WISCONSIN FAST PLANTS

Variation: Data Organization, Display, and Analysis

Measurable differences will be found among individuals in a group or population. It is therefore important to know how much variation in a particular phenotype (observable trait) might be expected so that it can be determined whether the variation observed experimentally may be viewed as *normal* for that population. Normal would be defined as that range of potential phenotypes that a population would exhibit in a specified range of environmental conditions.

The species *Brassica rapa*, of which Fast Plants are a specially bred stock, is inherently genetically variable. Within a population of Fast Plants one can observe considerable phenotypic variation in some traits such as plant height or intensity of purple stem color. For this reason, it is important to determine what is a normal range of phenotypes for Fast Plants.

Organizing and Displaying Data: Graphical Representation

When, for example, the heights of a population of 48 AstroPlants are measured in millimeters at Day 10 and recorded (Table 2), considerable variation can be noted. Height is measured from soil level to shoot apex.

Table 2: Height, in mm, of 48 Fast Plants measured at Day 10 (hypothetical data).

33	40	32	59	18	45	73	21
49	52	60	55	33	56	32	52
50	84	54	25	57	45	68	41
43	53	43	76	49	39	36	50
62	27	66	39	41	51	55	41
30	47	72	37	44	35	45	48

The "Stem and Leaf Table"

Simply listed as a set of 48 numbers, relatively little information can be gained from them other than to note that they are variable. An easy way to begin to organize the numbers is to put them into what is commonly known as a *stem and leaf table* (Table 3).

Table 3: Fast Plant height data from Table 2 organized into a stem and leaf table.

		digits, "leaf"
tens,	0	
"stem"	1	8
	2	7, 5, 1
	3	3, 0, 2, 9, 7, 3, 9, 5, 2, 6
	4	8, 3, 0, 7, 3, 9, 1, 4, 5, 5, 5, 1, 0, 8
	5	0, 2, 3, 4, 9, 5, 7, 6, 1, 5, 2, 0
	6	2, 0, 6, 8
	7	2, 6, 3
	8	4
	9	

To do this, note that each number is broken into "tens" and "digits." Examine each number, breaking it into its tens and digits, e.g., 48 becomes 4 (tens) and 8 (digits). Make a vertical column "stem" listing from zero to 9 that represents the tens. Then enter the digit from each number in the horizontal row "leaf" corresponding to the appropriate ten or stem position; e.g., 48 is listed as an 8 in row 4 in Table 3. Numbers in the range from 10 to 19 go in the "1" row, while numbers in the range from 20 to 29 go in the "2" row, etc.

Considerable information about the population of 48 plants begins to become apparent from the stem and leaf table. For example, it can be observed that the most plant heights in this data set fit into the "4" stem. The numbers representing the plant heights in the population are a set of size 48 (n = 48).

Department of Plant Pathology, 1630 Linden Drive, Madison, WI 53706 1-800-462-7417 wfp@fastplants.cals.wisc.edu

The Frequency Table

The set of 48 plant heights can be organized into groupings or *classes* representing a specified range of values or *class interval* (i). In this example the class interval is 10 mm: i = 10 mm. The number of plants having heights within a particular class interval (e.g., 20 to 29 mm) is the *class frequency* (f_i). The *relative frequency* of a class is determined by looking at the number of measurements in a class (f_i) relative to the number of measurements in the entire data set (n): f_i/n.

With the above information the set n = 48 can be arranged in a *frequency table* by counting and recording the numbers in each class (f_i) and calculating the proportion of numbers in each class to the total set (f_i/n). The relative frequency of the class interval 20 to 29 mm in the example set of 48 plant heights is $f_i/n = 3/48 = 0.06$.

Table 4: Frequency table of heights, in mm, of 48 Fast Plants at Day 10, grouped in classes of 10 mm intervals and relative frequency of each class.

class interval, i	0	10	20	30	40	50	60	70	80	90
class frequency, f_i	0	1	3	10	14	12	4	3	1	0
relative frequency, f_i/n	0	0.02	0.06	0.20	0.29	0.25	0.08	0.06	0.02	0

n = 48, i = 10
Note: relative frequency fractions should add up to 1, rounding numbers in this example reduced this to 0.98.

