Constructing Bar Charts and Histograms:

Bar charts are easy to construct if the underlying variable is ordinal (e.g. the number of female or male meerkats who are babysitting at some time) since you simply let the height of each bar be the number of observations associated with each nominal value. There is no issue in this case with defining the width of the range covered by each bar. If the underlying variable is ordinal but discrete (for example, the number of tomatoes produced per plant must be an integer), then the only decision to be made is how to lump together the possible outcomes in deciding the range covered by each bar. In this case, generally you should assign equal ranges to each bar and use 10-20 bars, assuming the underlying data set has sufficient points to allow for more than 1-2 points per bar. For example, if the number of tomatoes per plant varied from 1 to 50, you could use 10 bars each of width 5 (1-5, 6-10,... , 46-50). Difficulties arise here if the number of bars doesn't divide the range of the data evenly - there is no one solution to this, because one bar would either have to be wider or narrower than the others, or else you must extend the range of the data beyond the observations. The usual rule is to have one of the bars cover a wider range of values than the other bars.

Constructing a histogram for a continuous data set involves:

(1) Decide on the number of classes (bars) for the histogram. Typically choose this to be 10-20, but if there is a small dataset, choose this smaller. A rule of thumb might be that the number of observations divided by the number of classes should be at least 4.

(2) Choose the width of the classes (all widths will be the same) by dividing the range of the data (Maximum value observed minus minimum value observed) by the number of classes from step (1). Round this value up so that the class width has the same precision (number of significant digits) as the measurements. So if we are measuring body weights of black bears and the smallest weight is 16.0 kg and the largest weight is 118 kg, and we wish to use 10 classes, then the class width is 10.2 kg.

(3) Select the smallest observed value as the lower limit of the first class (e.g. the left hand value for the first class), and add multiples of the class width from step (2) to obtain the lower limits for each class. So in the black bear case, the lower class limits would be
16.0, 26.2, 36.4, 46.6, … , 107.8

(4) Find the upper class limits (the right hand value for each class) by adding the class width to the lower class limits and subtracting from this the smallest significant digit in the observations. So if the class width is 0.3 and the lower limit of the first class is 3.7 (and the finest precision of the original data was .1 unit), then the upper class limit for the first class is 3.7 + .3 - .1 = 3.9 and the next class would have lower class limit 4.0 (so that there is a gap between the upper class limit of one class and the lower class limit of the

next class, and this gap is the precision of the original data. In the black bear case, since the highest precision of the original data is .1 kg, the upper class limits are
26.1, 36.3, 46.5, …, 118
Note that the last class goes up through the highest value in the data set.

(5) For the boundaries between classes, use the value which is halfway between the upper class limit of one class and the lower class limit of the next class. So in the above example, the class boundary would be 3.95 and the lowest class would be 3.7-3.95 with the next class being 3.95-4.25 and so on. In the black bear example, the class boundaries would be
26.15, 36.35, 46.55, …, 118

(6) Count up the number of observations in each class and make the height of the bar for each class equal to these numbers. You can also construct a "frequency histogram" by dividing the height of each bar of the histogram (the number of observations falling in that class) by the total number of observations (the sample size). For a frequency histogram, the sum of the heights of all bars is one.

Linear Regression:

One objective in using descriptive statistics is to aid in understanding whether there is a relationship between two measurements associated with an observation (e.g. are leaf length and leaf width related, are weights of bats related to their age since birth, is respiration rate of an individual related to the temperature of the environment around that individual). The simplest type of relationship that would exist between two measurements is that they are proportional (so that doubling one value leads to a doubling in the other). A slight extension of this is that they are linearly related (e.g. if the measurements are L and W, then L= a*W + b where a and b are constants).

A regression is a formula that provides the "best fit" of a particular mathematical equation to a set of data. A simple regression occurs if there are only two variables being considered, and a linear regression is one in which the relationship is assumed to be linear and the "best line" which goes through the data on a standard x-y graph is obtained. This is not meant to imply that one of the variables necessarily "determines" the other one, but rather that any complex dependence between them may be adequately summarized as a linear one. Strictly speaking, when there is reason to believe that there is a causal relationship between the measurements (e.g. variation in weights of bats is due to differences in their ages, but variation in ages of bats is in no way "caused" by their weights) then the term regression is used, otherwise we say there is a "correlation". Thus variation in leaf length is correlated with leaf width, but not "caused" by it.

The standard method to find the "best fit" line through a data set is called least squares - it is the line which minimizes the sum of the squares of the differences between the height of the line (y-value) and height (y-value) of each data point, where this sum is taken over all data points. This best fit line minimizes the sum of the square residuals. There are formulas for the slope and y-intercept of the best fit line (the regression line) that can be calculated using the original data points.