1 The Nature of Data

Objectives

As you complete this lab exercise you will:

1. Calculate measures of central tendency for a data set and the different perspectives they provide on the same data set.
2. Understand the relationship between the mean and the variation for replicate values of a data set.
3. Examine the frequency distribution for three data sets.
4. Calculate the confidence intervals surrounding the mean of replicate values.

Ecologists collect data . . . and a lot of it. The data are extensive because (1) natural processes are complex; (2) ecological processes involve many variables; and (3) each variable can vary greatly. For example, plant growth is a complex process, the number of factors influencing growth is immense, and factors such as rainfall vary from day to day. To understand this kind of complexity and variation, ecologists analyze data to search for patterns and relationships among variables. In other words, they look at the nature of their data.

POPULATIONS AND SAMPLES OF POPULATIONS

Researchers gather data to describe and learn about large **populations**. Unfortunately, most populations are too big for us to measure a variable for every member of the population. For example, oak trees are too numerous for us to measure the length of *all* of their leaves. Instead, we **sample** the population of oak leaves. Sampling means that we take a relatively small number of measurements that represent the entire population. The characteristics we measure, such as leaf length, are **variables**, and the values for a variable are our **data**. Ecologists use these data to calculate **statistics** such as means, variances, etc. that describe our sample.

Statistics, such as a mean derived from samples, estimate variables of the entire population. For example, we measure the lengths of 20 sampled oak leaves and calculate a sample mean. This mean estimates the mean length of all the leaves in the population of oaks. We will never know the exact mean for the entire population so we use the sample mean ($\bar{x}$) to estimate the population mean (μ). Sample statistics are estimates of population statistics.

The major terms to quantitatively describe a set of data are:

population—the entire collection of possible measurements about which we wish to draw conclusions. All frogs in a pond, or all possible measurements of a forest's soil nitrate content, are examples of populations.

sample—a measurement(s) representing all possible measurements of a parameter of a population. It is a subset of all possible measurements in a population.

variable—a measurable characteristic of a biological entity. It may vary from one organism to the next, one environment to the next, or one moment to the next.

statistic—an estimate of a parameter based on a representative sample of a population. The mean of a set of values is a statistic.

DATA SETS

Data are usually organized into a **data set**, defined as a series of repeated measures of one or more variables. A variable might be the number of eggs in a robin's clutch, the concentration of nitrate in the soil, or the monthly rainfall on a prairie. You will quickly learn that the most noticeable attribute of a data set is its variation.

Procedure 1.1

Examine variation within a data set.

1. Examine the data set in table 1.1.
2. These data are measurements of the length of 20 leaves randomly selected from an oak tree. Notice that all leaves are not the same length.

3. Use the blanks provided in table 1.1 to rearrange and record the data from lowest to highest value.

Questions 1
Do the leaf lengths shown in table 1.1 appear to be simply random numbers? If not, what pattern or tendency do you detect? _____

_____

What factors might cause variation in leaf length for an oak tree? _____

_____

How would you sample leaves to test one of those influences on leaf length? _____

_____

TABLE 1.1

A SAMPLE DATA SET OF 20 REPLICATE MEASUREMENTS OF OAK-LEAF LENGTHS

Oak Leaf Lengths:

78	69	62	74	69	51	45	40	9	64
___	___	___	___	___	___	___	___	___	___
65	64	61	69	52	60	66	71	72	27
___	___	___	___	___	___	___	___	___	___

Measures of Central Tendency:

Mean = _____ Median = _____ Mode = _____

MEASURES OF CENTRAL TENDENCY

The most likely "pattern" revealed by examining the data set in table 1.1 is the **central tendency** of the values. They are not spread out randomly. They tend to be clustered around a central value somewhere in the 50s, 60s, or 70s rather than randomly scattered from 1 to 100. That shouldn't surprise you—oak leaves don't grow randomly. Their development has a pattern.

The three most common measures of central tendency are mean, median, and mode. The **mean** ($\bar{x}$) is the arithmetic average of a group of measurements. It is the sum of all the values ($\sum x_i$) divided by the number of values (N).

$$\text{mean} = \bar{x} = \sum x_i / N$$

The **median** is the middle value of a group of measurements that have been ranked from lowest to highest or highest to lowest. The **mode** is the value that appears most often in the data set.