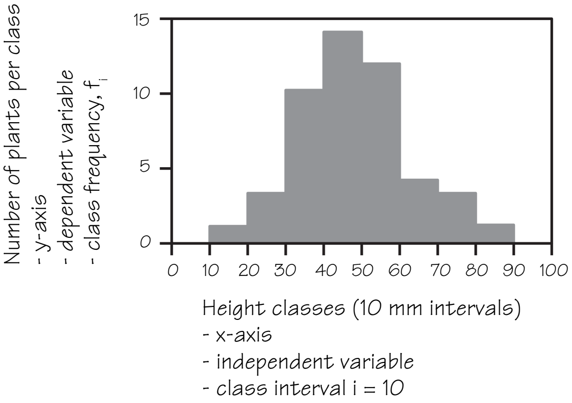
The Frequency Histogram

The relationship among the numbers in each class can be more effectively visualized by displaying them as a *frequency histogram* in which the data are treated as two variables, x and y, and plotted in relation to each other in a two-dimensional graph with the x and y axes at 90˚ to each other.

The first variable, the class interval (i), was chosen to be i = 10 is the *independent variable* as it was predetermined by choice. The independent variable is arrayed on the *x* or *horizontal axis* just as it appears in the frequency table (Table 4).

The second variable is the class frequency (f_i) and is known as the *dependent variable* because the number in the particular class (i) depends on the class chosen and is arranged and plotted on the *y* or *vertical axis* of the graph. When plotting the x and y axes of a graph it is important to consider the size or scale of a unit on each axis so that an effective symmetry is achieved in the presentation of the graph. Figure 1 is a frequency histogram of the data from the frequency table, Table 4.

Figure 1: Frequency histogram of heights, in mm, of 48 Fast Plants at Day 10, grouped in class intervals of 10 mm.



The relative frequency (f_i/n) from the frequency table can also be plotted as a *relative frequency histogram*. In this case the x-axis remains the same as in the frequency histogram and the y-axis is arrayed in units of decimal fractions. The appearance of the relative frequency histogram is similar to the frequency histogram, however what is being portrayed is the relative proportion of a class size in relation to the set.

Choosing the proper class interval can be important to the process of analyzing and understanding the information that is codified in the data set of plant height measurements. If the chosen class interval is too small or too large, certain relationships among the individuals within the set will not be evident.

For example if a class interval of i = 25 rather than i = 10 were chosen then the frequency histogram would appear as in Figure 2 or if a class interval of i = 2 were selected the frequency histogram would appear as in Figure 3.

Figure 2: Frequency histogram of heights, in mm, of 48 Fast Plants at Day 10, grouped in class intervals of 25 mm.

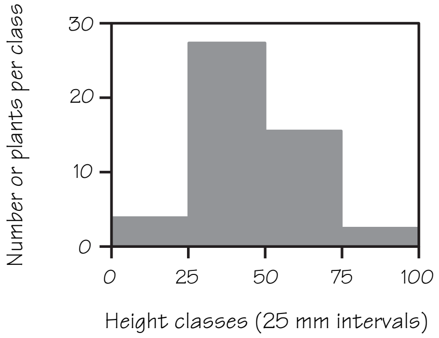
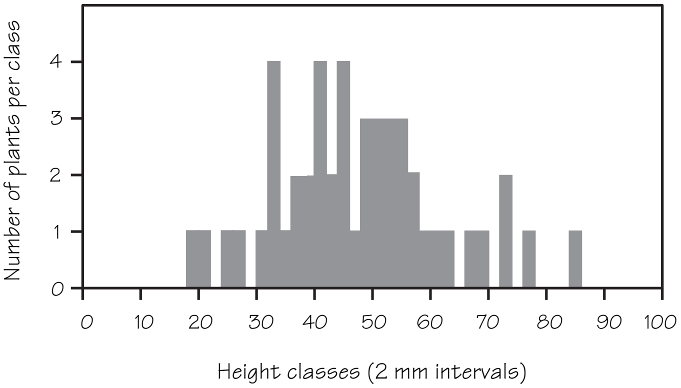


