Social Studies 201
September 20-22, 2004

Histograms

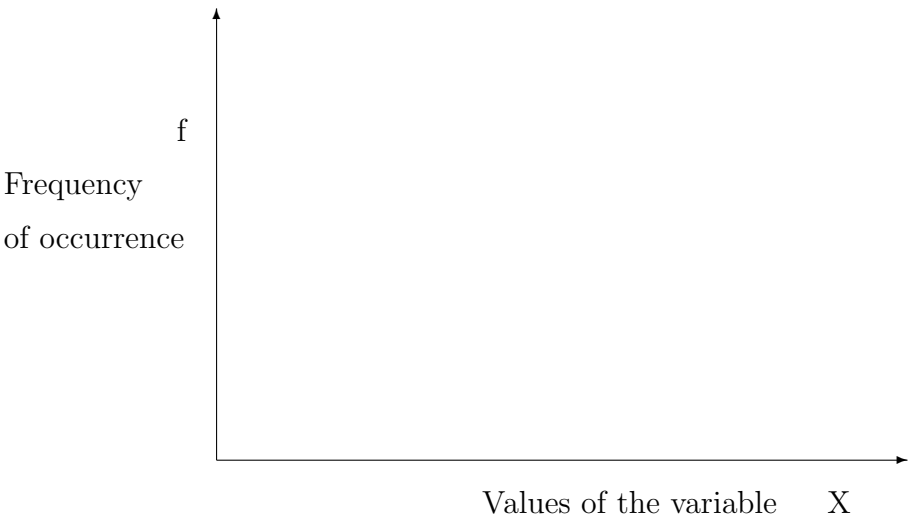
See text, section 4.8, pp. 145-159.

Introduction

From a frequency or percentage distribution table, a statistical analyst can develop a graphical presentation of the 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 when presenting a distribution as a diagram is to place the values of the variable on the horizontal axis and the frequency, or percentage of cases, on the vertical axis. The diagram of Figure 1 presents the values of the variable (X) on the horizontal axis, with the vertical axis representing the frequencies (f), or number of cases, taking on each value of X. For a diagram of a percentage distribution, percentages (P) replace frequencies (f) on the vertical axis.

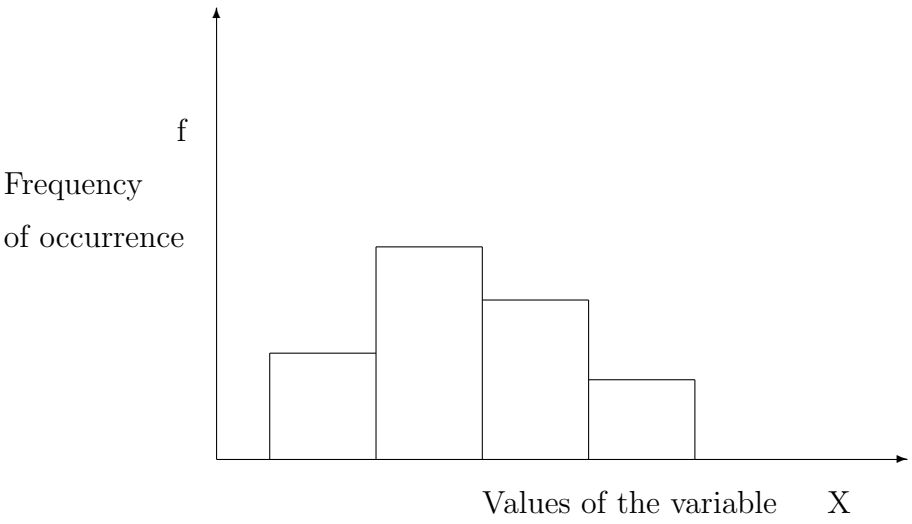
Figure 1: Conventional labelling of axes for diagram of frequency distribution



Sometimes the axes are reversed, with frequencies on the horizontal axis

and values of the variable on the vertical axis. But the most common convention concerning diagrams of frequency distributions is to use the approach of Figure 1.

Figure 2: Generic histogram of a frequency distribution



An example of how a histogram might appear is contained in Figure 2. Bars are drawn so the height of the bars represent the relative frequency of occurrence of the variable. A quick glance at the histogram shows that there are relatively few cases at the lowest and highest categories, with most cases in the middle two categories. The single most common category is that of the second category from the left.