The mean is the most common measure of central tendency, but the median and mode are sometimes useful because they are less sensitive to extreme values. To appreciate the differences among the measures of central tendency, complete Procedure 1.2.

Procedure 1.2

Examine measures of central tendency of an example data set.

1. Calculate and record the mean of the data set in table 1.1.

Questions 2
Are any of the leaf measurements the same as the mean?

_____

How many leaves were longer than the mean? _____

_____

How many leaves were shorter than the mean? _____

_____

Does the mean always describe the "typical" measurement? Why or why not? _____

_____

_____

2. Determine and record the median and mode of the data set in table 1.1.

Questions 3
Notice that in the sample, the mean differs from the median. What is responsible for this difference between the mean and the median? _____

_____

How would the mean change without the 9-mm leaf?

_____

_____

How would the median change without the 9-mm leaf?

_____

_____

How would the mode change without the 9-mm leaf?

_____

_____

VARIATION WITHIN A DATA SET

Measures of central tendency don't fully describe variation within a data set. Examine the two small sets listed in table 1.2.

Notice that the mean and the median are informative, but they do not describe variation in the data. The stream fish data set has considerably more variation even though the mean is the same as for the pond fish data set. Variation is best quantified by range, variance, standard deviation, and standard error.

TABLE 1.2

TWO SAMPLE DATA SETS WITH DIFFERENT LEVELS OF VARIATION

Number of Fish Collected in Five Replicate Seine-Net Samples:

Pond Fish Data Set: 25 28 30 32 35

mean ($\bar{x}$) = 30; median = 30

range _____ variance _____

standard deviation _____ standard error _____

Stream Fish Data Set: 10 20 20 25 75

mean ($\bar{x}$) = 30; median = 20

range _____ variance _____

standard deviation _____ standard error _____

The **range** is the difference between the smallest and the largest values of the data set—the wider the range, the greater the variation. The range of the pond fish data set is 25–35 = 10; the range of the stream fish data set is 10–75 = 65. The mean number of fish per sample is the same for both data sets, but the ranges indicate much more variation among the stream samples. Notice that the range can be artificially inflated by one or two extreme values, especially if only a few samples were taken.

Questions 4

Examine the values for the stream and pond fish data sets. In which data set is the variation most influenced by a single value? _____

_____

What is the best way to collect data and prevent a single sample from skewing the measures of central tendency and variation? _____

_____

Could two samples have the same mean but different ranges? Explain. _____

_____

Could two samples have the same range but different means? Explain. _____

_____

Variance measures *how* data values vary about the mean. Variance is much more informative than the range, and is easy to calculate (see the following example). First, calculate the mean. Second, calculate the deviation of each sample from the mean. Third, square each deviation. Then sum the deviations. The summation is called the *sum of squared deviations* (or *sum of squares*). Finally, divide the sum of squared deviations by the number of data points minus one to calculate the variance (S^2). The example uses the pond fish data set (table 1.2). Record the calculated values in table 1.2.

Number of Pond Fish Collected x_i	Mean $\bar{x}$	Deviation from the mean $(x_i - \bar{x})$	(Deviation)^2 $(x_i - \bar{x})^2$
25	30	−5	25
28	30	−2	4
30	30	0	0
32	30	2	4
35	30	5	25

Sum of squared deviations = 58
Variance = 14.5

The formulae for the sum of squared deviations and the variance are:

$$\text{sum of squared deviations} = \sum_{i=1}^{N}(x_i - \bar{x})^2 = 58.0$$

$$\text{variance} = \frac{\textit{sum of squared deviations}}{N - 1} = 58/4 = 14.5$$

where

N = total number of samples
$\bar{x}$ = the sample mean
x_i = measurement of an individual sample

