# TWO SAMPLE PROBLEMS FOR A DICHOTOMOUS VARIABLE WITH MISSING DATA

Janet Dixon Elashoff, Stanford University
amd Robert M. Elashoff, University of California, San Francisco

Introduction. Incomplete or missing data is a major problem in many fields. Data may be incomplete due to subject nonresponse or refusal to cooperate, transcription errors, random loss, or to a variety of other reasons. As a consequence, statistical techniques to deal with incomplete data are necessary. One common approach is simply to delete and ignore the incomplete cases. To select an appropriate technique, however, something must be known about the kind of observations which are missing and the variables influencing the loss of certain observations.

Two sample problems with binary data are considered in this study. These problems are formulated in section 2 along with specific probability models to describe the missing observations. In later sections, we describe and explain the difficulties encountered in parameter estimation for the sampled populations, discuss estimation and tests for differences and ratios of the parameters, and give recommendations for data analysis.

Problem formulation, Models, Notation. Many studies to compare the effectiveness of different treatments have nonrespondents. For example, suppose patients with a certain disease are assigned either an active drug ($x = 2$) or a placebo ($x = 1$) in a double blind study. The placebo has the same side effects as the active drug, but presumably it does not have the same curative or palliative effect as the active drug. A follow-up study is made and each patient is scored as improved ($y = 1$) or unimproved ($y = 0$) on a response variable y. Lack of improvement may cause some patients to drop out of the study or refuse to cooperate further. Improvement also may give patients a reason to drop out or a chance to leave the area. In either case the y measurements are unknown. Clearly, under these circumstances, a missing y may be influenced by whether or not the patient is improved but not directly by the drug the patient received. The goal of the clinical study is to compare the probability of improvement for the active drug ($p_2$) with the probability of improvement for the placebo ($p_1$).

A statistical model for such problems is defined in this way. A random sample of $N_i$, $i = 1, 2$, individuals is given treatment $x = i$, and $n_i$ of these individuals are observed on the binary response variable y. No measurement error exists. Denote by $p_i$ the probability that $y = 1$ when $x = i$. The value of x is known for each of the $(N_1 + N_2)$ individuals. Define

$q(x,y) = P(\text{an individual's y is recorded} | x,y)$.

We can distinguish four particular specifications of $q(x,y)$:

(1) $q(x,y) = q$,
the missing data occur at random;

(2) $q(x,y) = q_y$,
the probability of missing data depends on an individual's y value but not his x value;

(3) $q(x,y) = q_x$,
the probability that an individual's score is recorded depends only on the population he is from; or

(4) $q(x,y) = q_{xy}$,
the general case where the probability of missing data depends on both x and y.

The second specification, case 2, is more appropriate for problems like the drug example described above. We investigate the question of how much it matters whether we base our statistical analysis on case 1 or case 2.

Our goals are to estimate the population quantities $p_1$ and $p_2$, to estimate the comparative population measures

1. $D = p_1 - p_2$,
2. $R = p_1/p_2$,
3. $OR = \dfrac{p_1/(1-p_1)}{p_2/(1-p_2)}$

and to propose significance tests and confidence intervals for these three population measures.

Under cases 1 and 3, standard estimators and inference procedures for $p_1$, $p_2$, D, R, and OR may be used conditional on the observed sample sizes, $n_1$ and $n_2$. Under case 4, there is insufficient information to estimate $p_1$ and $p_2$ or the functions D, R, or OR since only 4 of the 6 parameters can be estimated from the data. Our study investigates the maximum likelihood estimators for the parameters $p_i$, D, R, and OR under case 2, examines asymptotic and small sample results for conditional and unconditional means, variances, and mean squared errors of the estimators, and compares their behavior to that of case 1 estimators.