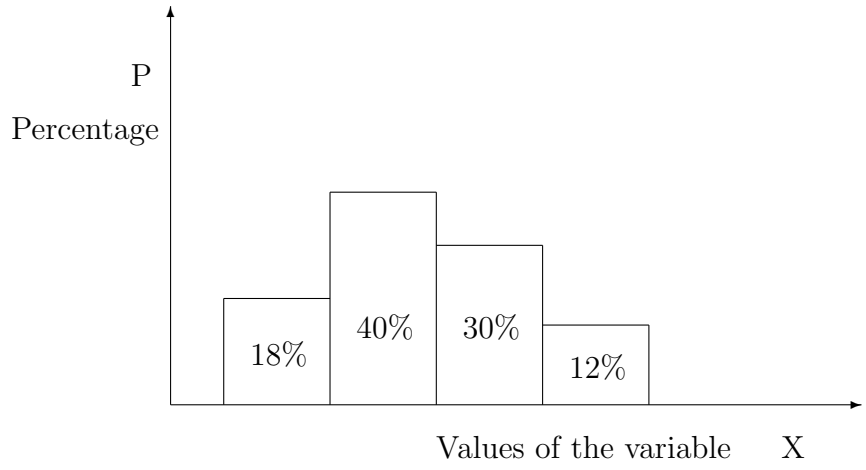
Guidelines and procedures for developing such a histogram are contained in the following notes. Various forms of line and bar charts can be presented using a similar format. See the text, section 4.7.1, pp. 121-126 for examples of line and bar charts. In this section of Module 2, only histograms are discussed.

Histograms

A histogram is a bar chart of a frequency distribution, where the height of each bar represents the relative frequency of occurrence of the cases in that interval. More exactly, the relative height of each bar of a histogram represents the frequency of occurrence per unit of the variable X. A histogram, such as that of Figure 2 provides a quick visual picture of how the cases are spread across the values of the variable, giving a good idea of the frequency or percentage distribution.

One characteristic of a histogram is that all the cases are represented by the areas of the bars in a histogram. This is most easily understood in terms of the histogram of a percentage distribution, where there are one hundred per cent of cases in total. This is illustrated in Figure 3. This is the same histogram as in Figure 2, but with percentages replacing frequencies on the vertical axis and approximate percentages of cases associated with each bar labelled inside the bar.

Figure 3: Histogram of a percentage distribution



In Figure 3, each bar is labelled with the approximate percentage of cases associated with each category. As can be seen there, the height of the bars represents the relative percentage in each category. In addition, the sum of the percentages in the four bars is one hundred per cent, that is

$18 + 40 + 30 + 12 = 100\%$. The percentage label in each bar indicates the percentage of cases having the values of X associated with that bar. Bar charts of this type are termed histograms.

Constructing and interpreting a histogram is straightforward when the intervals into which data are grouped are of equal width. In this case, the height of the bars represents the relative frequency of occurrence of the variable X, per unit of X. However, when the intervals used to group the variable are of different interval width, the height of the bars should be adjusted to take the different interval widths into account.

The following notes show how to construct a histogram in each of the above situations. The case of intervals of equal width is illustrated first, followed by a discussion of how to adjust the heights of the bars to take account of intervals of unequal width. The data used in each case is the distribution of incomes of Saskatchewan clerical workers, used for the stem-and-leaf display.

Histogram – intervals of equal width

When the values of a variable X have been grouped into intervals of equal interval width, the construction of a histogram is straightforward. The histogram is composed of vertical bars, with the height of each bar equal to the frequency, or percentage, of occurrence of the variable. The width of each bar is equal to the width of the interval so, in the case of equal interval widths, each bar has the same width. If the variable is continuous, the bars touch each other, so there are not gaps in the histogram. That is, the bars are constructed using widths representing the real class limits for the variable.

To illustrate how to construct a histogram in the case of equal interval widths, the frequency distribution of Table 1 will be used. This is a frequency distribution of incomes of a sample of sixty Saskatchewan clerical workers, originally used in the stem-and-leaf display of Example 2.10.