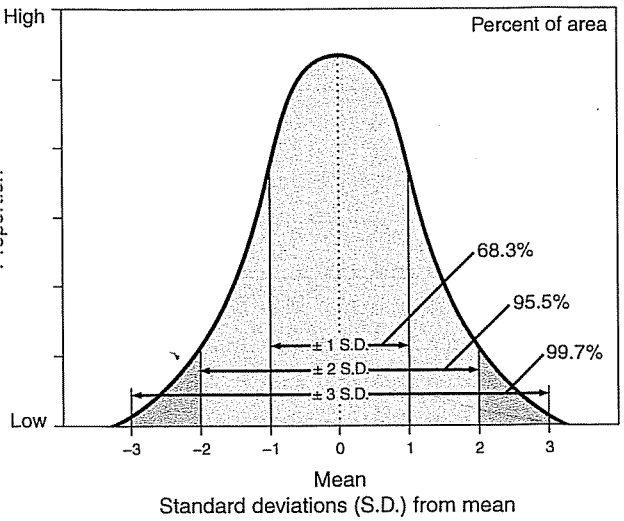
This formula for sum of squared deviations is really quite simple. The formula $(x_i - \bar{x})^2$ is the squared deviation from the mean for each value. The summation sign $(\sum_{i=1}^{N})$ means to sum all the squared deviations from the first one ($i = 1$) to the last one ($i = N$). The sum of squared deviations (58) divided by the number of samples minus one ($5 - 1 = 4$) equals the variance. The variance for these data is 58/4 = 14.5.

Variance is a good measure of the dispersion of values about the mean. A second, and more commonly cited, measure of variation is the standard deviation. The **standard deviation** (S) equals the square-root of the variance. For our example data set:

$$\text{standard deviation } (S) = \sqrt{\textit{variance}} = \sqrt{14.5} = 3.8$$

Standard deviation is usually reported with the mean in statements such as, "The mean number of fish per sample was 30 ± 3.8." The standard deviation helps us understand the spread of values in a data set. For normal distributions of measurements, the mean ± 1 SD includes 68% of the measurements. The mean ± 2 SD includes 95% of the measurements (fig. 1.1).

Another useful measure of the spread of data about a mean (i.e., variation) is the **standard error** ($S_{\bar{x}}$). This value measures the error from having a limited sample size (N). Clearly, a small sample (N) has more sampling error than does a large sample. The term "sampling error" doesn't mean that we have done something wrong. But we must document sampling error to use later in our calculations of confidence in the sample mean.



igure 1.1
lormal distribution graph.

tandard error is calculated as:

$$S_{\bar{x}} = S/\sqrt{N} = 1.70$$

/here

S = standard deviation
N = total number of samples

Procedure 1.3

Calculate four measures of variation.

1. Complete table 1.2 by calculating the range, variance, standard deviation, and standard error of the stream fish data set.

Question 5
The range indicates greater variation in the stream fish data et. Do the other three measures of variation also indicate greater variation? _____

FREQUENCY DISTRIBUTION

Some data sets are better understood if they are displayed in a graph called a **frequency distribution** (fig. 1.2). Frequency distributions summarize data at a glance and reveal subtleties that might not be apparent in calculations of central tendency. The abscissa (*x* axis) is plotted as *Data Class* and the ordinate (*y* axis) as *Frequency of Occurrence* in each Data Class. The raw data for the frequency distribution in figure 1.2 are below the graph. The shape of the curve reveals the nature of variation in the data set. A broad and flat curve reveals high variation. A narrow and high-peak curve reveals less variation.

Frequency distributions often have gradually tapering "tails" of frequencies toward each end of the curve (fig. 1.2). These tails produce a "bell-shaped" curve called a **normal distribution**. A normal distribution is common for ecological data—most values are near the mean and fewer values are at the extremes. Many variables are normally distributed, and many statistical tests used by ecologists assume that the variable is normally distributed.

Procedure 1.4

Examine a frequency distribution of heights.

1. Examine the data presented in figure 1.2 for the height of 119 female college students.

Questions 6
Are the heights distributed as you expected? How so?

_____

_____

Do you see evidence of central tendency in this data set?

