ON ESTIMATING DISTRIBUTIONAL CHARACTERISTICS FROM CATEGORIZED DATA

Lynn Roy LaMotte and Edward A. Blair, U. of Houston
College of Business, 4800 Calhoun Blvd., Houston, Texas 77004

1. INTRODUCTION

This paper concerns quantitative data obtained in categorical form. Examples include categorical income data, categorical frequency-of-use data, and categorical expenditure data.

Survey researchers often apply categorical measures to quantitative phenomena. A categorical measure provides less information than a "continuous" measure; however, for this very reason, a categorical measure may be viewed as less threatening to respondents or easier to answer, while still providing adequate information for the purpose at hand. In a mail survey, the researcher may feel that respondents are more likely to check a category than to write a numerical estimate.

Subsequent to gathering data via categories, a researcher may wish to estimate distributional characteristics, such as the sample mean and standard deviation, that could be obtained accurately only from detailed, uncategorized data. It may be desired to estimate bin frequencies for ranges of values other than those used to gather the data.

Most elementary statistics textbooks prescribe a procedure for obtaining sample means and standard deviations from grouped quantitative data. The same procedure is cited in professional research texts and is used by working researchers In this procedure, bins are replaced by bin midpoints and calculations accomplished as if only the midpoints had occurred as values.

In this paper we present a procedure for smoothing histograms for categorized quantitative data. In addition to providing a smoother histogram, the procedure provides an attractive alternative to the midpoint method for calculating numerical distributional characteristics.

2. THE MIDPOINT METHOD

With the midpoint method for calculating sample moments from grouped data, each value in a bin is replaced by the midpoint of the bin. This has the effect of replacing the values in the data set by a discrete relative frequency distribution. Sample moments are then calculated from this discrete relative frequency distribution in the same way that moments are calculated from a discrete probability distribution. Open-ended intervals, such as "forty or over", cause some ambiguity as to what value should be used in place of the midpoint.

Two points can be made about the midpoint method. First, distributional characteristics, such as moments, calculated with the midpoint method may differ considerably from results that would have been obtained, had the values not been replaced by bin midpoints. Second, results from a given set of responses may give different moments, depending on the bin definitions used.

Figures 1 and 2 are histograms for the same set of values. All values are nonnegative integers. In Figure 1, the histogram includes a bar for each integer value represented in the data set, so that the histogram is exact except for the representation of discrete values by bars. This representation is for visual purposes only and does not affect any calculated values. In Figure 2, values are grouped into four ranges. The representation in Figure 2 of the values in the rightmost range as being uniformly distributed across that range is not appealing. In that range, it appears that lesser values should have greater relative frequencies than greater values. Admittedly, this objection is difficult to state precisely, and if we had only Figure 2 we would have to concede that the original data that yielded the histogram could have been distributed exactly as shown in Figure 2. Still, we probably would feel more comfortable with a representation like Figure 1, in which the relative frequencies blend more smoothly with their neighbors.

3. A NEW METHOD FOR SMOOTHING HISTOGRAMS

In the remainder of this paper we present a simple method for smoothing histograms in hopes of recovering some of the shape of the original or underlying set of values in the data set. With this method, our purpose is to investigate how well it recovers descriptive statistics, such as the sample mean and sample standard deviation, and to begin to investigate the effects of bin definitions on such descriptive statistics.

The procedure smooths the CDF of the grouped data, producing a new, smoothed CDF from which the smoothed histogram is computed directly. Beginning with a table of endpoints and cumulative relative frequencies for the grouped CDF, additional points and approximated relative frequencies are found to minimize a smoothness criterion. The criterion is the sum of squared second divided differences. The additional points and relative frequencies must satisfy the monotonicity constraints of a CDF. The resulting problem may be stated as a quadratic programming problem with linear inequality constraints. For the results shown in this paper, we have used the IMSL constrained quadratic programming routine QPROG.

4. THE METHOD ILLUSTRATED

The examples shown next are based on the distribution shown in Figure 1. This distribution is, in fact, a mixture of three negative binomial distributions. Figure 2 shows a particular grouping of the distribution from Figure 1, with bins defined as 0, 1, 2-6, and 7-30. Our method was used to smooth the histogram of Figure 2, resulting in the histogram shown in Figure 3 with bins 0, 1, ..., 30.

Figure 4 shows the histogram when the relative frequencies of Figure 1 are grouped in bins 0, 1-3, 4-6, and 7-30. Applying our smoothing algorithm to the distribution in Figure 4 results in the distribution shown in Figure 5. Figure 5, obtained by smoothing the grouped relative frequencies shown in Figure 4, matches the original distribution shown in Figure 1 better than Figure

3 does. The difference between Figure 5 and Figure 3 is particularly striking because the bins for Figure 2 and Figure 4 differ only in that the second bin includes only 1 in Figure 2 while it includes 1-3 in Figure 4.

Table 1 shows means and standard deviations for ten different sets of bins for the Figure 1 distribution. The ten bins are defined by manipulating the endpoints of the second and third bins, leaving the other three endpoints fixed to define four bins. The bins are 0, 1-e_2, e_2-e_3, e_3-30, with e_2 = 1, 2, 3, 4, 5 and e_3 = 6, 8. Grouped means and standard deviations were obtained with the midpoint method from the grouped relative frequencies. To obtain the smoothed means and standard deviations, each grouped histogram was smoothed by our algorithm into bins 0, 1, ..., 30, then the mean and standard deviation calculated with midpoints of these bins and the smoothed relative frequencies.