Estimation of the $p_i$. Under case 1, when $q(x,y) = q$, and the missing observations occur at random, the $(N_i - n_i)$ missing observations can be ignored and the remaining observations regarded as random samples of size $n_i$. Let $r_i$ be the number of individuals for whom $y = 1$ out of the $n_i$ actually observed in population i. Then the ml estimators of $p_1$, $p_2$, and q are

(5) $\hat{p}_{1i} = r_i/n_i$

(6) $\hat{q} = (n_1 + n_2)/(N_1 + N_2)$.

Standard distribution theory applies to these estimators.

When $q(x,y) = q_y$ and case 2 holds, the ml estimators are derived by the following argument. Let

$\alpha_1 = q_1 p_1$, $\alpha_2 = q_1 p_2$;

$\beta_1 = q_0(1-p_1)$, $\beta_2 = q_0(1-p_2)$.

Then, the likelihood of the two samples becomes

(8) $L = C \prod_{i=1}^{2} \alpha_i^{r_i} \beta_i^{(n_i - r_i)} [1-\alpha_i-\beta_i]^{(N_i-n_i)}$

After differentiating L with respect to the $\alpha_i$ and $\beta_i$ and setting the resulting equations to zero, we find

(9) $\hat{\alpha}_i = r_i/N_i$, $\hat{\beta}_i = (n_i-r_i)/N_i$, $i = 1, 2.$

Since the parameter set $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ is a one-to-one transformation of $(p_1, p_2, q_1, q_0)$ when $p_1 \neq p_2$ and neither $q_1$ nor $q_0$ equal to zero, the ml estimators of $p_i$, $q_0$, $q_1$ can be obtained from equations (7) and (9) as

$$\hat{p}_{2i} = (r_i/N_i)[N_2(n_1-r_1)-N_1(n_2-r_2)](n_1r_2-r_1n_2)$$

(10) $\hat{q}_0 = (n_1r_2-n_2r_1)/(N_1r_2-N_2r_1)$

$$\hat{q}_1 = (n_1r_2-n_2r_1)/(N_2(n_1-r_1)-N_1(n_2-r_2)).$$

We note that these estimators of $p_1$, $p_2$, $q_1$, $q_0$ break down when $p_1 = p_2$. Clearly, when $p_1=p_2=p$, we have only a single sample from which it is impossible to obtain even a consistent estimator of p under case 2. Mathematically, when $p_1=p_2=p$, then $\alpha_1 = \alpha_2 = \alpha$ and $\beta_1 = \beta_2 = \beta$ and although ml estimators of $\alpha$ and $\beta$, or $q_1p$ and $q_0(1-p)$, exist the parameter set $(\alpha,\beta)$ is not a one-to-one transformation of $(p, q_0, q_1)$ and ml estimators for $p, q_1, q_0$ do not exist.

In practice, additional difficulties arise with the use of this ml estimation method even when $p_1 \neq p_2$. Samples in which $(\hat{\alpha}_1/\hat{\alpha}_2) < 1$ and $(\hat{\beta}_1/\hat{\beta}_2) < 1$ or in which both ratios are greater than one lead to some of the $(\hat{p}_1, \hat{p}_2, \hat{q}_0, \hat{q}_1)$ being negative. This situation not infrequently occurs when $p_1$ is close to $p_2$ or the $n_i$ are small. Such a difficulty implies that the ml estimators are not very precisely determined. If we constrained the $\hat{\alpha}_i$, $\hat{\beta}_i$ so that the preceding inequalities would not occur, we would essentially be setting $p_1 = p_2$ in which case p cannot be estimated.

Under case 2 where $q(x,y) = q_y$ both $\hat{p}_{1i}$ and $\hat{p}_{2i}$ have asymptotic normal distributions for $p_1 \neq p_2$. The mean of $\hat{p}_{1i}$ is $\theta_i$ which equals $p_iq_1/(p_iq_1 + (1-p_i)q_0)$ and the asymptotic mean of $\hat{p}_{2i}$ is $p_i$. The asymptotic conditional and unconditional variances of $\hat{p}_{1i}$ and $\hat{p}_{2i}$ are given in Elashoff and Elashoff (1971). The asymptotic variance formulas for $\hat{p}_{2i}$ contain the term $(p_1-p_2)^2$ in the denominator indicating that for $|p_1-p_2|$ small the variance of $\hat{p}_{2i}$ will be large.