_____

Do the data appear normally distributed? _____

_____

In what way do the data deviate from normality? _____

_____

_____

2. Examine the mean, median, and mode provided for the data set in figure 1.2.

Question 7
One criterion for a normal distribution is that the mean, median, and mode are equal. Are they equal for the data in figure 1.2? _____

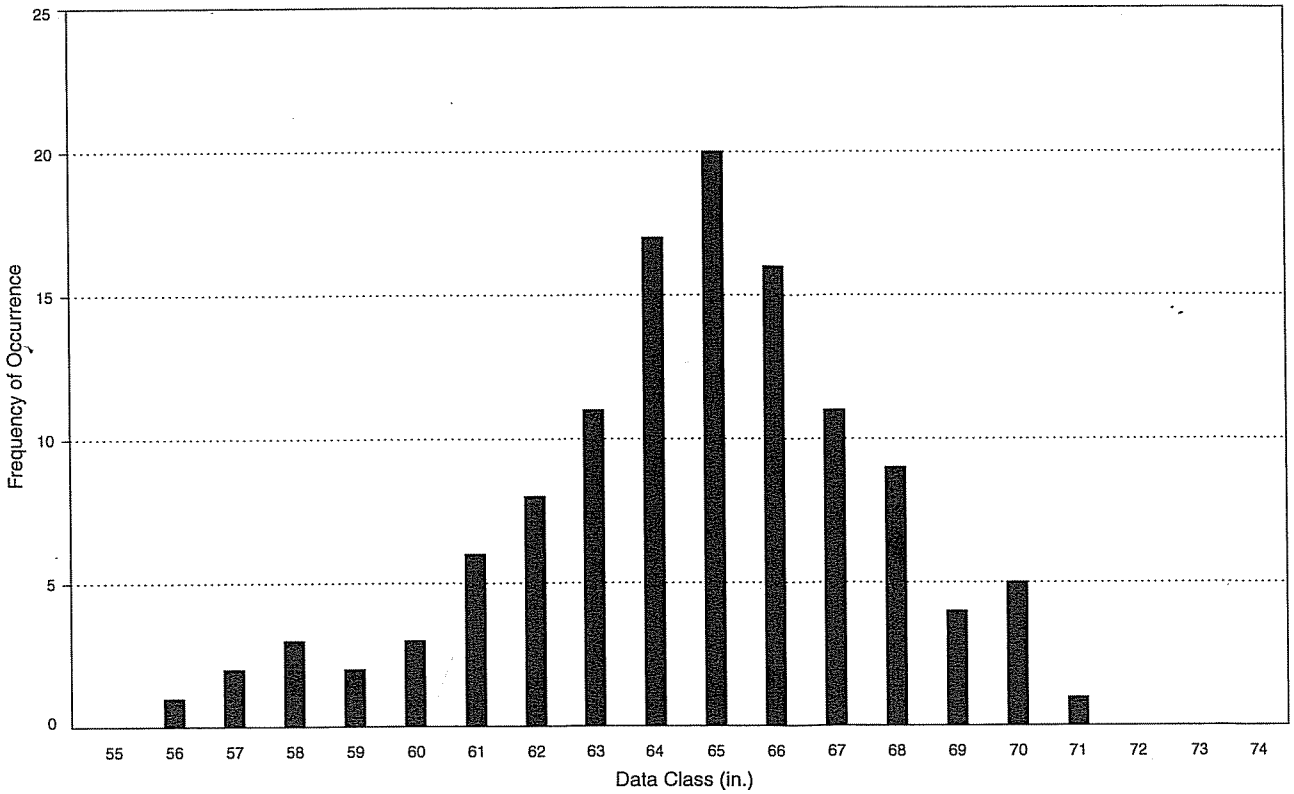
_____

3. Calculate and record the variance, standard deviation, and standard error for the data. The mean and sum of squared deviations are provided to speed your calculations.