Table 1: Frequency distribution of income, sixty Saskatchewan clerical workers, 2000

Income in Thousands of Dollars (X)	f
0-9	3
10-19	13
20-29	15
30-39	17
40-49	9
50-59	2
60-69	1
Total	$n = 60$

The intervals of Table 1 match the groupings initially used to construct the stem-and-leaf display for these data. These groupings are 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, and 60-69, so the intervals are of equal width. Each interval is to represent ten thousand dollars of income, but these apparent class limits make it appear that the intervals represent only nine thousand

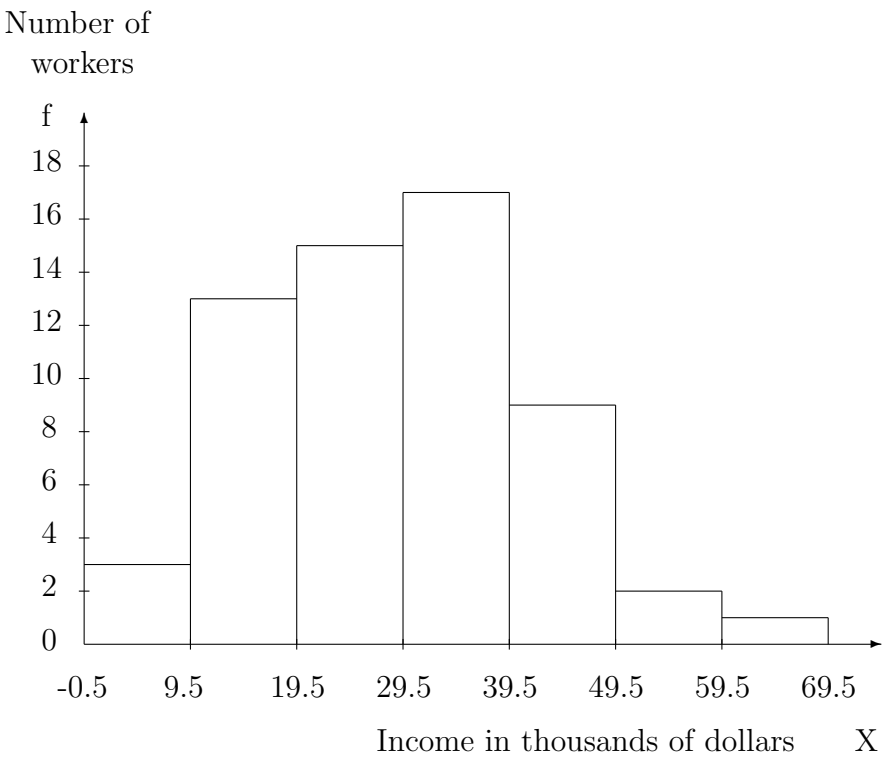
dollars of income. The ten unit width for each interval is obtained by calculating the real class limits. This is done in Table 2. This is the same frequency distribution as presented earlier, but with real class limits identified.

Table 2: Frequency distribution of income, sixty Saskatchewan clerical workers, 2000; apparent and real class limits

Apparent class limits for X	Real class limits for X	Frequency f
0-9	-0.5-9.5	3
10-19	9.5-19.5	13
20-29	19.5-29.5	15
30-39	29.5-39.5	17
40-49	39.5-49.5	9
50-59	49.5-59.5	2
60-69	59.5-69.5	1
Total		$n = 60$

The data in Table **??** are used to construct the histogram of Figure 4. The vertical axis represents the frequency of occurrence for each value of X and the horizontal axis represents the values of X. The width of each bar is equal to ten thousand dollars, representing the fact that the values of X are grouped into intervals of ten thousand dollars each. The height of each bar equals the frequency of occurence of each category or interval. For example, the bar for an income of 9.5 to 19.5 is thirteen high because there are thirteen individuals with income in this interval. For the interval of 59.5 to 69.5, the bar is only one high since there is only one individual in this category of income.

Figure 4: Histogram of income distribution for sixty Saskatchewan clerical workers, 2000



In words, this histogram can be described as follows. There are relatively few individuals with incomes at the lowest income level of less than ten thousand dollars. The greatest number of individuals have incomes from 9.5 to 39.5 thousand dollars, with considerable, although lesser, numbers of individuals having incomes of 39.5 to 49.5 thousand dollars. The peak number of individuals occurs in the interval representing those with incomes from 29.5 to 39.5 thousand dollars. As income increases above this, there are fewer households at each successively higher income level.

Summary. When a variable is grouped into intervals of equal width, a histogram is a bar chart with height equal to the frequency, or percentage, of occurrence. The width of the bars is equal to the common interval width. If the variable is continuous, it is best to use real, rather than apparent, class limits. Since all cases are represented by the bars, and interval widths are equal, the relative number of cases in each interval is proportional to the size of each bar.