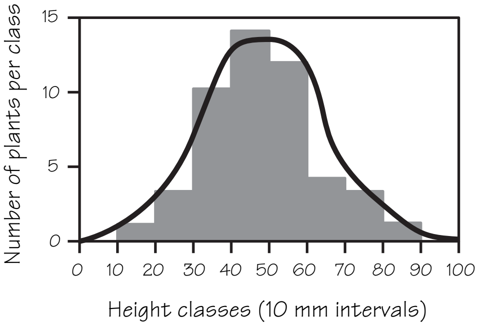
Figure 3: Frequency histogram of heights, in mm, of 48 Fast Plants at Day 10, grouped in class intervals of 2 mm.



The Normal Curve

The outline of a frequency histogram roughly depicts a curve known as a *frequency curve*. Frequency curves can assume various different shapes. Interpretation of the shapes can give insight into underlying phenomena conditioning the expression of the phenotype's contribution to the curve. For instance, the data on plant height recorded in the data chart (Table 2), organized in a frequency table (Table 4), and displayed in the frequency histogram (Figure 4) depicts what is referred to as the *normal distribution curve* or the *normal curve*. A *bell-shaped normal distribution* is commonly observed for many phenomena and is the basis for certain kinds of statistical summarization and interpretation.

Figure 4: Frequency histogram of heights, in mm, of 48 Fast Plants at Day 10, grouped in class intervals of 10 mm.



Organizing and Displaying Data: Numerical Representation

r

Range

There are various ways of describing or summarizing the variation in heights of the 10 day old Fast Plants recorded in Table 2 and displayed in Figure 4. One way is in terms of *range* (r). Range extends from the shortest plant to the tallest plant and is defined as: "r = the difference between the largest and smallest numbers in a set of data." Here again the stem and leaf diagram is useful in identifying the range, r = 84 - 18 = 66 mm. The range identifies the upper and lower limits of a data set and is helpful in determining the limits of the x-axis on a graph. When measuring a population of Fast Plants over several days of growth it is interesting to observe what happens to the range of plant heights. Does the range stay the same, decrease or increase? Why?

Mean, Median and Mode: Measures of Center

Another way to summarize the variation represented in a set is in terms of *averages*. Continuing with our example, the average or *arithmetic mean* (*x*) is the sum of the measurements divided by the total number of measurements, n:

x

$x = \sum i\ {}^{x}i /_{n} = (x_1 + x_2 + x_n)(1/n)$

When phenotypes are distributed normally, the mean can be a useful way of summarizing or representing the set. The mean or average is a way of representing a data set using a single number. In our example the mean is:

$x = (x_1 + x_2 + x_n)(1/n) = (2212)(1/48) = 47.13$

Another way of identifying a central point in the data set is to identify the *median* (md), or middle value of a set. The median is the highest value divided by two, in our example:

md

md = 84/2 = 42

Notice that the median differs from the mean by approximately 5 mm (47 - 42 = 5).

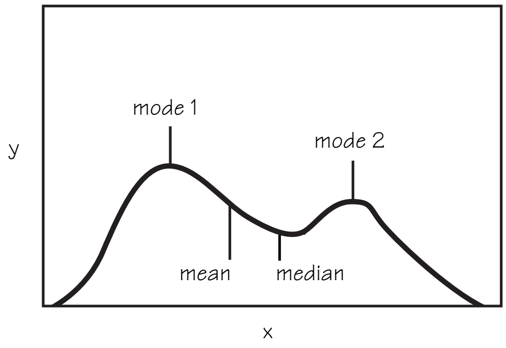
Yet another way of representing the data set with a single number is to use the *mode* (mo). The mode is the measurement with the highest frequency. Again, by scrutinizing each "leaf" of the stem and leaf diagram, you will observe that the number 45 mm appears three times. All others appear less frequently. This would be the mode for our example:

mo

mo = 45.