Procedure 1.5

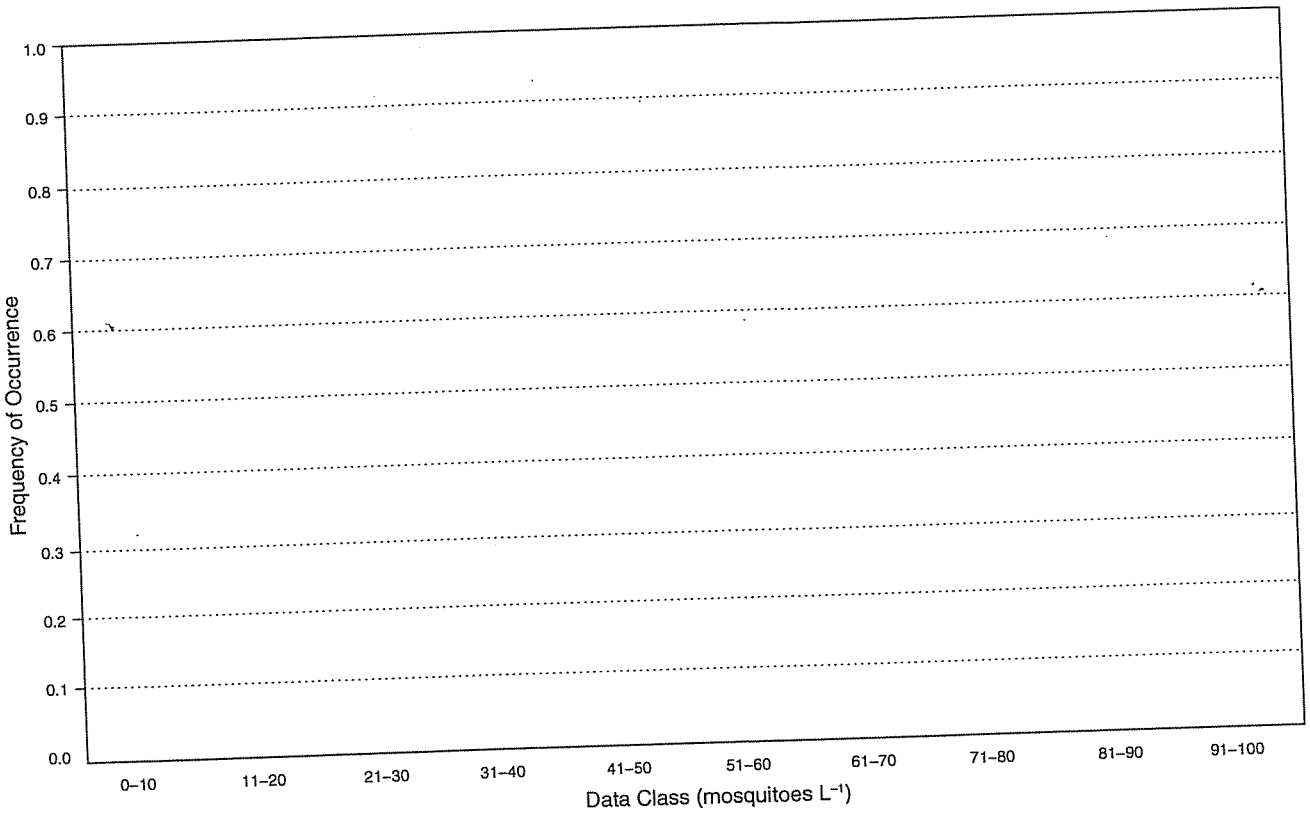
Calculate and graph the frequency distribution for a data set for mosquito larvae occurrence in tree-hole cavities.

1. Some mosquito species lay their eggs in tree-hole cavities that hold small volumes of water (< 1 L). Examine the data in figure 1.3 showing the densities of mosquito larvae (number per liter) found in 108 tree-hole cavities.
2. Plot a frequency distribution in the graph provided in figure 1.3.
3. Calculate and record in figure 1.3 the mean, median, mode, variance, standard deviation, and standard error for this data set.



Raw Data (height in inches)							Data Class	Frequency of Occurrence
							Height (in.)	Number of Students
56	62	64	65	66	68		55	0
57	62	64	65	66	68		56	1
57	62	64	65	66	68		57	2
58	62	64	65	66	68		58	3
58	62	64	65	66	68		59	2
58	63	64	65	66	68		60	3
59	63	64	65	66	68		61	6
59	63	64	65	66	68		62	8
60	63	64	65	66	68		63	11
60	63	64	65	67	69		64	17
60	63	64	65	67	69		65	20
61	63	64	65	67	69		66	16
61	63	64	65	67	69		67	11
61	63	65	66	67	70		68	9
61	63	65	66	67	70		69	4
61	63	65	66	67	70		70	5
61	64	65	66	67	70		71	1
62	64	65	66	67	70		72	0
62	64	65	66	67	71		73	0
62	64	65	66	67			74	0
N = 119	Mean = 64.6	Median = 65	Mode = 65	Sum of Squared Deviations = 1078.6		Variance =	Std. Dev. =	Std. Error =

Figure 1.2
A frequency distribution of 119 height measurements of college-aged women.