Histogram – intervals of unequal width

When the values of a variable X have been grouped into intervals with different interval widths, the construction of a histogram is much less straightforward than in the case of equal interval widths. The problem is that a statistical analyst can always make a bar look larger by widening the interval used to group the values of the variable X. If the different width of the bars is not taken into account, the bar chart gives a distorted view of the distribution. A corrected bar chart can be constructed, using what are termed denities, or relative frequencies of occurrence.

In the notes in this section, an incorrect histogram is presented to illustrate the distortion that grouping can cause. Then the method of using densities to produce a more accurate portrayal of the distribution is discussed.

Construction of an incorrect histogram

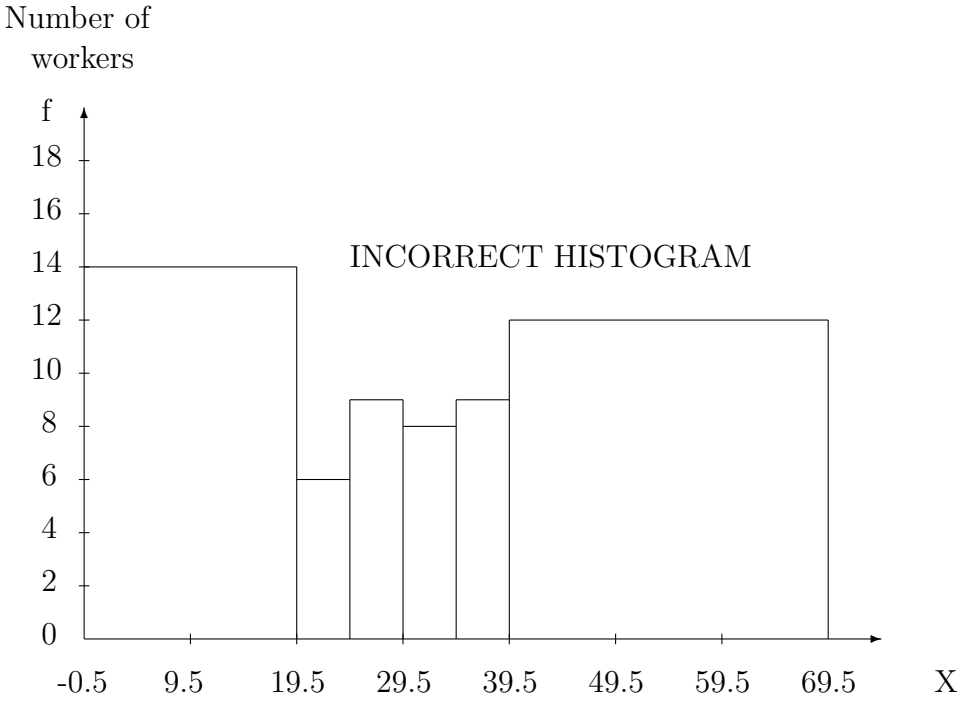
In order to illustrate the problem associated with intervals of different width, the distribution of income for sixty Saskatchewan clerical workers is again used. But for this example, the grouping of incomes is changed from earlier presentations, with some narrower and some wider intervals. The frequency distribution with intervals of different width is presented in Table 3. This is the same frequency distribution as in Table 2, but with a reorganization of the cases into different intervals. Such a regrouping might make sense in that narrower intervals of five thousand dollar width are used where there are more cases (20 to 40 thousand dollars), thus providing greater detail where X is more concentrated. In contrast, wider intervals are used at the smallest and largest values of X, where there were fewer cases in each of the original intervals.

Table 3: Distribution of income, sixty Saskatchewan clerical workers, 2000 apparent and real class limits

Apparent limits	Real limits	f
0-19	-0.5-19.5	16
20-24	19.5-24.5	6
25-29	24.5-29.5	9
30-34	29.5-34.5	8
35-39	34.5-39.5	9
40-69	39.5-69.5	12
Total		$n = 60$

Using the frequency distribution of Table 3, and not adjusting the frequencies for different interval width, results in the incorrect histogram of Figure 5. In this figure, the bars have the width of the intervals and frequencies of Table **??**, without making any adjustments.

Figure 5: Incorrect histogram of income distribution for sixty Saskatchewan clerical workers, 2000



If no adjustment for interval width is used, the incorrect histogram of Figure 5 is the result. This incorrect histogram comes from the same data as produced the correct histogram of Figure 4. While the regrouping might be expected to produce a slightly different histogram, the incorrect histogram gives a completely different, and inaccurate, view of the distribution. That is, rather than showing a peak at the middle income levels, as should occur, the incorrect histogram shows peaks at the lowest and highest levels of income. The shape of the correct histogram in Figure 4 is completely distorted in Figure 5 to give an incorrect picture of the distribution.