For tests of the null hypothesis $H_0 : p_1 = p_2$ under cases 1 or 2 $\theta_1 = \theta_2$ if and only if $p_1 = p_2$ (for neither $q_0 = 0$ or $q_1 = 0$), and thus a test of $p_1 = p_2$ may be carried out by standard methods such as the Fisher-Irwin test conditional on $n_1$, $n_2$, and $r_1 + r_2$.

Estimation of D, R, and OR. The population measures D, R, and OR, or log OR are frequently used quantities for comparing $p_1$ and $p_2$. In this section we discuss their estimation when $q(x,y)=q_y$. Although inferences about D, R, or OR will usually be based on conditional variance formulas, a detailed numerical study of asymptotic formulas and small sample behavior conditional on possible $n_1$ and $n_2$ pairs is unwieldy and consequently discussions will focus on unconditional results. To evaluate the usefulness of asymptotic formulas for comparisons and to examine the behavior of the estimators in small samples, exact unconditional means and variances were calculated. For example, the exact mean of $\hat{D}_1$ is

$$(11) \sum_{n_2=1}^{N_2} \sum_{n_1=1}^{N_1} \sum_{r_2} \sum_{r_1} \hat{D}_1 \frac{p(r_1)p(r_2)p(n_1)p(n_2)}{1 - P\{n_1=0 \text{ or } n_2=0\}}$$

where

$$p(r_1) = \binom{n_i}{r_i} \theta_i^{r_i}(1-\theta_i)^{(n_i-r_i)} \text{ and } p(n_i) =$$

$$\binom{N_i}{n_i} [p_iq_1+(1-p_i)q_0]^{n_i}[1-p_iq_1-(1-p_i)q_0]^{(N_i-n_i)}$$

Note that results were obtained conditional on $n_1 \neq 0$, and $n_2 \neq 0$ and that for $n_1r_2 = n_2r_1$ we defined $\hat{D}_2 = 0$. Calculations were made for $N_1=N_2=$ 20, 50, for $p_1, p_2 = .10, .25, .50, .75, .90$ and $q_1,q_0 = .50, .75, .90, 1.0$. Summary results may be found in Elashoff and Elashoff (1971).

Estimators for D are obtained by substitution of $\hat{p}_{1i}$ or $\hat{p}_{2i}$ in $D = (p_1-p_2)$ and are given by

$$(12) \hat{D}_1 = \left(\frac{r_1}{n_1} - \frac{r_2}{n_2}\right)$$

for case 1, and

$$(13) \hat{D}_2 = \left(\frac{r_1}{N_1} - \frac{r_2}{N_2}\right) \frac{N_2(n_1 - r_1) - N_1(n_2 - r_2)}{(n_1r_2 - n_2r_1)}$$

for case 2, when $p_1 \neq p_2$.

Both $\hat{D}_1$ and $\hat{D}_2$ have asymptotic normal distributions under case 2. Asymptotic conditional and unconditional means and variances are given in Elashoff and Elashoff (1971).

The estimator $\hat{D}_2$ does not exist when $p_1 = p_2$; the presence of the term $(\theta_2-\theta_1)^4$ in the denominator of the conditional variance of $\hat{D}_2$ demonstrates that $\hat{D}_2$ will have a large variance when $p_1$ is near $p_2$. $\hat{D}_2$ is a consistent estimator of D while $\hat{D}_1$ is not consistent unless $q_1 = q_0$ or $p_1 = p_2$; the bias in $\hat{D}_1$ is $(\theta_1-\theta_2) - (p_1-p_2)$ independent of N.

Examination of unconditional asymptotic and small sample results indicate that neither $\hat{D}_1$ nor $\hat{D}_2$ provides a good estimate of D in general. Unless $q_1 = q_0$ or $p_1 = p_2$, $\hat{D}_1$ may have considerable bias, and unless $N|p_1-p_2|$ is large $\hat{D}_2$ has a relatively large variance.

