

# **AMS 7: Statistical Methods For the Biological, Environmental and Health Sciences**

## **1: Introduction and Descriptive Methods**

**David Draper**

Department of Applied Mathematics and Statistics  
University of California, Santa Cruz

`draper@ams.ucsc.edu`

`http://www.ams.ucsc.edu/~draper`

© 2008 David Draper (all rights reserved)

# Outline

- **Introduction: populations and samples; parameters and statistics (estimates)**
- **Data types:** qualitative and quantitative variables; nominal and ordinal; discrete and continuous; interval and ratio; dichotomous
- **Descriptive methods** for a single variable
  - **Graphical:** histograms, bar charts
  - **Numerical:** measures of **center** (mean, median, mode) and **spread** (standard deviation, variance)
- Using the **normal distribution** descriptively

# 1.1 Introduction

**Statistics** is the **study of uncertainty**: how to **measure it**, and **how to make choices in the face of it**.

**Uncertainty** is a state of **incomplete** or **imperfect information** about something of interest to you, for example

the **percentage**  $\theta$  of the deer who lived on the UCSC campus as of 27 July 2008 who have **chronic wasting disease**.

I notice that I **don't know** the value of  $\theta$  **exactly**; I have the impression that  $\theta$  is **rather small**, since the deer on campus seem **relatively healthy**, but I have **substantial uncertainty** about its **precise value**.

I can **reduce my uncertainty** by **gathering data** on the disease status of campus deer; **how** should this data-gathering be done?

The **set**

$\mathcal{P} = \{\text{the deer who lived on the UCSC campus as of 27 July 2008}\}$

is an example of a **population**: a collection of **subjects** or **elements** (in this case, deer) of interest to me.

There is an **aspect** of each of these population subjects that I'm curious about: if I encountered one of these deer, the **question** I would ask is "**Does this deer have chronic wasting disease or not?**"

Things that can be **measured** on population subjects are called **variables**; in this case the variable of interest takes on only two values, **{yes, no}** (such variables are called **dichotomous** or **binary**).

# Populations and Samples; Parameters and Statistics (Estimates)

---

We'll see soon that a **handy** way to work with dichotomous variables is to assign **1** and **0** to their two possible values (hence the term **binary**); for example, with the variable (**chronic wasting disease or not**) the coding (**1 = yes, 0 = no**) is particularly useful.

A **numerical summary** of a population is called a **parameter**;  $\theta$  is an example of one possible parameter of interest about the population  $\mathcal{P}$  above (others might include the **average weight** of the deer who are **more than three months old**).

If I had enough **time** and **money** (and a way of ensuring that I could **find** all the deer and mark them **uniquely**, so that I didn't **double-count** any individual), in principle I could perform a **complete census** of the entire population, obtaining the **disease status** for each individual, and at the end of this census **I would no longer have any uncertainty** about the parameter  $\theta$ .

In practice people **rarely** have enough time and money to perform a **complete census** of a population  $\mathcal{P}$ ; instead it's natural to choose a **subset**  $\mathcal{S}$  of  $\mathcal{P}$  and evaluate the variable(s) of interest **only on the population subjects in the subset**.