In order to correct for the different interval widths, the concept of density is introduced in these notes. Calculating and using these densities produces a new histogram that is similar in shape to the original histogram of Figure 4.

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 of occurrence, is:

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

The reason for obtaining the density is to adjust for the different width of the intervals, so the frequency of occurrence per unit of the variable is presented consistently across the different values of the variable. That is, the density represents the frequency of occurrence per unit of the variable.

In the case of a percentage distribution, densities are obtained in the same manner, but with the percentages replacing frequencies. In this case, the densities represent the percentage of cases per unit of the variable X.

When making a diagrammatic presentation of a frequency, or percentage, distribution with unequal interval widths, it is best to calculate densities prior to constructing the histogram. The bars of histogram are drawn using the interval widths as given in the table of the distribution, but drawing the bars with heights equal to densities, rather than frequencies or percentages.

The result this has on a histogram is illustrated using the example of incomes of sixty Saskatchewan clerical workers. A frequency distribution with unequal interval widths was presented in Table 3. When densities were not obtained for this distribution, the diagrammatic result was to produce the misleading histogram of Figure 5. In Table 4, the same frequency distribution is presented, and the desnities associated with each interval are calcualted; Figure 6 presents the histogram using these densities.

When obtaining densities for a distribution of a continous variable with unequal interval widths, make sure you determine interval widths using the real class limits. If the apparent class limits had been used to determine interval width in Table 4, the intervals would be widths such as nineteen (from 0 to 19), four (from 20 to 24), and so on. This would result in losing a unit of income when moving from one interval to the next. By constructing and using the real class limits, this gap is eliminated and the proper interval widths are obtained. For example, for 20-24, the real class limits are 19.5 to 24.5, implying an interval width of five thousand dollars.

Table 4: Distribution of income, sixty Saskatchewan clerical workers, 2000; unequal interval width, apparent and real class limits

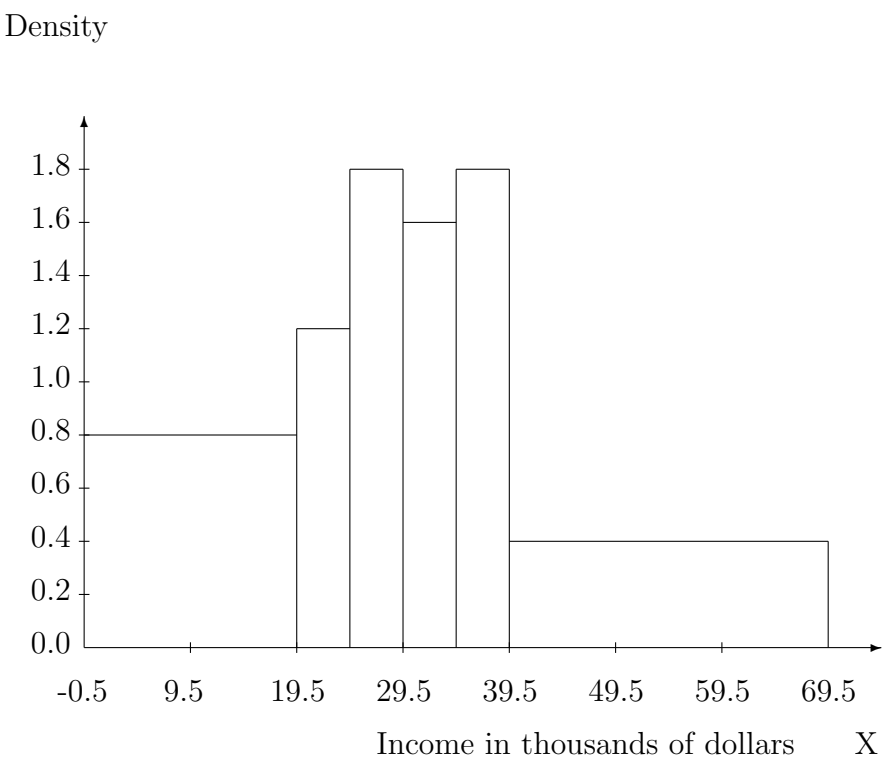
Apparent class limits	Real class limits	Interval width	f	Density
0-19	-0.5-19.5	20	16	0.8
20-24	19.5-24.5	5	6	1.2
25-29	24.5-29.5	5	9	1.8
30-34	29.5-34.5	5	8	1.6
35-39	34.5-39.5	5	9	1.8
40-69	39.5-69.5	30	12	0.4
Total			$n = 60$	

