LBL-20963

Lawrence Berkeley Laboratory

UNIVERSITY OF CALIFORNIA

Computing Division

RECEIVED
LAWRENCE
BERKELEY LABORATORY
MAR 20 1986
LIBRARY AND
DOCUMENTS SECTION

To be presented at the Conference of the Society for Epidemiologic Research, Pittsburg, PA, June 11-13, 1986

TWO ISSUES CONCERNING THE ANALYSIS OF GROUPED DATA

S. Selvin

February 1986

TWO-WEEK LOAN COPY
This is a Library Circulating Copy which may be borrowed for two weeks.

Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098

LBL-20963

LBL-20963

Two Issues Concerning the Analysis of Grouped Data

Steve Selvin

Computer Science Research Department
University of California
Lawrence Berkeley Laboratory
Berkeley, California 94720

February, 1986

TWO ISSUES CONCERNING THE ANALYSIS OF GROUPED DATA

Steve Selvin[1]

ABSTRACT

Selvin, Steve. (Department of Biomedical and Environmental Health Sciences, University of California, Berkeley) Two issues concerning the analysis of grouped data.

Simple statistical models are used to illustrate two important issues arising in the analysis of grouped data. The consequences are explored of grouping continuous data and analyzing the resulting contingency table. Specifically, an expression for the loss of power is derived when an odds ratio is used to assess risk measured by a continuous variable. Also explored are the consequences of employing correlation and regression coefficients to analyze summary variables derived from grouped data (ecologic data). An expression is given that demonstrates the magnitude of a bias (ecologic fallacy) resulting from analyzing a specific type of grouped data.

keywords: bias; ecologic fallacy; grouped data; odds ratio

[1] This research was supported by the Office of Health and Environmental Research, U.S. Department of Energy under contract DE-AC03-76SF00098.

Grouped data generally arise in two ways -- data aggregated by the investigator into categories and data consisting of summary measures characterizing predefined categories (sometimes called ecologic data). Two statistical models are proposed to illustrate and discuss specific questions in connection with analyzing these types of data. The models provide a forum where the properties of specific analytic strategies are precisely delineated and, then using mathematical/statistical tools, the consequences of employing a particular analytic approach is evaluated. A statistical model forces one to define unambiguously the problem at hand.

Two general questions addressed by the proposed models are;

> What are the consequences of analyzing a set of continuous data with methods designed for contingency tables?

and

> What are the consequences of analyzing a set of summary measures with methods designed for continuous data?

The first question concerns the loss of efficiency resulting from treating continuous variables as categorical data (statistical power); where the second question concerns a bias resulting from uncritical application of specific measures of association (ecologic fallacy).

I. Loss of statistical power from grouping continuous data

For generally unclear reasons, many investigators judge that when an observation is not measured precisely then using continuous measures will not gain much precision over an analysis based on data grouped into a few categories. A statistical model gives some idea of the consequences of this decision.

Assume that a continuous variable labeled X, related to the disease under investigation, has a normal distribution. For example, X could represent the level of an individual's blood pressure. Further, assume a dichotomous risk factor exists such as educational level -- high school education (R = 0) versus college education (R = 1). Let variables X and R define a model population where X is normally distributed with mean μ_0 when R=0 and with mean μ_1 when R = 1. Both normal distributions are assumed to have the same variance, say $\sigma^2 = 1.0$ for convenience. If a sample of k individuals is randomly selected from this population, the expected data would produce a 2 by 2 contingency table with cell frequencies P_{ij} or

	$\bar{D}$ (no disease)	D (disease)
R = 0	$P_{00} = k(1 - p)\Phi$	$P_{01} = k(1 - p)(1 - \Phi)$
R = 1	$P_{10} = kp(1 - \Phi)$	$P_{11} = kp\Phi$

The symbol $\bar{D}$ represents all observations such that $X \leq \frac{1}{2}(\mu_0+\mu_1)$ and D represents $X>\frac{1}{2}(\mu_0+\mu_1)$. The new variable D results from grouping the continuous variable X into non-diseased and diseased categories. A familiar example of such a practice is defining individuals as non-hypertensive (say, $X \leq 140$) and hypertensive ($X>140$) based on their systolic blood pressure. The prevalence of the risk factor R is represented as p (i.e., $p = P(R = 1)$) and

$$\Phi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{1}{2}(\mu_0+\mu_1)} e^{-\frac{1}{2}t^2} dt.$$

This statistical structure is depicted in figure 1.0.

