Histograms*

David Lane

This work is produced by The Connexions Project and licensed under the
Creative Commons Attribution License †

Abstract

Introduction to histograms.

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psychology test. The test consists of 197 items each graded as "correct" or "incorrect." The students' scores ranged from 46 to 167.

The first step is to create a **frequency table**. [1] Unfortunately a simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table 1: Grouped Frequency Distribution of Psychology Test Scores.

Grouped Frequency Distribution of Psychology Test Scores

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

*Version 2.4: Jun 20, 2003 12:00 am GMT-5

†http://creativecommons.org/licenses/by/1.0

[1] Frequency Table: A table containing the number of occurrences in each class of data; for example, the number of each color of M&Ms in a bag. Frequency tables often used to create histograms and frequency polygons. When a frequency table is created for a quantitative variable, a grouped frequency table is generally used.

Table 1

To create this table, the range of scores was broken into intervals, called **class intervals**. The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval is counted to obtain the **class frequencies**. There are three scores in the first interval, 10 in the second, etc. Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too "choppy." More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. A histogram of these data is shown in Figure 1.

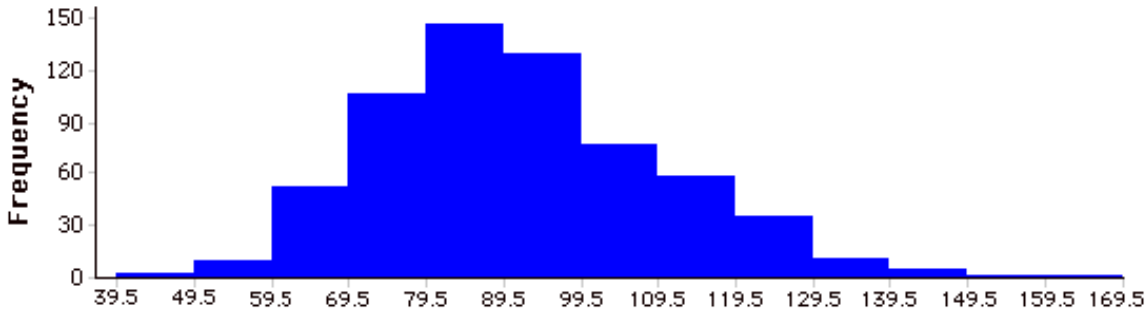

Figure 1: Histogram of scores on a psychology test.

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not symmetric; the scores extend to the right farther than they do on the left. The distribution is therefore said to be **skewed**[2]. (We'll have more to say about the shape of distributions in our discussion on summarizing distributions.)

In our example the observations are whole numbers. Histograms can also be used when the scores are measured on a more continuous scale such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, 4000, 5000, etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance, and is the practice of many computer programs that create histograms. Note also that some computer programs label the middle of each interval rather than the end points.

Histograms can be based on **relative frequencies** instead of actual frequencies. Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. In this case, the Y axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class

[2] A distribution is skewed if one tail extends out further than the other.

frequency by the total number of observations, and then (b) plotting the quotients on the Y axis (labeled as proportion).

There is more to be said about the widths of the class intervals, sometimes called **bin widths**. Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. There are some "rules of thumb" that can help you choose an appropriate width. (But keep in mind that none of the rules is perfect.) **Sturgis' rule** is to set the number of intervals as close as possible to $1+\log_2 N$, where $\log_2 N$ is the base 2 log of the number of observations. The formula can also be written as $1+3.3\log N$ where $\log N$ is the log base 10 of the number of observations. According to Sturgis' rule, 1000 observations would be graphed with 11 class intervals since 10 is the closest integer to $\log_2 1000$. We prefer the Rice rule, which is to set the number of intervals to twice the cube root of the number of observations. In the case of 1000 observations, the Rice rule yields 20 intervals instead of the 11 recommended by the Sturgis' rule. For the psychology test example used above, Sturgis' rule recommends 10 intervals while the Rice rule recommends 17. In the end, we compromised and chose 13 intervals for Figure 1 to create a histogram that seemed clearest. **The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.**

To provide experience in constructing histograms, we have developed an interactive demonstration. The demonstration reveals the consequences of different choices of bin width and of lower boundary for the first interval.

Interactive histogram[3]

[3] http://www.ruf.rice.edu/~lane/stat_sim/histogram/index.html