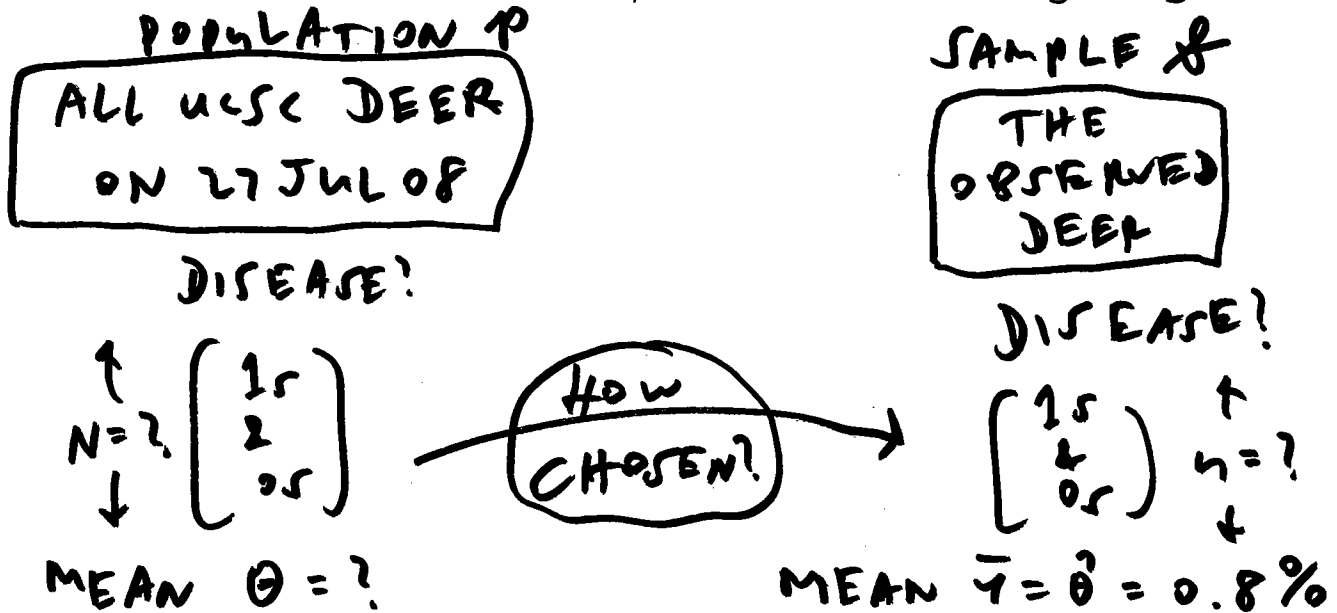
Such a subset is called a **sample** from the population  $\mathcal{P}$  — if the sample is chosen well, it seems like a good idea to use the data in the sample to make an **estimate** of (an educated guess at) the population parameter  $\theta$  of interest.

An estimate  $\hat{\theta}$  of a population parameter  $\theta$  is also sometimes called a **statistic**.

# Populations and Samples

Let's let  $N$  stand for the **number of subjects** in the **population  $\mathcal{P}$** , and  $n$  denote the **number of subjects** in the **sample  $\mathcal{S}$** .

Then both the **population** and the **sample** can be thought of as **data sets**, which can be further visualized as **rectangular tables\*** with **one row** for each **subject** and **one column** for each **variable**, as in the following diagram:



In this class we'll be looking a lot at diagrams like this one: such diagrams are the basis of both **probability models** and **statistical models**, both of which are crucial to the process of **quantifying uncertainty**.

To fill out a diagram like this I need to specify the following **ingredients**:

- In the **box** above the population data set  $\mathcal{P}$  I describe the subjects in  $\mathcal{P}$  by saying to myself “**There’s one row in the population for each \_\_\_\_\_**” and filling in the blank; for example, here there’s **one row for each deer living on the Santa Cruz campus on 27 July 2008**.

\***Math note:** the official name for such a rectangular table is a **matrix**.

# Random Sampling

- Above each **column** in the population data set I write the name of the **variable** summarized by that column (in this part of the class we'll typically work with only **one variable at a time**); here the variable of interest is the answer to the question (**chronic wasting disease?**).
- I identify the **number**  $N$  of subjects in the population if I know it (here I'm **not sure** how many deer there were on the UCSC campus in July 2008, so I just put a **question mark**).
- Then I do the same three things for the **sample** data set: in the box above  $S$  I describe the **subjects** in the **sample** (here I might just say "**the observed deer**"); above each **column** in the sample data set I write the name of the **variable** summarized by that column (this will be the **same** as in the population); and I identify the number  $n$  of **subjects** in the **sample** if I know it (here we haven't yet talked about **how large the sample should be**, so again I just put a **question mark**).

There's one crucial thing about this concept of **using the sample data to estimate a parameter** of interest in the population: I said above that this is a good idea "**if the sample is chosen well,**" and we need to figure out what this means.

Evidently, if the sample is to serve as a **good stand-in** for the rest of the population, the basic principle we want to follow is to try to make the **sample** and the unsample (the part of the population not chosen in the sample) **as similar as possible in all relevant ways.**

The simplest way to achieve this goal turns out to be to **draw the sample at random from the population** (so that all subsets have an **equal chance** of being chosen).

# SRS and IID Sampling

So the last (and perhaps **most crucial**) step in filling out the **diagram** above is as follows:

- Finally, in the **circle** above the **arrow** from the population to the sample I describe the **sampling method** (in this case, random).