The expected odds ratio calculated from data sampled from these normal populations and classified into a 2 by 2 table is $O = [\Phi/(1 - \Phi)]^2$ or $\ln[O] = 2\ln[\Phi/(1 - \Phi)]$. The variance of $\ln[O]$ is approximately:

$$V^2 = \text{variance}(\ln[O]) = [kp(1 - p)\Phi(1 - \Phi)]^{-1}$$

A typical α-level test of the hypothesis,

$$H_0 : O = 1.0 \text{ versus } H_1 : O > 1.0,$$

is often used to assess the role of the risk factor R. Employing $\ln[O]$, which has an approximately normal distribution, gives,

$$P(\ln[O] > Z_{1-\alpha} V_0 \mid H_0) = \alpha$$

where $Z_{1-\alpha}$ is the $(1 - \alpha)$-th percentile of a standard normal distribution.

Now if the power of this test is set at a level $1 - \beta$ or $P(\text{reject } H_0 \mid H_1) = 1 - \beta$, then the necessary sample size k to achieve a power of $1 - \beta$ is approximately

$$k = \frac{[2Z_{1-\alpha} + Z_{1-\beta}/\sqrt{\Phi(1-\Phi)}]^2}{p(1-p)\,\Phi(1-\Phi)}.$$

The comparison of two mean values based on a sample from a population consisting of two normal distributions ($\sigma^2 = 1$) is a classic problem in statistics. An α-level hypothesis test is

$$H_0 : \mu_0 = \mu_1 \text{ versus } H_1 : \mu_0 < \mu_1$$

and

$$P((\bar{X}_1 - \bar{X}_0) > Z_{1-\alpha} v \mid H_0) = \alpha.$$

When sampling is conducted without knowledge of the risk factor, the approximate variance of the difference between two mean values is

$$v^2 = \text{variance}\,(\bar{X}_1 - \bar{X}_0) \doteq [np(1 - p)]^{-1}$$

where n represents the total sample size. The sample size necessary for α-level test with statistical power of 1-β is then

$$n = \frac{[Z_{1-\alpha} + Z_{1-\beta}]^2}{p\,(1-p\,)(\mu_0 - \mu_1)^2}$$

The efficiency ratio (k/n) contrasting the two approaches is (for the special case of $\alpha = \beta$) given by

$$k/n = \frac{\left[2 + 1/\sqrt{\Phi(1-\Phi)}\,\right]^2 / \,(ln\,[0])^2}{4/(\mu_0 - \mu_1)^2} > \frac{\pi}{2}\ .$$

Furthermore, if the odds ratio is less than 2.0 or $0 < (\mu_1 - \mu_0) < 1.2$ then, approximately, k/n $= \pi/2 = 1.571$. For this range, the efficiency ratio implies that if n = 100 observations are necessary to achieve a specific level of α and β, then k = 157 observations are required when continuous data is dichotomized and analyzed with an odds ratio for the same error rates of α and β.

Using an odds ratio to analyze dichotomized continuous data indeed reduces the probability of detecting the influence from a risk factor when it exists. On occasions, grouping continuous data into a table protects the analysis against affects from outliers. Outliers (out and out outliers) should be eliminated from a data set but should not dictate the analytic approach. Another motivation for preferring an odds ratio approach has to do with presentation and simplicity of measurement which are also not persuasive reasons for chosing a specific analytic strategy. In some cases, the judgement that the data can only be roughly measured is said to jus-

tify the analysis of continuous data using contingency table techniques. This question has not been fully explored [1] and, clearly, the presence of measurement error hurts any analytic approach. So the best that can be said for using a 2 by 2 table to assess a risk factor associated with a continuous variable is that equivocal gains are paid for by a definite loss of statistical power. One last point: if a continuous variable is divided into more than two categories, the loss of power is reduced as the number of categories is increased [2].

II. Bias incurred by applying correlation and regression techniques to grouped data

The fact that an analysis of summary measures derived from grouped data does not always reflect the behavior of the individuals who make up the group was noted by Robinson [3] and subsequently called the ecologic fallacy. In one form the ecologic fallacy appears as a bias in correlation and regression coefficients.

