A PROCEDURE FOR DECIDING WHICH TREATMENTS TO ACCEPT IN A SCREENING STUDY INVOLVING PSEUDO-BERNOUlli DATA

Ralph A. DeMasi, University of North Carolina
Dana Quade, University of North Carolina

I. Pseudo-Bernoulli Data

Pseudo-Bernoulli data is simply data which can be represented on both dichotomous and continuous scales. For example, in an experiment one may observe one of two outcomes—a success or a failure. In addition, if one observes a success, an observation indicating the quality of success may also be made. This second response measure is on a continuous scale and may or may not be normally distributed. This type of data contains two kinds of meaningful information: the proportion of successes and the "location" of the successes, where the location may be the mean or median. Data of this sort are common in bone grafting studies. These studies are often conducted to determine which of several grafting substances improves healing of the bone after a fracture. The "healing" is assessed in two ways: by noting whether the bone healed at all (a yes/no observation) and evaluating the strength of those bones which healed (a continuous response). This situation also occurs in other fields, such as business and industry. For example, in the manufacture of certain electrical components, one may observe whether the component works at all, and if it works, one may wish to quantify how "well" it works. In business applications one may be interested in whether a particular sale occurs on a given day, and if the sale does occur, the dollar amount or amount of goods sold would also be of interest. For the purposes of this paper we will assume that higher values indicate a better response.

To illustrate the situation, a relative frequency distribution for a hypothetical case is shown in Figure 1. This represents a treatment for which the successes have responses centered around 100 units, but 13% of the trials are failures. No single parameter summarizes such a situation: neither the 100, nor the 13, nor any compromise such as the mean of 87 obtained by counting the failures as zeros. One needs to consider explicitly the success rate and the location of the successes.

There are several methods of analysis for studies involving pseudo-Bernoulli data. One approach is to evaluate the proportion of successes using a binomial test, ignoring the actual values for the successes. Since this technique does not distinguish between a response that is 1 unit and one that is 100 units, it is not a very good test (Conover 1980). A second approach is to omit the failures and perform a location test on the successes, employing a parametric or nonparametric method, depending on the shape of the data. This technique has a weakness too because it fails to consider the failures that have occurred. A third approach is to count the failures as zeros and perform a location test. Still another approach is to fit a model to the data; perhaps an exponential model would be appropriate in some situations.

There are several problems with these approaches. First, it is impractical to spend time and money collecting data and then throw half of it away. Also, the assumptions of both the parametric and nonparametric techniques may not be valid. For example, a nonparametric rank test may not perform well in the presence of numerous ties (Lehmann 1975). Perhaps the biggest disadvantage of existing techniques is that it is difficult to decide on the overall "goodness" of the treatment. For example, is a treatment having a high success rate but a low mean more favorable than a treatment having a low success rate but a high mean? This is difficult to assess since there is no technique which simultaneously evaluates the proportion of successes and the location of the successes.

These ideas are nicely illustrated with figures. Figure 2 represents two hypothetical situations. Treatment 1 has a very high success rate, but its successes have a low mean, while treatment 2 has a low success rate, but a high mean. If we use a binomial test on the number of failures, treatment 1 will be preferred; but a location test on the successes will favor treatment 2. Analyses based on counting the failures as zeros may be unable to distinguish between the treatments since their overall means are about the same. In addition, such analyses may be inapplicable because of the obvious nonnormality and the high proportion of tied observations.

II. Theoretical Development

Thus far the emphasis has been on data analysis. However, the technique developed here is to be used for experimental design. In particular, the method developed can be used to assist the investigator in planning a study in which the proportion of successes as well as the mean value of the successes will be analyzed. The technique will be developed for the special case of a screening study in which several treatments are compared to a standard. It will also be assumed that the continuous response measure is distributed with known variance. Making these assumptions will simplify the development of the technique; however, as will be shown later in the paper, extensions can be made to accommodate screens in which these assumptions do not hold.

The goal of the screen is to select the "best" treatment or "best" subset of treatments, where best is defined by some reasonable criteria. This problem, generally known as ranking and selection, has received considerable attention in the fields of Operations Research and Industrial Statistics (Nechhofer et al 1968, Gibbons et al 1977, Gupta and Sobel 1960).