To **literally take a random sample** of size  $n$  of deer from  $\mathcal{P}$ , you'd have to

(a) make a **list** of all the **population subjects** (deer), with **unique identifying tags**;

(b) choose  $n$  of these tags at random (using, for example, **pseudo-random numbers** generated by computer) **without replacement** (the sampling method **at random without replacement** is called simple random sampling (SRS), as opposed to **at random with replacement**, which is called independent identically distributed (IID\*) sampling); and

(c) measure the **variables** of interest on the sampled deer by finding the ones with the **chosen tags**.

In practice people would often instead use a **simpler** method that's not literally SRS (for example, if the deer were well distributed spatially, you could **partition** the UCSC campus into  $n$  non-overlapping and exhaustive spatial subsets and have  $n$  people each get data on the **first deer they encounter** in their subset on a given day) and then argue that their simpler method was **like** what you would get with SRS.

\***Textbook note:** IID is a term not mentioned in Triola & Triola.

## 1.2 Data Types

It's useful to have a **classification** of the various **types of data** that variables can keep track of (because some methods of analysis are definitely **not appropriate** for some data types).

**Example 1:** **Genetic phenotype.** **Eye color** in an animal you're studying may take on only **two values** (brown, blue) that have **no unique place on the number line** (earlier we called such variables **dichotomous** or **binary**); similarly, **hair color** might take on **four values** (predominately brown, black, red or white).

Variables like this are said to occur on a **nominal** scale of measurement (so dichotomous variables with values like {yes, no} are **special cases** of nominal variables).

**Example 2:** **Success in running a maze** might be recorded

1 (**very slow**), 2 (**slow**), 3 (**moderate**), 4 (**fast**),  
5 (**very fast**)

There are still **no unique places on the number line** for such values, but (unlike example 1) there's a **natural ordering** to these values.

Variables like this are said to occur on an **ordinal** scale.

Some other names for both **nominal** and **ordinal** variables are **qualitative** and **categorical**; **ordinal** variables are also sometimes called **ordered categorical** variables.

## Data Types (continued)

**Example 3:** **Size of a plant.** Two measures of the **size** of a plant (which, in turn, is a measure of its **competitiveness**) would include its **height** (in centimeters (cm)) and the **number of leaves** it has.

Unlike the situations in Examples 1 and 2, the values taken on by these variables **do have unique places on the number line**, and in fact there are two important characteristics of the numerical values of these variables:

- (a) there's a **constant size interval** between any adjacent units on the measurement scale, so that the concept of **1 unit** means the same thing anywhere on the scale (for example, plants *A*, *B*, *C* and *D* are (respectively) 14, 15, 62, and 63 cm high; the **amount** by which *B* is **taller** than *A* is the same as the **amount** by which *D* is **taller** than *C*); and
- (b) there's a **true zero** on the measurement scale with a **direct physical meaning** (in this case, absence of height, or no leaves at all) — this allows us to make meaningful statements about **ratios** (for example, plant *C* is  $\frac{62\text{cm}}{15\text{cm}} \doteq 4.1$  times taller than plant *B*).

Variables like this are said to occur on a **ratio** scale.

**Example 4:** **Growing temperature** at which a plant produces the most buds. Temperature (measured either in  $^{\circ}\text{C}$  or  $^{\circ}\text{F}$ ) does have a **constant size interval** but **lacks a true zero**, so (contrary to statements you see in the newspaper or on TV) when it's  $80^{\circ}\text{F}$  outside you can't correctly say that it's **twice as hot** as when it's  $40^{\circ}\text{F}$ .

Variables like this are said to occur on an **interval** scale.

# Data Types (continued)

Some other names for both **ratio** and **interval** variables are **quantitative** and **numerical**.

**One last distinction:** plant **height** and **number of leaves** are different in that with plant height, **conceptually** (with finer and finer measuring instruments) there are **no possible gaps between the possible values**, whereas with number of leaves, **distinct structural gaps** exist (it doesn't make sense to talk about  $4\frac{1}{2}$  leaves).

Quantitative variables with **gaps** between the possible values are called **discrete**; quantitative variables with **no conceptual gaps** between the possible values are called **continuous**.

