Social Studies 201
January 16, 2004

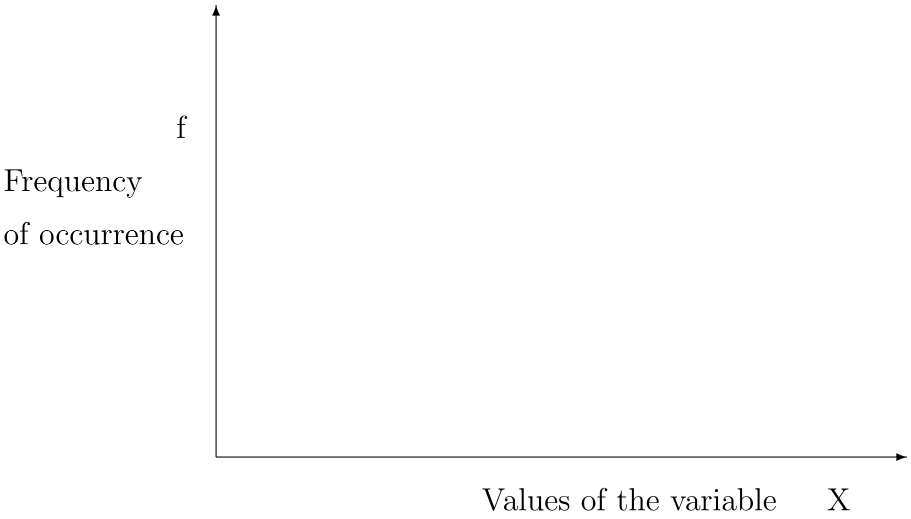
Histograms and Density

1. Introduction

From a frequency or percentage distribution table, a statistical analyst can develop a graphical presentation of a frequency distribution. One common way to do this is with a bar chart or histogram. These notes outline procedures for constructing these.

The conventional statistical approach to presenting a frequency distribution diagram is to place the variable on the horizontal axis and the frequency, or percentage of cases, on the vertical axis. As in Figure 1, the values of the variable (X) are placed along the horizontal axis and the vertical axis represents the number, or percentage, of cases.

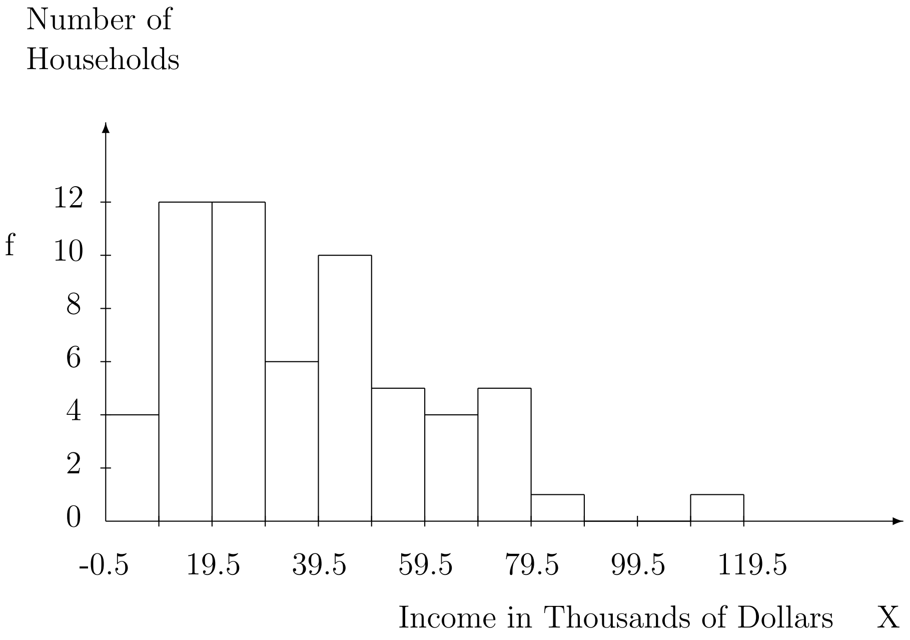
Figure 1: Conventional labelling of axes for diagram of frequency distribution



While the axes could be reversed, so frequencies are on the horizontal and values of the variable on the vertical, the convention in most statistical presentations is as in Figure 1.

Various forms of line and bar charts can be presented using this format. See section 4.7.1, pp. 121-126, of the text for examples of line and bar charts.