To illustrate this bias as it applies to correlation coefficients the following statistical structure is proposed. Let $X_{1i} = U_i + cV_i$ and $X_{2i} = U_i + cW_i$ where U, V and W are independent, normally distributed random variables with expectations = 0.0 and variances = 1.0. Then, it follow that

$$\text{variance}(X_1) = \text{variance}(X_2) = 1 + c^2 = 1/\rho$$

$$\text{with covariance}(X_1, X_2) = 1.0$$

giving the correlation between X_1 and X_2 as ρ when $c = \sqrt{(1-\rho)/\rho}$. That is,

$$\text{correlation } (X_1, X_2) = \frac{1}{1+c^2} = \rho.$$

To model the behavior of X_1 and X_2 in the context of an ecologic study envision N pairs

(X_{1i}, X_{2i}) distributed into k groups with n pairs per group (N = nk). These k groups are summarized by k pairs of mean values $(\bar{X}_{1j}, \bar{X}_{2j})$ calculated for each group based on n observations.

When the k groups are formed without regard to the values of X_1 or X_2 (say, at random), then the correlation calculated ($\bar{r}$) employing the k pairs directly reflects ρ or, in other words, $\bar{r}$ is approximately equal to ρ for large values of k. Note that $\bar{r}$ is based on a sample of size k reducing its precision compared to a correlation coefficient based on sample of size N (ungrouped data) but, nevertheless, $\bar{r}$ is a consistent estimate of the correlation between individual pairs of observations.

If the grouping of the N pairs is based, at least to some extent, on the values of X_1 or X_2 an entirely different picture emerges. Consider the following special and admittedly extreme case. Say k groups are formed on the basis of X_1. The N pairs (X_{1i}, X_{2i}) are ordered according to values of X_1 and then formed into k groups. Again each group is summarized by $(\bar{X}_{1j}, \bar{X}_{2j})$. For this situation, then

$$\text{variance } (\bar{X}_1) \doteq \text{variance } (X_1) = 1/\rho \ ,$$

$$\text{covariance } (\bar{X}_1 , \bar{X}_2) \doteq \text{covariance } (X_1 , X_2) = 1.0$$

and

$$\text{variance } (\bar{X}_2) \doteq \frac{1}{\rho}[\rho^2+(1-\rho^2)/n\,].$$

Therefore, the correlation based on the grouped data is approximately

$$\text{correlation } (\bar{X}_1 , \bar{X}_2) = \bar{\rho} \doteq \rho[\rho^2 + (1-\rho^2)/n\,]^{-1/2}$$

Figures 2.0 shows the relationship between ρ based on the population of individuals and $\overline{\rho}$ based on a set of groups sampled from that population. The failure of ρ to equal $\overline{\rho}$ illustrates the ecologic fallacy. The figure shows that $\overline{\rho} >> \rho$ for many moderate values of ρ. This type of spurious association has been noted in correlations calculated from grouped data where coefficients exceeding 0.90 are observed [4].

The same sort of bias can influence regression analysis. Consider again the variables X_1 and X_2 entered as independent variables into a typical bivariate regression analysis or

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

where e_i represents a series of independent and normally distributed error terms with the same variance.

If k groups of size n are randomly formed from a set of N values (Y_i, X_{1i}, X_{2i}), then estimates of the regression coefficients are produced which are not influenced by the grouping process. In other words, the regression analysis based on randomly grouped data reflects, with some loss of precision, the underlying linear relationship. Even if, as before, the data are ordered into k groups based on X_1, the regression analysis employing means calculated from each group also produces accurate estimates of the parameters of the linear model. However, if the variable used to form the groups is not included in the analysis, the resulting estimates of the coefficients are biased. If X_1 is again used to order the data, for example, but not included in the regression analysis; then employing k pairs of mean values $(\overline{Y}_j, \overline{X}_{2j})$ gives an estimate of the regression coefficient associated with X_2 as b_2 + bias where the bias is $\overline{\rho}^2/\rho$. As in the case of simple correlation coefficients, this bias can be of a considerable magnitude.

The discussed statistical model illustrates the fact that a correlation coefficient calculated from grouped data is misleading when interpreted as a measure of the correlation that would be observed if the non-aggregated data were available. A similar bias in the regression coefficients is demonstrated when the coefficients are estimated from grouped data and the analytic model fails to include measures that reflect the grouping process. This type of bias, sometimes called the ecologic fallacy, can be considered as a special case of incomplete model bias [5]. Therefore, a fundamental question associated with applying linear models to ecologic data or, for that matter any grouped data, is whether the process underlying the formation of the groups is measured and included in the model. If the answer is yes, the estimated regression equation may be of value. If the answer is no, the estimated regression equation has little value with respect to understanding the relationships among the individuals that make up the analyzed groups.