As is characteristic with normally distributed data, the mean, median and mode tend to be in proximity. With some natural phenomena which are not normally distributed there may be more than one mode, hence the terms *bimodal* and *trimodal* (Figure 5). In other distributions the mode may be widely separated from the mean and median (Figure 6).

Figure 5: Example of a bimodal frequency curve.



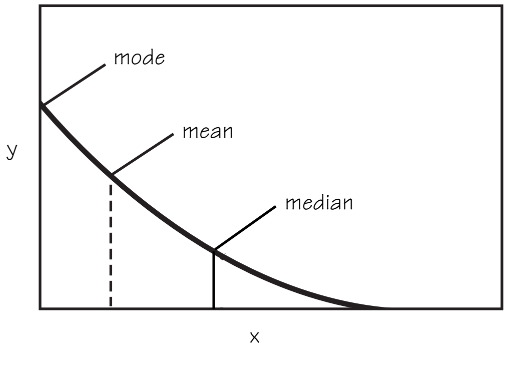


Figure 6: Example of a frequency curve with widely spread mode, mean and median.

Standard Deviation and Variance

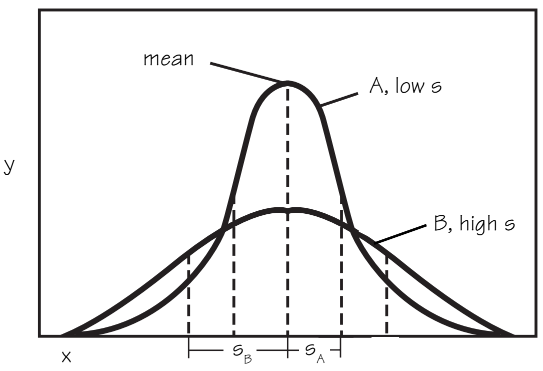
Although the mean is probably the most useful value in representing a set of measurements, the mean does not give an indication of the way in which the values of the set are distributed around the mean. In other words, how the shape of the bell in the normal frequency curve appears. The *standard deviation* (s) is a statistical notation that provides an indication of whether the measures of phenotype are widely distributed around the mean. When s is relatively high the normal curve is broad; when s is low the curve is relatively narrow, or tightly distributed around the mean (Figure 7).

The standard deviation is the square root of the *variance* (s^2), which is the sum of the squared deviations of each value from the mean *x* divided by n-1, the set size minus one.

$$s^2 = \frac{\sum i\,(x_1 - x)^2}{n-1}$$

Though the standard deviation is a tedious calculation to make with a pencil and paper, most hand calculators with a statistical capability will have functions that automatically provide the mean (*x*), variance (s^2) and standard deviation (s).

Figure 7: Example of normally distributed frequency curves depicting high A and low B standard deviations.



Statistical Summaries

For our data set of 48 height measures of 10 day old Fast Plants, the summarized statistical data are given in Table 5.

From the statistical summaries and graphical displays of the data sets students will be able to better understand the variation that will become evident in all aspects of the Fast Plants investigations. Throughout, the activities of measuring and recording, organizing and displaying are important. In order to communicate your observations, results and conclusions effectively with others, it is important that you compare the same sorts of data in the same terms of reference.

Table 5: Statistical summary of height data of 48 Fast Plants from Table 2.

number in set	n = 48
range	r = 66 mm
mean	*x* = 47.13 mm
standard deviation	s = 14.27 mm

Data Sheets and Tables

Data sheets or tables need to be organized so as to receive descriptive information in a logical and orderly manner that will minimize the likelihood of entry errors and that will aid in later summarization and analysis. For each activity, examples of student and class data sheets have been provided. With most of the experiments, the data sheets also contain columns for data summation and statistical analyses. Calculators with graphical capabilities may be useful to students in analyzing data.