In order to develop the procedure some basic notation will be introduced. We will let $P$ denote the proportion of successes and $N$ denote
the mean of the successes; the subscripts i and 0 denote the i'th sample and the standard, respectively. Using this notation, there are four possible sets of outcomes for any one treatment:

1. the treatment has both the proportion and the mean better than the standard, i.e., \( P_i > P_0 \) and \( M_i > M_0 \).

2. the treatment has only the proportion better than the standard, i.e., \( P_i > P_0 \) and \( M_i < M_0 \).

3. the treatment has only the mean better than the standard, i.e., \( P_i < P_0 \) and \( M_i > M_0 \).

4. the treatment has neither the proportion nor the mean better than the standard, i.e., \( P_i < P_0 \) and \( M_i < M_0 \).

The technical development can be simplified if one further assumption is made: that the standard treatment is not in the experiment. The standard values can be supplied by historical data or they can be hypothesized ones. Another way to simplify the presentation is to conceive of the screen as a two-stage procedure. In stage one, each sample proportion is compared to the standard proportion. If the proportion is "good enough", it remains for stage 2; otherwise it is discarded. In stage 2 each sample remaining from stage 1 has its mean compared to the standard, if the sample mean is good enough, then the treatment is selected (i.e., it passes the screen).

What is the criterion to help us decide to pass each treatment? Since this is a statistical technique, it seems appropriate to base the criterion for acceptance on the probability of accepting each treatment given its particular proportion and mean. To determine if this is a good criterion, we note that the probability of the treatment passing the screen should be a strictly non-decreasing function of \( P \) and \( M \). That is, the higher the proportion and mean, the higher the chance of passing the screen. The goal of the procedure is to maximize the probability of selecting the treatments which are truly better than the standard and to minimize the probability of accepting the treatments which are truly not better than the standard.

The simplest method to evaluate the "goodness" of each treatment is to form a descriptive criterion for passing the screen. Suppose we select at stage 1 if and only if the sample proportion for each treatment is better than the standard proportion. Then if the true proportion equals the standard proportion, the probability of selection at stage 1 is 1/2. Also suppose that the treatment is selected at stage 2 if and only if the sample mean is better than the standard mean. Then if the true treatment mean equals the standard mean, the probability of selection at stage two is again 1/2. Now, using the conditional law of probability, \( P(A) = P(B) = P(A)P(B) = Pr(A) + Pr(B) \), the probability of the treatment being selected in the probability of passing both stages, which is \((1/2)(1/2) = 1/4\).

Using an analytical criterion, we can do a better job of maximizing the probability of acceptance for the i'th treatment. In stage 1, we let \( b \) equal the probability of success for the i'th treatment in any one trial. We then specify a critical value, \( c \), which is the minimum number of successes needed for the i'th treatment to pass the screen. Then, \( \sum \binom{n}{x} (b)^x (1-b)^{n-x} \), equals the standard mean, the probability of the treatment the conditional law of probability, \( Pr(A \cap B) \), the probability of selection at stage 1 is 1/2. Also proportion equals the standard proportion, the probability of the treatment the conditional law of probability, \( Pr(A \cap B) \), the probability of selecting the treatments which are truly not better than the standard.

The simplest method to evaluate the "goodness" of each treatment is to form a descriptive criterion for passing the screen. Suppose we select at stage 1 if and only if the sample proportion for each treatment is better than the standard proportion. Then if the true proportion equals the standard proportion, the probability of selection at stage 1 is 1/2. Also suppose that the treatment is selected at stage 2 if and only if the sample mean is better than the standard mean. Then if the true treatment mean equals the standard mean, the probability of selection at stage two is again 1/2. Now, using the conditional law of probability, \( Pr(A) + Pr(B) \), the probability of the treatment being selected in the probability of passing both stages, which is \((1/2)(1/2) = 1/4\).