Figure 2: Histogram of income distribution for sixty Saskatchewan households, 1994



2. Histograms

A histogram is a bar chart where the height of each bar represents the relative frequency of occurrence of the cases, that is the frequency of occurrence per unit of the variable X. When data are grouped into intervals of equal width for X, the histogram is equivalent to a bar chart. In the case where the vales of X are of different interval width, the presentation of the bar chart is modified. Each of these cases is examined here.

a. Intervals of equal width

The bar chart of the stem and leaf display of January 14, 2004 is in Figure 2. Here the data have been grouped into intervals of ten thousand dollars, and the height of the bars represents the frequency of occurence for each category or interval. For example, the bar for an income of 9.5 to 19.5 is

12 units high because there are twelve households with income in this range. For the interval 109.5-119.5, there is only one household, so the bar there is only one unit high.

In words, this histogram can be described as follows. There are relatively few households at the lowest income level of less than ten thousand dollars, with the greatest number of households at incomes from 10-29 thousand dollars, and considerable numbers of households with incomes of 30-49 thousand dollars. As income increases, there are relatively fewer households at higher income levels.

b. Intervals of different width

In some cases it may make sense to reduce or expand the number of intervals into which the data are grouped. In the above example, there are relatively few households with incomes of eighty thousand or more dollars, but the intervals continue to 110-119. In some data sets, there may be values even greater than this, with the result that there would need to be many intervals to account for all possible values. In order to handle such situations, it may make sense to collapse the intervals, so there are fewer, but wider, intervals for values that occur infrequently. Where there are many more cases in some intervals than in others, it may make sense to split such an interval into two or three intervals, to provide greater information on the distribution of cases within such an interval.

As an example, In Table 1 the household incomes from the stem and leaf display are grouped into intervals of ten thousand dollars each for several of the intervals. But the original 10-19 thousand dollar interval is split into two (10-14 and 15-19) and for higher incomes, the cases are grouped into wider intervals. The intervals of 50-59 and 60-69 are combined into a single interval from 50 to 69 with nine households. The nine cases in this interval come from 4 households with income of 50-59 and 5 households with income of 60-69. Similarly, the intervals of 70-79, 80-89, and 90-99 have been grouped together into the 70-99 interval, and the highest incomes have been grouped into the interval 100-119.

Using these new intervals makes the presentation of the data more detailed for the 10-19 thousand dollar interval and more efficient for the higher income intervals, in that there are fewer intervals at the upper income levels. When presenting the data from Table 1 graphically, it is advisable to compensate for the different interval widths. In order to provide an accu-

Table 1: Distribution of income, sixty Saskatchewan households, 1994

Income in Thousands of Dollars (X)	Real Class Limits	f
0-9	-0.5-9.5	4
10-14	9.5-14.5	8
15-19	14.5-19.5	4
20-29	19.5-29.5	12
30-39	29.5-39.5	6
40-49	39.5-49.5	10
50-69	49.5-69.5	9
70-99	69.5-99.5	6
100-119	99.5-119.5	1
Total		$n = 60$

rate visual presentation of how household income is distributed, the density or relative frequency of occurrence of the number of cases in each interval should be calculated. This is done as follows.

c. Density

The density of occurrence of a variable is the number of cases per unit of each value of the variable. For each interval, the density, or relative frequency, is:

$$\text{Density} = \frac{\text{frequency}}{\text{interval width}}$$

The idea of obtaining the density is to adjust for the different width of the intervals, so the frequency of occurrence of the variable per unit of the variable is presented consistently across the different values of the variable. Table 2 provides a calculation of the densities for the grouping in Table 1.

By dividing the frequency of occurrence by the interval width, the densities represent the relative frequencies of occurrence of the variable, after

Table 2: Distribution of income, densities or relative frequency of occurrence

Income in Thousands of Dollars (X)	Real Class Limits	width (w)	f	Density=f/w
0-9	-0.5-9.5	10	4	4/10=0.40
10-14	9.5-14.5	5	8	12/10=1.60
15-19	14.5-19.5	5	4	12/10=0.80
20-29	19.5-29.5	10	12	12/10=1.20
30-39	29.5-39.5	10	6	6/10=0.60
40-49	39.5-49.5	10	10	10/10=0.10
50-69	49.5-69.5	20	9	9/20=0.45
70-99	69.5-99.5	30	6	6/30=0.20
100-119	99.5-119.5	20	1	1/20=0.05
Total		$n = 60$		

taking account of interval width. For example, in the first interval from 0 to 9 thousand dollars, there are 4 cases spread across 10 units of the variable, so the density is $4/10 = 0.40$. For the two narrower intervals of 10 to 14 and 15 to 19 thousand dollars, the interval width is five thousand dollars in each case. The densities here are $8/5 = 1.60$ over the 10 to 14 range and $4/5 = 0.80$ over the 15 to 19 thousand dollar range. In the case of the 70-99 thousand dollar interval, there are 6 cases, but they are spread across thirty thousand dollars of income. In this case, the density is $6/30 = 0.20$. That is, the density or relative frequency of occurrence, for the 0-9 interval is double that for the 70-99 interval. In this case there are relatively more households in the lower than in the higher interval.

