\infty x d\Phi(x)\\ &= \alpha M^{-1}(a)\end{aligned}$$

So,

$$\lim_{N\to\infty} E\ N^{-1}\Sigma_{j=1}^{n} X_{2j} = \alpha M^{-1}(a) \tag{3.10}$$

Combining the above and noting that $N/n \to \alpha^{-1}$ gives the desired result for $\bar{Y}_2(\alpha)$. By symmetry, the result for $\bar{Y}_1(\alpha)$ follows.

The proof of part (b) is similar. Since

$$E\ s_2^2 = E\ n^{-1}\Sigma_{j=1}^{n} Y_{2j}^2 - E\bar{Y}_2^2(\alpha) \tag{3.11}$$

and, by virtue of (3.5),

$$\begin{aligned}En^{-1}\ \Sigma_{j=1}^{n} Y_{2h}^2 &= \rho^2 En^{-1}\ \Sigma_{j=1}^{n} X_{2j}^2 + En^{-1}\ \Sigma_{j=1}^{n} \varepsilon_j^2\\ &= \rho^2\ En^{-1}\ \Sigma_{j=1}^{n} X_{2j}^2 + (1-\rho^2)\end{aligned} \tag{3.12}$$

it is sufficient to consider

$$En^{-1}\ \Sigma_{j=1}^{n} X_{2j}^2.$$

By a small modification of the proof of Stigler's Theorem, it follows that

$$\lim_{N\to\infty} EN^{-1}\ \Sigma_{j=1}^{N} J\left(\frac{j}{N+1}\right) X_{(j)}^2 = \int_0^1 J(u)\Phi^{-2}(u)du \tag{3.13}$$

with the previously mentioned conditions on $J(\cdot)$. Using the change of variables as given for part (a), it follows that

$$\begin{aligned}\int_{1-\alpha}^1 \Phi^{-2}(u)du &= \int_a^\infty x^2 d\Phi(x)\\ &= \alpha(1+\rho^2 aM^{-1}(a)),\end{aligned} \tag{3.14}$$

and hence from (3.12),

$$\lim_{n\to\infty} En^{-1} \Sigma_{j=1}^{n} Y_{2j}^2 = \rho^2(1+\rho^2 aM^{-1}(a)) + (1-\rho^2). \tag{3.15}$$

Consider now the last term in (3.11). Using (3.5), it follows that

$$\begin{aligned} E\bar{Y}_2^2(\alpha) &= E(n^{-1}\Sigma_{j=1}^{n}(\rho X_{2j}+\varepsilon_j))^2 \\ &= E(n^{-1}\rho^2\Sigma_{j=1}^{n} X_{2j})^2 + E(n^{-1}\Sigma_{j=1}^{n} \varepsilon_j)^2 \\ &= \rho^2 E(n^{-1}\Sigma_{j=1}^{n} X_{2j})^2 + n^{-1}(1-\rho^2) \end{aligned} \tag{3.16}$$

Now,

$$E(n^{-1}\Sigma_{j=1}^{n} X_{2j})^2 = (E(n^{-1}\Sigma_{j=1}^{n} X_{2j}))^2 + var(n^{-1}\Sigma_{j=1}^{n} X_{2j}) \tag{3.17}$$

The last term in (3.17) approaches zero as $N\to\infty$ since

$$\begin{aligned} \lim_{N\to\infty} N\, var(N^{-1}\Sigma_{j=1}^{n} X_{2j}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(\Phi(x))J(\Phi(y))\,[\Phi(x\wedge y)-\Phi(x)\Phi(y)]dxdy \\ &\le \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [\Phi(x\wedge y)-\Phi(x)\Phi(y)]\, dxdy \\ &= N^{-1} var\,(N^{-1}\Sigma_{i=1}^{N} X_i) \\ &= 1. \end{aligned}$$

The first equality is a consequence of Stigler's [5] Theorem 1 and the inequality is valid since the integrand is nonnegative. Therefore, from (3.10) and (3.17), it follows that

$$\lim_{N\to\infty} E(n^{-1}\Sigma_{j=1}^{n} X_{2j}) = M^{-2}(a) \tag{3.18}$$

Combining the above with (3.11), (3.15) and (3.16) gives

$$\begin{aligned} \lim_{N\to\infty} Es_2^2(\alpha) &= \rho^2 + \rho^2 aM^{-1}(a) + (1-\rho^2) - \rho^2 M^{-2}(a) \\ &= 1+\rho^2(aM^{-1}(a)-M^{-2}(a)), \end{aligned}$$

the desired result. By symmetry, the result for $s_1^2(\alpha)$, follows. Q.E.D.