Using an analytical criterion, we can do a better job of maximizing the probability of acceptance for the i'th treatment. In stage 1, we let \( b \) equal the probability of success for the i'th treatment in any one trial. We then specify a critical value, \( c \), which is the minimum number of successes needed for the i'th treatment to pass the screen. Then, \( \sum \binom{n}{x} (b)^x (1-b)^{n-x} \), equals the standard mean, the probability of the treatment the conditional law of probability, \( Pr(A \cap B) \), the probability of selection at stage 1 is 1/2. Also proportion equals the standard proportion, the probability of the treatment the conditional law of probability, \( Pr(A \cap B) \), the probability of selecting the treatments which are truly not better than the standard. Then, if the true treatment mean is greater than the standard, the conditional probability that the i'th treatment passes stage 2, given \( k \) successes in stage 1, is the probability of rejecting the null hypothesis given that the alternative hypothesis is true. This is the power of the second stage test, denoted by \( P_2(s,m) \). Using the conditional law of probability, the probability of the i'th treatment passing the screen is then

\[ P(b,m) = \sum \binom{n}{x} (b)^x (1-b)^{n-x} \]

In stage 2, our analytical criterion appears in the form of a hypothesis test: Here we test the null hypothesis that the true mean for the i'th treatment is less than or equal to the standard mean. Then, if the true treatment mean is greater than the standard, the conditional probability that the i'th treatment passes stage 2, given \( k \) successes in stage 1, is the probability of rejecting the null hypothesis given that the alternative hypothesis is true. This is the power of the second stage test, denoted by \( P_2(s,m) \). Using the conditional law of probability, the probability of the i'th treatment passing the screen is then

\[ P(b,m) = \sum \binom{n}{x} (b)^x (1-b)^{n-x} \]

In particular, for a normal \( z \)-test for testing a sample mean against the standard, the hypotheses are \( H_0: \mu = \mu_0 \) vs. \( H_1: \mu > \mu_0 \), and the \( z \)-statistic is

\[ z = \frac{x - \mu_0}{\sigma} \]

The power is then \( Pr(\text{Reject } H_0 | H_1 \text{ true}) = Pr( Z > z ) \), or

\[ \text{Power} = Pr( Z > \frac{Q' - \mu_0}{\sigma} ) \]

where \( Q' \) is the \( 1-\alpha \) upper tail critical value for forming the rejection region for \( H_0 \):

\[ Q' = \sigma \cdot z_{\alpha} \]

Appendix II illustrates the calculation of \( P(b,m) \) for a hypothetical case.

One way to represent the probability of acceptance for varying \( b \) and \( m \) is to plot 2-dimensional operating characteristic curves. These curves are analogous to the usual operating characteristic curve (or power function) for a hypothesis test — for example, the normal \( z \)-test for a mean. If we fix the sample size, the first stage critical value, the level of significance for the second stage test, and the null hypothesis for the second stage test, we can calculate the probability of acceptance for
varying \( b \) and \( m \). These probabilities can then be represented using contour plots, where contours of probabilities are plotted with \( b \) on the vertical axis and \( m \) on the horizontal axis. The choice of which parameter appears on which axis is completely arbitrary. With ordinary one-dimensional power functions we can vary one of alpha, \( n \), and the alternative mean while holding the others fixed and examine the effects on the power of the test. Similarly, with these two-dimensional contour plots we can vary the first stage critical value \( c \) as well as the other factors (holding the others fixed) and examine the effect of each factor on the probability of passing the screen.

Before examining any actual plots, it will be illustrative to view the power function for an "ideal" procedure. Figure 3 represents the ideal power function for a test of \( m = m_0 \) versus upper one-sided alternatives (Mendenhall et al. 1986). Here we can see that when \( m \) is less than or equal to \( m_0 \), the probability of rejecting the null hypothesis is zero. However, for alternative values of \( m \) greater than \( m_0 \), the probability of rejection is 1, even for very small differences between the hypothesized mean and the alternative mean. Clearly this is an ideal situation because we want to reject the null hypothesis whenever the true mean is greater than the one hypothesized in the null hypothesis.

In the bottom half of Figure 3 is sketched the ideal operating characteristic for screening procedures involving pseudo-Bernoulli data. Ideally, we wish to reject null and accept the treatment whenever the proportion of successes is greater than or equal to \( c \) and the alternative mean is greater than the hypothesized one. Thus, any alternative treatments having parameters falling in the shaded region of the figure would pass the screening procedure.