Raw Data (mosquitoes L^{-1})									Data Class	Frequency of Occurrence
									Mosquitoes L^{-1}	Number of Cavities
68	0	50	12	42	24	22	11		0–10	36
24	15	8	29	5	2	99	38		11–20	25
58	56	21	11	32	18	7	23		21–30	17
54	4	26	93	13	29	10	30		31–40	9
3	15	10	1	14	26	0	19		41–50	7
10	5	51	3	2	9	5	17		51–60	6
13	19	71	60	20	20	1	0		61–70	2
7	49	73	17	63	48	14	1		71–80	4
11	44	28	75	30	2	8	12		81–90	0
36	13	17	9	27	6	46	6		91–100	2
8	4	16	53	16	15	41				
5	22	35	9	7	34	37				
19	21	18	74	33	3	8				
39	4	25	5	2	28	31		Mean =	Median =	Mode =

Figure 1.3
Survey of the density (mosquito larvae L^{-1}) of the mosquito, *Aedes triseriatius*, occurring in tree-hole cavities (N = 108).

Questions 8

Is this variable normally distributed? _____

_____

How many values were greater than the mean? How many were less? _____

_____

What value best describes the central tendency of this data set? _____

_____

Procedure 1.6

Collect an original data set and calculate its measures of central tendency and variation.

1. Follow your instructor's directions to gather an original data set.
2. Record the raw data in figure 1.4.
3. Calculate and record in figure 1.4 the mean, median, mode, variance, standard deviation, and standard error for these data.

Questions 9

Is the variable in your original data set normally distributed?

_____

_____

How many values were greater than the mean? _____

How many were less? _____

What value best describes the central tendency of this data set? _____

_____

POPULATION MEANS AND CONFIDENCE INTERVALS

Population statistics can never be known exactly because we only measure samples of the population, not every member. Therefore, values such as the population mean (μ) must be estimated by the sample mean ($\bar{x}$). If variation is low, then we have high confidence in the sample mean as an estimator of the population mean.

A **confidence interval** is a range of values within which the true population mean occurs with a particular probability. Ecologists usually express their sample means with **95% level of confidence**, also called a **95% confidence interval**. For example, a sample mean ($\bar{x}$) may be 64.6 cm (see figure 1.2 for a sample of height measurements for college-aged women). After calculations, we are 95% confident that the *population* mean lies between 64.0 and 65.1.

The 95% confidence interval surrounding a *population* mean is calculated as:

$$\mu = \bar{x} \pm t_{0.05}(S_{\bar{x}})$$

where

μ = population mean
$\bar{x}$ = sample mean
$t_{0.05}$ = value from student's t table at the 95% confidence level
$S_{\bar{x}}$ = standard error

The value of $t_{0.05}$ is selected from a student's t table available in most statistics textbooks. The appropriate student's t value is determined by the degrees of freedom ($N - 1$) and the confidence level, which in this case is 95% (= 0.05 probability that the population mean occurs outside the range). For example, consult a student t table and you will find that the appropriate $t_{0.05}$ value for a sample of 30 ($N = 30$) and for a 95% confidence interval is 2.045.

To calculate the 95% confidence interval using the oak leaf measurements in table 1.1, first calculate the mean and standard error (see table 1.1).

$\bar{x}$ = mean = _____
$S_{\bar{x}}$ = standard error = _____
$N = 20$
DF = degrees of freedom = $(N - 1) = 19$

A student's t table shows that the critical value of $t_{0.05}$ is 2.09.

Therefore, the 95% confidence intervals surrounding the mean of 20 oak leaf samples is:

$$\mu = \bar{x} \pm 2.09(S_{\bar{x}}) = \_\_\_\_\_$$

With this confidence interval, we can state that there is a 95% probability that the population mean of the oak leaves is somewhere between _____ and _____.

Procedure 1.7

Calculate the confidence intervals for example and original data sets.