In small samples, $\hat{D}_2$ is biased; both the bias and the variance of $\hat{D}_2$ decrease as N increases, so for sufficiently large N, $\text{mse}(\hat{D}_2) < \text{mse}(\hat{D}_1)$ unless $p_1 = p_2$ or $q_1 = q_0$. We note however that for $N_1 = N_2 = 50$, $\text{mse}(\hat{D}_2) < \text{mse}(\hat{D}_1)$ only for $|p_1-p_2| > .4$ and $|q_1-q_0|$ large.

Let us now consider the estimation of R. Maximum likelihood estimators for $R = p_1/p_2$ are

$$(14)\quad \hat{R}_1 = \frac{r_1 n_2}{r_2 n_1}$$

and

$$(15)\quad \hat{R}_2 = \frac{r_1 N_2}{r_2 N_1}$$

for cases 1 and 2, respectively. Under case 2, $R_1$ and $R_2$ have asymptotic normal distributions with means and conditional and unconditional variances as given in Elashoff and Elashoff (1971).

The estimator $\hat{R}_1$ is consistent if and only if $q_1 = q_0$ or $p_1 = p_2$, otherwise the bias is

$$(16)\quad \text{bias}\,(\hat{R}_1) = (q_1-q_0)(p_2-p_1)\frac{\theta_1}{p_2 p_1}\ ;$$

$\hat{R}_2$ is consistent. Note that the ratio of the asymptotic conditional variances, var $\hat{R}_1$/var $\hat{R}_2$, equals

$$\left(\frac{\tau_2}{\tau_1}\right) = \left(\frac{n_2 N_1}{N_2 n_1}\right)^2$$

which should approach

$$\left(\frac{p_2 q_1 + (1-p_2)q_0}{p_1 q_1 + (1-p_1)q_0}\right)^2$$

for large N. Thus var $\hat{R}_1$/var $\hat{R}_2 < 1$ for $(q_1 - q_0)(p_2 - p_1) < 0$ or $(q_1 - q_0)(1 - R) < 0$. The ratio of asymptotic unconditional variances varies with $(q_1 - q_0)(p_2 - p_1)$ in a similar way but is generally smaller than the ratio of conditional variances.

Asymptotic unconditional formulas for the mean squared errors of $\hat{R}_1$ and $\hat{R}_2$ were compared for N = 200 for the parameter sets defined earlier. Except for cases where $q_1 = q_0$ or $p_1 = p_2$, when $\hat{R}_1$ is unbiased, $\text{mse}(\hat{R}_1)/\text{mse}(\hat{R}_2)$ was generally greater than .84 and frequently greater than 1.0, which suggests that use of $\hat{R}_2$ will prove generally satisfactory in large samples.

Both $\hat{R}_1$ and $\hat{R}_2$ are biased in small samples. The bias in $\hat{R}_2$ is independent of $p_1$ and decreases slowly with increasing N and increasing $q_1$ (it is almost unaffected by $q_0$). The range of percentage bias in $\hat{R}_1$ is similar to that of $\hat{R}_2$ when $q_1 = q_0$ but generally larger when $q_1 \neq q_0$. For investigators interested in using $\hat{R}_2$, the estimator can be corrected for bias using standard methods.

Comparisons of exact mean squared errors for $\hat{R}_1$ and $\hat{R}_2$ when N = 20 and N = 50 demonstrate that asymptotic formulas provide good indications of the size of $\text{mse}(\hat{R}_1)/\text{mse}(\hat{R}_2)$ in small samples. The ratios of exact to asymptotic unconditional variances are quite similar for $\hat{R}_1$ and $\hat{R}_2$. The exact variances are generally larger than the asymptotic variance formulas for both $\hat{R}_1$ and $\hat{R}_2$ except for $p_1 = p_2$ and N = 20; for N = 50, the ratios vary from 1.0 to 3.7, being close to 1.0 for $R < 1.0$ and larger for $R > 1.0$.