REFERENCES

1. Cochran, W., Planning and Analysis of Observational Studies. John Wiley & Sons, New York, 1983.

2. Cox, DR., Note on Grouping. Am. Stat. Assoc. J. 1957;19:543-549.

3. Robinson, WS., Ecologic correlations and the behavior of individuals. Am Sociol Rev 1950;15:351-7.

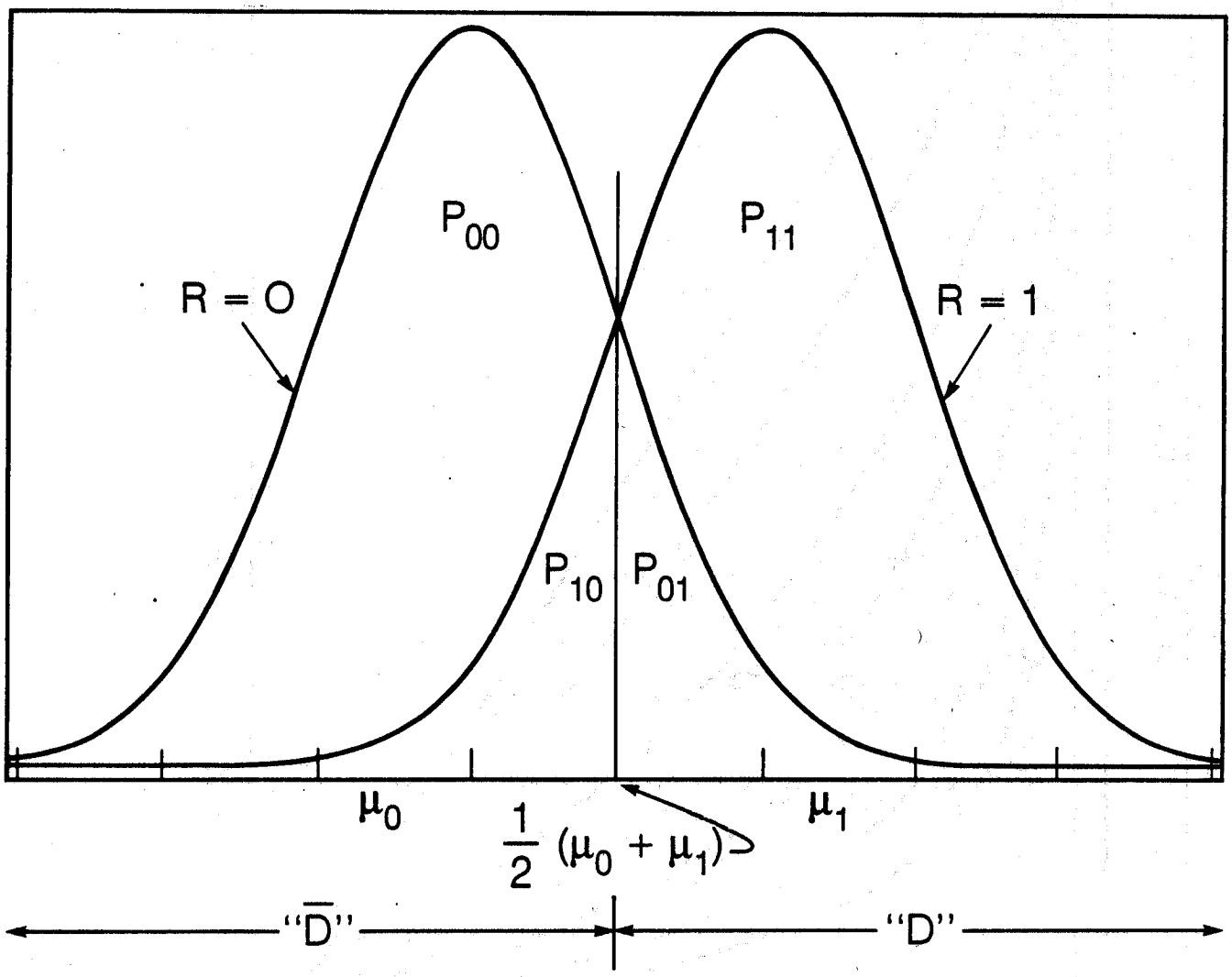
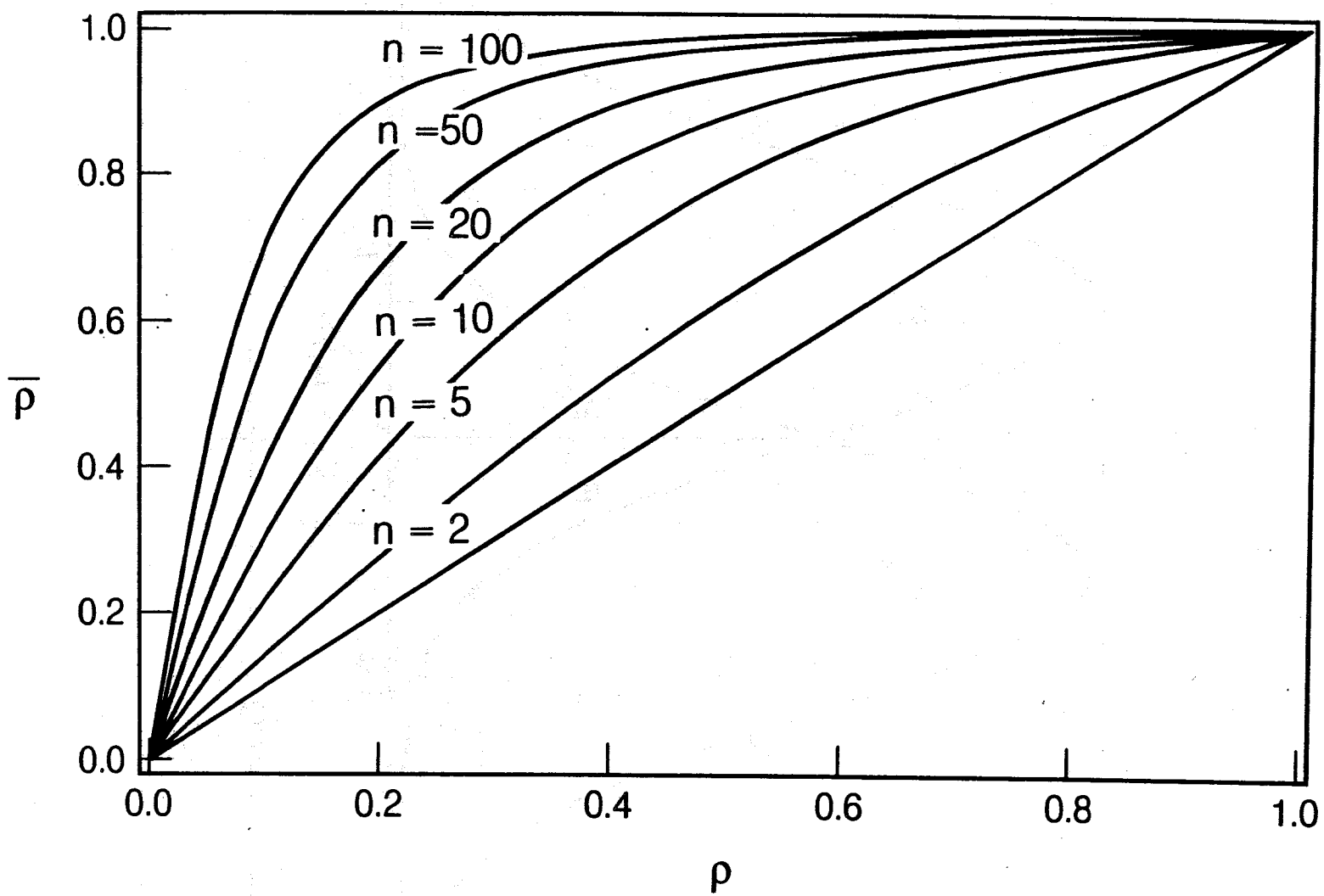
4. Kasl, SV., Mortality an the business cycle: some questions about research strategies when utilizing macro-social models and ecologic data. Am J. Public Health;69:784-788.

5. Draper, NR. and Smith H., Applied Regression Analysis. John Wiley & Sons, New York, 1966.

TITLES TO FIGURES

Figure 1.0 Two normal populations with equal variance and different mean values associated with two levels of a risk factor R.

Figure 2.0 An illustration of the relationship between ρ and $\overline{\rho}$ (n is the sample size).


P_{00}
P_{11}
R = O
R = 1
P_{10}
P_{01}
μ_0
$\frac{1}{2}(\mu_0 + \mu_1)$
μ_1
"$\overline{D}$"
"D"

ρ plotted against $\overline{\rho}$