Now that we've seen what an ideal procedure would look like, let us turn to some hypothetical situations. Figure 4 represents a screen for a sample size of 10 and a known standard deviation of 10. The critical value for the first stage screen is 6, the standard mean is 5, and the level of significance is .05. The reference one-lines drawn at \( (5, .6) \) represent the standard. The contours labeled .1 through .9 represent values where the treatment would be accepted with the indicated probability. For example, for this treatment to have a .80 chance of passing the screen it must have at least 8 successes and a mean greater than 13. In designing a study, if one wished to have a reasonable chance of detecting a smaller difference \( s \), he could consider increasing the sample size.

The contour plot in Figure 5 demonstrates the effect of increasing sample size on the probability of the treatment passing the screen. It is identical to the previous plot, except that the sample size has doubled. As one would expect, the contours have shifted closer to the standard. Now, instead of an 80% chance of passing the screen, a treatment with the same proportion of successes (i.e., 14) and a mean of 13 would have approximately a 95% chance of passing. Figure 6 shows the contour plot for the same experimental design but for a sample size of 100. Overlaying the three plots as illustrated in Figure 7 displays the shifting contours nicely. Note that the .95 contour for \( n=100 \) approaches the ideal plot shown in Figure 3; as \( n \) approaches infinity, the contours approach the standard.

These graphical techniques can enhance the design stage of the pilot study by allowing visual calculation of sample sizes required to detect a specified difference with a desired probability. These plots also demonstrate the effect of the other factors on \( P(b,m) \). For example, increasing the significance level causes the contours to shift inward, while lowering it will cause the contours to shift outward. Also, the value of \( c \) will cause the contours to shift downward, while increasing it will cause the contours to shift upward. For example, raising \( c \) to 9 out of 10 may cause the .90 contour to shift off the plot. Changing the standard mean also will vary the position of the contours; raising it will cause the contours to shift right, while lowering it will cause the contour to shift left.

III. Computer Software

One might guess that computing and plotting the probabilities would be a very tedious and time-consuming task. For this reason, considerable effort was made to illustrate how one can use statistical programming techniques to produce operating characteristic contour plots for a variety of experimental situations. The programs were developed utilizing elements of SAS BASICS (R) and SAS GRAPH (R) software version 6.03 running on an IBM PC (SAS Language Guide 1985, SAS/Graph 1988). The program consists of a short DATA step for calculating the probabilities and a PROC step for plotting them. The PROBBNML function calculates the binomial probabilities required for stage 1 and the PROBNORM function calculates the power of the second stage test. The GCONTOUR procedure in SAS graph allows one to specify the levels of the contours appearing on the plot. In addition, the ANNOTATE facility allows the user to enhance the appearance of the output to his or her specifications. The program requires the user to specify the level of significance, the sample size, the critical value for the first stage, the standard mean and the known standard deviation, and an upper limit for the alternative mean. The program produces one plot for each experiment if the sample sizes for the treatments are equal. In the case of unequal sample sizes, the program produces one plot for each size \( n \). All contour plots in this paper were produced using SAS on an IBM Personal Computer and printed on a Hewlett Packard laser printer.

An upper limit is necessary for the second stage power calculations because the program must know when to stop the power calculations. That is, it must know how far the mean hypothesized in \( H_0 \) is from the one under \( H_1 \). Do loops with increments of .1 for \( b \) and 1 for \( m \) were used for all plots in this paper. However, the user can easily specify smaller increments for \( b \) if he wishes a smoother contour. The difference in the increments does not affect the shape of the plot appreciably. To illustrate this, the program

1370
which produced the contour plot appearing in Figure 4, except for the increment in one DO loop. The plot in Figure 4 was produced using an increment of .1 for b and 1 for m, while the one in Figure 6 was produced using increments of 1 for b and 1/2 for m. Changing the increments for b also had no substantial effect on the smoothness of the contours. Thus, computer resources can be minimized with little sacrifice in the precision of the contours. This may be very important when operating in the mainframe environment, where CPU costs are high, and in the PC environment, where the real time may be long.