On the whole then, $\hat{R}_2$ should provide a reasonable estimator of R for N not too small.

The estimators of OR for case 1 and case 2 both reduce to

$$(17)\quad \widehat{OR} = \frac{r_1(n_2 - r_2)}{r_2(n_1 - r_1)}.$$

This estimator is asymptotically unbiased under both cases. The asymptotic conditional and unconditional variances of $\sqrt{N_1 + N_2}\ \widehat{OR}$ under case 2 are given in Elashoff and Elashoff (1971).

The independence of the form of the estimator from q(x,y) suggests that the use of $\widehat{OR}$ will be robust to q(x,y). Although asymptotically unbiased, $\widehat{OR}$ may have a substantial bias for $N_1 = N_2 = 20$. Generally the bias is of the order of 20% to 50% of OR, although it does not contribute appreciably to the mean square error. The behavior of $\widehat{OR}$ in small samples does not seem to depend particularly on $|p_1 - p_2|$ or $|q_1 - q_0|$. For $N_1 = N_2 = 20$, the exact variance may be from 2 to 5 times larger than the asymptotic variance for the parameter sets investigated.

To estimate OR, $\widehat{OR}$ (or a modification to reduce bias) can be used for either case 1 or 2. Uniformly most-accurate confidence intervals can be constructed for OR using the noncentral distribution of $r_1$, $r_2$ conditional on $(r_1 + r_2)$, $N_1$, $N_2$ (see Lehmann, 1959). This noncentral distribution is the same for both cases.

Some authors prefer log OR to OR. Of course the estimator of log OR has the same property of invariance under cases 1, 2, and 3 as does the estimator of OR. Haldane (1955), Anscombe (1956) and Gart and Zwiefel (1967) have recommended the substitution of $\hat{p}_i + \frac{1}{2n_i}$ for $\hat{p}_i$ and $(1-\hat{p}_i) + \frac{1}{2n_i}$ for $(1-\hat{p}_i)$ to reduce the bias of the estimator of the logit. This would result in the estimator

$$(18)\quad \log \widehat{OR} = \log \frac{(r_1 + 1/2)(n_2 - r_2 + 1/2)}{(r_2 + 1/2)(n_1 - r_1 + 1/2)}$$

Uniformly most accurate confidence intervals for log OR could be constructed in the same way as for OR.

Conclusions. We have studied two-sample problems with dichotomous data in which the probability that an individual score will be missing depends on the value of that score.

Estimation of the $p_i$ and $q_1$, $q_0$ and estimation of $D = p_1 - p_2$ break down unless $N|p_1-p_2|$ is large. The case 2 estimator of R does not perform

especially well for $N|p_1-p_2|$ small. This suggests preceding attempts to estimate D or R by a test of $H_0 : p_1 = p_2$, since conditional upon $n_1$, $n_2$, and $r_1 + r_2$ such a test may be carried out by standard methods even when case 2 holds. Alternatively, since the estimator of OR is the same under cases 1 and 2, estimation of OR or log OR rather than of D or R should be considered when more than a small fraction of the data is missing.

Acknowledgements. The authors would like to thank B. L. Welch for his helpful comments.

REFERENCES

Anscombe, F. J. (1956). "On estimating binomial response relations." Biometrika, 43, 461-464.

Elashoff, J. D. and Elashoff, R. M. (1971). "Missing data problems for two samples on a dichotomous variable." Research and Development Memorandum No. 73, Stanford Center for Research and Development in Teaching, Stanford University, Stanford, Calif.

Gart, J. J. and Zweifel, J. R. (1967). "On the bias of various estimators of the logit and its variance with application to quantal bioassay." Biometrika, 54, 181-187.

Haldane, J. B. S. (1955). "The estimation and significance of the logarithm of a ratio of frequencies." Annals of Human Genetics, 20, 309-311.

Lehmann, E. L. (1959). Testing statistical hypotheses. New York: John Wiley and Sons.