Densities for Table 4 are obtained by dividing the frequencies by interval widths. This means the densities represent the frequency of occurrence per unit of the variable X. For example, across the interval from 0 to 19, there are 16 cases, spread across twenty thousand dollars. For this interval, this means a density of $16/20 = 0.8$ per thousand dollars. Across the 20 to 24 thousand dollar interval, an interval representing five thousand dollars of income, there are 6 cases, implying a density of $6/5 = 1.2$. While there are fewer cases in the 20-24 interval than in the 0-19 interval, the 20-24 interval is much narrower, so the cases are more densely packed in the 20-24 than in the 0-19 interval. Other densities are obtained in a similar manner.

Figure 6 is a histogram for this distribution. The values of X are again on the horizontal axis, but the interval widths are equal to those associated with the real class limits. Densities are placed on the vertical axis, and the height of each bar is equal to the density for the interval, as listed in the last column of Table 4.

When constructing a histogram using densities, make sure you define the density, either in the table or in the text accompanying the table. While the histogram produced by using densities is as correct a portrayal as possible of the distribution in the table, it assists a reader examining the table to know exactly how the densities have been calculated. In Figure 6, the unit for density is defined in a note at the bottom of the figure.

Figure 6: Correct histogram of income distribution for sixty Saskatchewan clerical workers, 2000; unequal interval width



Note: Density is the frequency of occurencer per thousand dollars of income.

The histogram of Figure 6 provides a similar general picture of the distribution of clerical incomes as did the original histogram of Figure 4. That is, the histogram of Figure 6 reaches its peak in the centre, with bars of less height at both lower and higher incomes. That is, the greatest concentration of workers is at the middle incomes with fewer at the very lowest incomes, and even fewer at the very highest incomes. In Figure 4, intervals were of equal width, so each bar has the same width. In Figure 6, there are more bars, and thus more detail, at the middle income level, but wider bars and less detail at the lower and higher incomes.

The essential aspect of Figure 6 though is that it is similar in shape to the original histogram of Figure 4. This was the intention in using densities, and that result has been achieved.

This example demonstrates that when groupings for the variable X do not use equal interval widths, it is necessary to adjust for these different widths. If this is not done, a distorted picture of the distribution is presented. The example also illustrates how to correct this problem, that is, by graphing the densities. This produces a histogram giving a proper picture of how the variable is distributed.

Using densities. In this course, densities are used for only two purposes – drawing histograms when intervals have unequal width and determining the mode (see Module 3). The application to histograms is discussed above. Previewing Module 3, the mode is the value of the variable at the peak of the distribution. In order to properly determine the peak of a distribution organized into unequal interval widths, the densities need to be obtained. This will be discussed again in Module 3.

Conclusion to histograms

Histograms provide a convient way to present distributions visually. These may provide a reader with a quick and comprehensive view of the structure of a distribution. Following are a few concluding notes about constructing and using histograms.

Constructing histograms. If a table of a distribution uses intervals of equal width, the construction of the histogram is straightforward, with the height of each bar equal to the frequency or percentage of cases in the each interval.

When a table of a distribution contains intervals of unequal width, it is best to calculate densities prior to constructing a histogram. The bars are then constructed with width equal to the interval widths and the height of the bars is equal to the density of cases for the interval. The size of the bars then represents the relative frequency of occurrence of the variable.

Using histograms. For later parts of the semester, the frequency or percentage distribution table provides a more useable format than a histogram. But in terms of allowing a reader to capture the essence of the distribution, the histogram is probably more useful. As a result, histograms or other forms of diagrams are commonly used when presenting data in the media. Pie charts, bar charts, line diagrams, and other charts that present the data in diagrammatic are widely used to visually portray distributions.

In later parts of this course, most of the work is conducted using tables of distributions, rather than diagrams. However, from time to time, diagrams are used to illustrate various aspects of distributions. When diagrams of frequency distributions are presented in later modules, these are generally histograms. However, they may not appear to be the same type of histogram as examined here in that the the diagrams are often smoothed, rather than presented as chunky bar charts. But the principles used later in the course are the same as those presented here for histograms. That is, in the smoothed histogram, all cases (one hundred per cent) are represented by the distribution. In addition, the areas in the diagram associated with particular intervals represent the relative occurrence of the variable for the interval.