IV. Application

Let us now consider an application which demonstrates how an investigator might use this technique in designing a screening study. The application is to a pilot study designed to assess the potential efficacy of five grafting substances believed to improve fracture healing (Hopp et al. 1989, Bolander and Ballan 1986). The experimental protocol involved surgically fracturing the femur in five groups of 10 rats. A different grafting substance for each group is then applied to each rat femur. After six weeks, the animals are sacrificed, the femurs are harvested, and the fracture healing is assessed by two types of observations. First, the investigator notes whether or not the fracture has healed at all (this is called a union and is based on gross examination of the fracture site). If the fracture has healed, the strength of healing is measured by the force to re fracture in newtons per kg body weight. In planning the study, the investigator wishes to know which alternative treatments will have an 80% chance of being accepted if the standard treatment has a 70% success rate, a mean strength of 50 N/kg, a standard deviation of 13 and the level of significance is set at .05.

The contour plot shown in Figure 9 indicates that treatments having 8 or more successes and a mean of 60 or higher will pass the screen with 80 probability. This plot was produced by the SAS program in Appendix III, attachment (a). These treatments are identified as having potential efficacy and will be studied in a future, larger study. If the investigator desires a higher probability of acceptance, s/he may increase the sample size, increase the level of significance, or allow for a more lenient standard criterion.

V. Extensions and Modifications

We have seen an application for the procedure handling the simplest special case. However, extensions can be made to accommodate experiments in which the standard treatment is also in the experiment, and not supplied by historical or hypothesized values. For cases where the standard deviation is unknown, the second stage test would involve using the t-test in place of the normal z-test. This is very easy to do by replacing the 1-α percentile of z with the 1-α percentile of t\(_{n-1}\) to form the rejection region in the second stage hypothesis test. The second stage test can also be extended to the case where the continuous response measure is not normally distributed. Thus, the t-test or z-test can be replaced by a nonparametric analog such as the Wilcoxon rank sum test.

VI. Summary and Conclusions

This paper has presented a statistical design technique for use in studies where one wishes to make a continuous response measure, but when a certain proportion of the observations are failures. This situation occurs frequently in public health, as well as in other fields. Since both the proportion of successes and the location of the successes yield important, yet different, types of information, the investigator should take care in designing such studies to have a reasonable power of detecting a pre-specified difference of clinical significance. If several experimental treatments are compared to a standard treatment, which can be one of the treatments in the experiment, the probability of each treatment passing the screen can be computed and plotted for varying proportions and means. The SAS programming language is used to compute and plot the probabilities of each treatment passing the screen. These graphical techniques can enhance the design phase of the screening study by allowing visual calculation of the effects of factors such as sample size and significance level on the probability of each treatment passing the screen.

REFERENCES


Appendix I: Figures

Figure 1
RELATIVE FREQUENCY DISTRIBUTION
PERCENT

Figure 3a
POWER FUNCTION
FOR AN IDEAL PROCEDURE

Figure 4
CONTOUR PLOT SHOWING PROBABILITY OF SELECTION

Figure 2
PSEUDO-BERNOULLI DATA
TWO HYPOTHETICAL SITUATIONS
1. TREATMENT YIELDING HIGH SUCCESS RATES, REGARDLESS OF QUALITY OF RESPONSE.

Figure 3b
OPERATING CHARACTERISTIC
FOR AN IDEAL PROCEDURE

Figure 5
CONTOUR PLOT SHOWING PROBABILITY OF SELECTION

n=10, c=6, Mo=5
α = 0.05

n=20, c=12, Mo=5
α = 0.05
Figure 6
CONTOUR PLOT SHOWING PROBABILITY OF SELECTION

Figure 7
EFFECT OF SAMPLE SIZE ON \( P(b,m) \)
FOR THE .90 CONTOUR

Figure 8
CONTOUR PLOT SHOWING PROBABILITY OF SELECTION
M INCREMENTED BY \( \sqrt{2} \)

Figure 9
OPERATING CHARACTERISTIC CONTOUR PLOT
Appendix II: Calculation of \( P(b, m) \) For the One Sample Case

\[
P(b, m) = \sum \left( C^n_x \cdot b^x \cdot (1-b)^{n-x} \cdot \frac{1}{\sqrt{n}} \right)
\]

\[
P(b, m) = \sum \left[ C^n_x \cdot b^x \cdot (1-b)^{n-x} \cdot \frac{1}{\sqrt{n}} \right]
\]