1. Calculate the 95% confidence interval for the population mean of stream fishes per seine haul from table 1.2.

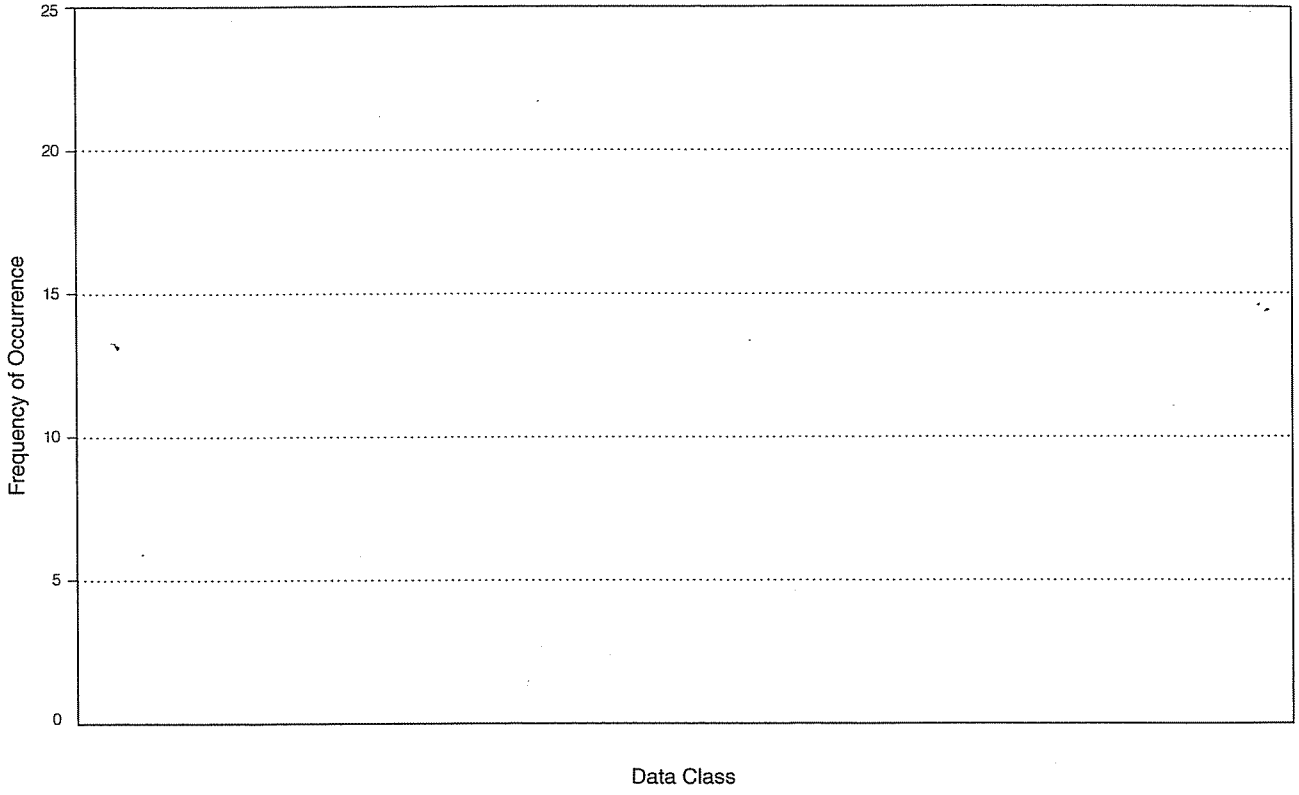
   $\mu = \bar{x} \pm t_{0.05}(S_{\bar{x}}) =$ _____ ± _____

2. Calculate the 95% confidence interval for the population mean of pond fishes per seine haul from table 1.2.

   $\mu = \bar{x} \pm t_{0.05}(S_{\bar{x}}) =$ _____ ± _____

3. Calculate the 95% confidence interval for the population mean height of female students whose sample is presented in figure 1.2.

   $\mu = \bar{x} \pm t_{0.05}(S_{\bar{x}}) =$ _____ ± _____



Raw Data									Data Class	Frequency of Occurrence
									Mean =	Median =	Mode =
									Variance =	Std. Dev. =	Std. Error =

Figure 1.4
Original data set for analysis.

Questions for Further Thought and Study

1. Replicate samples are central to good experimental design. How would the frequency distribution of heights in figure 1.2 differ if only five or six measurements were made?

2. Do you suspect that any biological variables have a perfectly normal distribution? Why or why not?

3. What is the relationship between variation in a data set and the width of the confidence intervals surrounding the estimate of the population mean?