These densities are especially useful when graphing the distribution from such a table as a histogram. If the densities are used, the bars of the histogram accurately represent the relative frequencies of occurrence of the variable. For the frequency distribution of Table 2, Figure 3 presents the histogram. Note that the horizontal axis is the same as in the case of Figure 2, but the vertical axis is labelled with the densities of occurrence. If the densities had not been used, the diagram would be inaccurate, over-representing wider intervals. But by calculating densities, each bar is the proper height, representing the relative frequency of occurrence of cases for each interval.

Figure 3: Histogram of income distribution for sixty Saskatchewan households, 1994

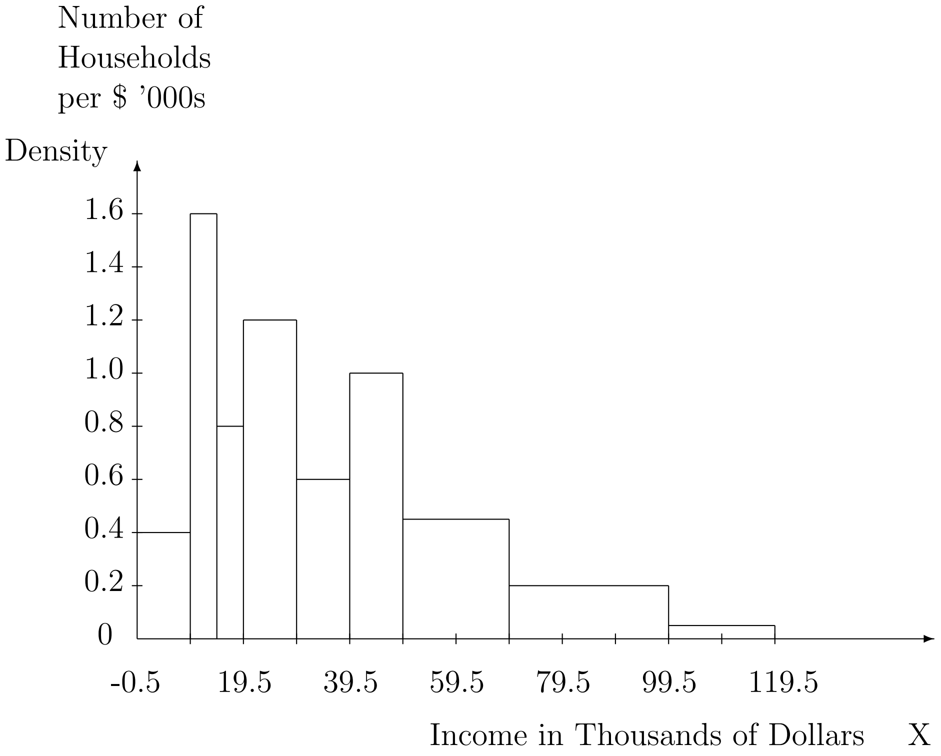


Figure 3 contains the histogram of income distribution for the sample of sixty households, where incomes are grouped into the intervals of Table 2. Comparing Figures 2 and 3, it can be seen that the histograms have the same general shape, even with intervals of different width. This is the way such a diagram of a frequency distribution should be – the basic shape of the curve should not change merely as a result of the manner the data are grouped. The only real difference in Tables 1 and 2 is that the grouping of Table 2 has more detail at 10-19 thousand dollars and not as much detail at higher income levels, as compared with Table 1. As a result, the bars on the right end of Figure 3 are wider.

3. Conclusion

If all intervals for a variable X are of the same interval width, then the histogram is constructed with bars equal in height to the frequency of occurrence of each interval. But where the intervals into which data are grouped have different interval width, it is best to calculate and graph the densities on the vertical axis. This presents a more accurate representation of the actual distribution of the values of the variable.

One other use for densities is the calculation of the mode. While this is covered in more detail in Chapter 5, the mode, or most common value, is the value of the variable where the histogram reaches its peak. In order to accurately determine this, when intervals are of unequal size, densities can be used to determine where the peak of the histogram occurs.

Last edited January 17, 2004.