Table for Calculation of \( P(b, m) \)

<table>
<thead>
<tr>
<th>X</th>
<th>P1</th>
<th>SEM</th>
<th>C1</th>
<th>P2</th>
<th>SEM*P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>.084</td>
<td>4.08</td>
<td>11.7</td>
<td>.43</td>
<td>.030</td>
</tr>
<tr>
<td>7</td>
<td>.201</td>
<td>3.78</td>
<td>11.2</td>
<td>.37</td>
<td>.074</td>
</tr>
<tr>
<td>8</td>
<td>.302</td>
<td>3.54</td>
<td>10.8</td>
<td>.23</td>
<td>.124</td>
</tr>
<tr>
<td>9</td>
<td>.369</td>
<td>3.33</td>
<td>10.5</td>
<td>.14</td>
<td>.118</td>
</tr>
<tr>
<td>10</td>
<td>.407</td>
<td>3.16</td>
<td>10.7</td>
<td>.08</td>
<td>.051</td>
</tr>
</tbody>
</table>

\( p = .05 \), \( \alpha = .05 \), \( m = 30 \), \( \sigma = 80 \), \( b = 5 \), \( m = 20 \)

SUPPLIED VARIABLES:
- C = CRITICAL VALUE FOR 1ST STAGE SCREEN
- MO = STANDARD ON HYPOTHESIZED MEAN
- M = OBSERVED NUMBER OF SUCCESSES IN SAMPLE
- N = PROBABILITY OF SUCCESS ON ANY TRIAL
- S = SAMPLE SIZE
- CLEVEL = CONFIDENCE LEVEL FOR S-TEST
- MAUPPER = UPPER LIMIT FOR SECOND STAGE POWER
- PWOT = PROPORTION OF SUCCESSES IN FIRST STAGE

DERIVED VARIABLES:
- SEM = S.E.M. BASED ON N-M SUCCESSES
- Z = 2-STATISTIC UNDER Ho:mu=ma vs Ha:mu=ma
- CRITZ = CRITICAL VALUE FOR Z-STATISTIC UNDER Ha
- CRITXBAR = CRITICAL VALUE FOR SAMPLE MEAN
- P1 = P1(b, m) = PROBABILITY OF M SUCCESSES AT THE 1ST STAGE
- P2 = P2(x, m) = PROBABILITY OF PASSING SECOND STAGE SCREEN
- PRM = PROBABILITY OF PASSING BOTH STAGES OF THE SCREENING STUDY
- SUMPBM = PROBABILITY OF PASSING BOTH STAGES OF THE SCREENING STUDY

/*
DATA NEW;
***SET CONSTANT VALUES***;
N=10;
S=1;
C=7;
MO=50;
MAUPPER=8;
CLEVEL=9.95;
CRITZ=PROBIT(CLEVEL) ;
PNOT=C/N;
DO MA=0 TO MAUPPER BY 1;
DO B=0 TO 1 BY .1;
SIMPBM:
DO N=0 TO N BY 3;
P1=PROBNOML(B,N,M)-PROBNOML(B,N,M-1);
SEM=5/(N**.5);
CRITXBAR=CRITZ*SEM;Z=(CRITXBAR-MA)/SEM;
P2=1-PROBNORM(Z) ;
SIMPBM=SIMPBM+SUMBM(UTILBM,PRM);
IF N=0 THEN OUTPUT;
END;
END;

GOPTIONS DEVICE=PS2EGA COLORS=(YELLOW RED BLUE CYAN) ;
LIBNAME GRAPHS 'C: \USERS\RAD\MYSAS\JOBS\BIOS\MPH\GRAPHS' ;
TITLE1 "CONTOUR PLOT SHOWING PROBABILITY OF SELECTION"
FOOTNOTE1 "alpha=. 05 ,N=10,Mo=50";
PROC GCONTOUR GOUT=GRAPHS.CONTOUR;
PLOT B*MA=SUMPBM/LEVELS= .1, .2, .3, .4, .5, .6, .7, .8, .9, .95, .99
DEC=.50, .6-9 hextt=50 vref=.7 NOLEGEND;
RUN;