Table 1 illustrates clearly that means and standard deviations are influenced considerably by bin definitions. Further, for the example used to construct Table 1, means and standard deviations calculated after smoothing are consistently closer to the "true" values than results from the grouped distributions.

FIGURE 1

PDF for Mixture of Neg. Binoms.

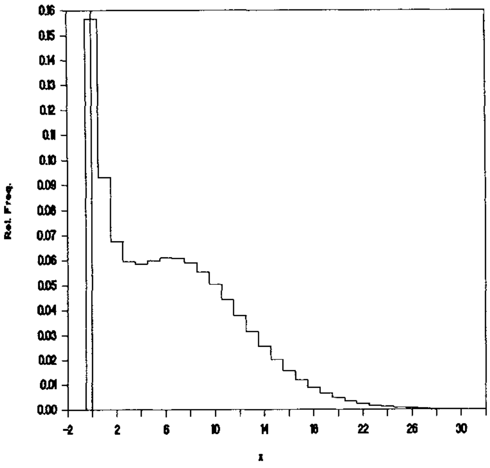

FIGURE 2

Grouped Histogram

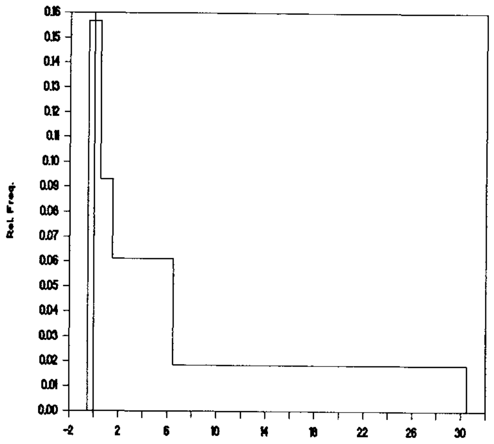

FIGURE 3

Smoothed Histogram

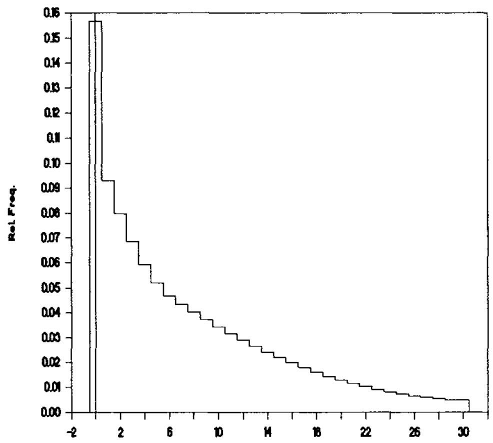

FIGURE 4

Grouped Histogram

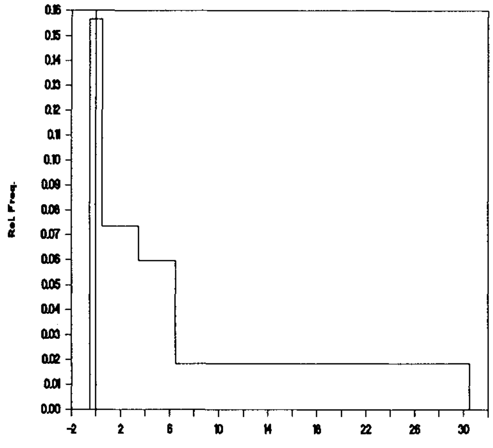

FIGURE 5

Smoothed Histogram

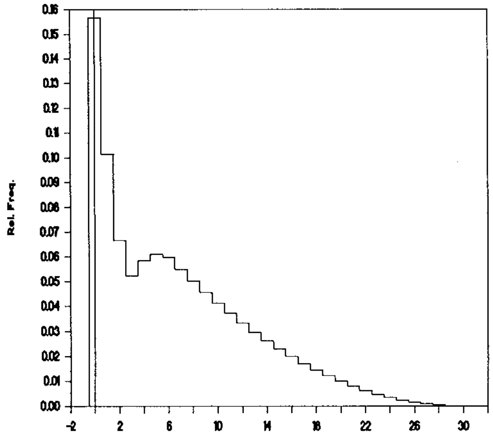

TABLE 1
Grouped and Smoothed Means and Standard Deviations
for Ten Sets of Bins

Second	Third	Grouped		Smoothed	
Endpnt	Endpt	Mean	St.Devn.	Mean	St.Devn.
1	6	9.526	9.377	7.527	7.313
2	6	9.523	9.377	7.009	6.425
3	6	9.544	9.360	6.760	6.022
4	6	9.568	9.343	6.650	5.909
5	6	9.589	9.330	6.643	5.987
1	8	8.537	8.699	6.745	6.399
2	8	8.526	8.700	6.498	5.699
3	8	8.548	8.685	6.390	5.462
4	8	8.573	8.669	6.305	5.367
5	8	8.594	8.657	6.230	5.347
True Values		6.300	5.309	6.300	5.309