4. THE CHOICE OF α

Suppose now that one wanted to compare the two sample means $\bar{Y}_1(\alpha)$ and $\bar{Y}_2(\alpha)$. Although these random variables are not independent and the variance of their difference is not necessarily well

approximated by $(s_1^2(\alpha)+s_2^2(\alpha))/n$, the seemingly natural comparison statistic is

$$t(\alpha) = \frac{n^{\frac{1}{2}}(\bar{Y}_1(\alpha) - \bar{Y}_2(\alpha))}{(s_1^2(\alpha)+s_2^2(\alpha))^{\frac{1}{2}}} \tag{4.1}$$

Finding the large sample distribution of $t(\alpha)$ is beyond the scope of the present study. It should be noted, however, that Stigler's [5] Theorem 1 can be used to find the variance of the difference for large samples and that asymptotic normality follows from his Theorem 2.

It is reasonable to prefer values of α which will make $t(\alpha)$ as far from zero as possible. Note that in practice, α is chosen before the observations are taken. To this end, we let

$$g(a) = \lim_{N\to\infty} N^{-1} t^2(\alpha)$$

and note that as a consequence of the theorem of the previous section,

$$g(a) = \frac{2\alpha\rho^2 M^{-2}(a)}{1+\rho^2(aM^{-1}(a)-M^{-2}(a))}$$

$$= \frac{2\rho^2(1-\Phi(a))}{M^2(a) + \rho^2(aM(a)-1))} . \tag{4.2}$$

Differentiating (4.2) with respect to α and setting the result equal to zero gives

$$\rho^2 = \frac{M^2(a)\ (1-2aM(a))}{(1+a^2)\ M^2(a) - 1} \tag{4.3}$$

Although an explicit formula for α in terms of ρ would be more convenient, (4.2) is adequate for present purposes. Note that $g(0)=\rho^2/(\pi/2-\rho^2)$, $g'(0)>0$ for $\rho \geq 0$, and $g(a)$ approaches zero as a gets large. By selecting values of a and evaluating (4.3), we obtain the values of ρ^2 for which those particular choices of a are optimal. For $a \leq 0.61$ and $a \geq 1.10$, (4.3) yields values outside the interval $[0,1]$. Some selected values are presented in Table 1.

TABLE I

Optimal Values of a for Selected Squared Correlations

ρ^2	a	α(100%)
0.063	0.62	26.7
0.135	0.63	26.4
0.259	0.65	25.8
0.519	0.71	23.9
0.789	0.83	20.3
0.904	0.93	17.6
0.955	1.00	15.9
0.999	1.09	13.8

5. CONCLUSIONS

For very small values of ρ^2, a 27% rule appears to be reasonable. However, for the moderately small values often encountered in practice 25% is a more convenient choice. Of course, these figures are based upon normality assumptions and limiting expressions for the sample statistics. The results given above are useful then, only in the sense that they give a rough indication of what one might expect to occur in practice.

The results do indicate that the current practice of using the upper and lower thirds of the sample (a 33 1/3% rule) could probably be improved upon by using the upper and lower fourths (a 25% rule.)

Generalizations to more complex situations, such as those involving several factors, are conceptually similar but also involve other considerations. As noted by Pearson [3] in a biological context, "If it be advantageous for a species to have a certain group of its organs of definite size, falling within a definite range and related to each other in a definite manner, then these changes cannot take place without modifying not only the size but the variability and correlation of all other organs correlated with these, although these organs themselves be not directly selected."

BIBLIOGRAPHY

[1] Kelley, T. L. (1939). The Selection of Upper and Lower Groups for the Validation of Test Items. Journal of Educational Psychology 30, 17-24.

[2] Mosteller, F. (1946). On Some Useful "Inefficient" Statistics. Ann. Math. Statist. 17, 377-408.

[3] Pearson, K. (1903). On the Influence of Natural Selection on the Variability and Correlation of Organs. Philosophical Transactions of the Royal Society of London A 200, 1-66.

[4] Ross, J. and Weitzman, R. A. (1964). The Twenty-Seven Per Cent Rule. Ann. Math. Statist. 35, 214-221.

[5] Stigler, S. M. (1974) Linear Functions of Order Statistics With Smooth Weight Functions. Ann. Statist. 2, 676-93.

Key Words: bivariate normal, optimal design, correlation.