**Why these distinctions matter.** Suppose I choose to **code** the **age** of some animals I'm observing in the following way, when **storing** the values of this variable in a **computer**:

**less than 1 year old = 0, between 1 and 2 years old = 1, between 2 and 3 years old = 2, between 3 and 4 years old = 3, ...**

Suppose further that I choose to **code** the hair color of these animals in the following way:

**brown = 0, black = 1, red = 2, white = 3**

Here's the data set I get (written, to save space, in a **transposed** fashion in relation to the convention on page 5 above: here the **rows** are the **variables** and the **columns** are the **subjects** (animals)):

							Mean
Animal Identifier	45	333	167	2	...	501	243.9
Age	2	0	1	1	...	3	1.7
Hair Color	0	2	3	2	...	1	1.2

## 1.3 Descriptive Methods

As we'll soon discuss, it's sometimes both **useful** and **meaningful** to **summarize** a variable by taking its **mean** (just add 'em up and divide by how many there are); the computer has done this for us in the table above in the **final column**.

The problem is, of course, that the **mean** is **meaningful** only for the **age** variable (because it's **quantitative** [ratio, discrete]; the other two variables are **qualitative** [nominal]).

**The point:** The **right way** to **analyze** a variable often depends on the **scale** on which it's measured.

---

### **1.3.1 Graphical descriptive methods.** Example:

**butterfly wing lengths.** Zar (1999) gives data from a sample of  $n = 24$  immature monarch butterflies, in which the **variable** of interest (we might call it  $y$ ; T&T would call it  $x$ ) is **wing length** (in cm):

4.4 3.6 4.1 3.3 3.5 3.8 4.5 4.3 4.3 4.0 4.1 3.6  
4.0 4.0 3.8 3.8 3.9 4.2 4.2 4.1 3.7 3.9 4.0 3.9

(This is just **shorthand** for a **data set** with  $n = 24$  **rows** (**subjects = butterflies**) and 1 **column** (**variable = wing length**), written in this manner to save space.)

How might we **summarize** this variable in a way that would allow us to **see patterns** (**graphical summaries**) and to capture **most of the information it contains** in **fewer than 24 numbers** (**numerical summaries**)?

# Raw Frequency Distribution

As long as the **order** in which the data values were listed above is **not relevant**, the first step would be to **sort** the data from **smallest** to **largest**:

3.3 3.5 3.6 3.6 3.7 3.8 3.8 3.8 3.9 3.9 3.9 4.0  
4.0 4.0 4.0 4.1 4.1 4.1 4.2 4.2 4.3 4.3 4.4 4.5

Now we can see that there are a number of **duplicate values** (caused by **rounding** the wing length measurement to the nearest cm).

This suggests a **further summary** in which we keep track of the **values** of the variable and the **raw frequencies** (the numbers of times those values are attained):

Value	Frequency
3.3	1
3.4	0
3.5	1
3.6	2
3.7	1
3.8	3
3.9	3
4.0	4
4.1	3
4.2	2
4.3	2
4.4	1
4.5	1
Total	$n = 24$

This is called a **raw frequency distribution** (or **frequency table**) for the variable  $y$  (sometimes people just refer to the **distribution** of  $y$ , or ask “How is  $y$  **distributed?**”).

# Raw Frequency Histogram

The **table** on the previous page is not as easy to **absorb** as it would be if we could display it **graphically**.

Since it has **two columns** or **dimensions**, it's natural to make a plot in which one dimension (**horizontal**, say) is the **values** the variable takes on and the other (**vertical**, say) is the **raw frequencies** — the resulting graph is a (raw frequency) **histogram** of the variable  $y$ :

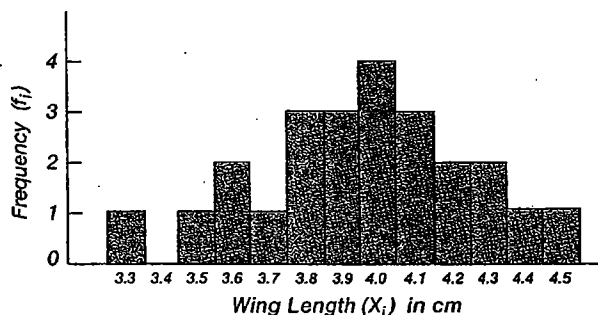


Figure 3.1 A histogram of the data in Example 3.2. The mean (3.96 cm) is the center of gravity of the histogram, and the median (3.975 cm) divides the histogram into two equal areas.

A **histogram** is a special case of a **bar graph**: a plot with **locations** identified along the **horizontal** axis corresponding to **values** a variable takes on and **bars** over those locations with **heights** give by the (raw) **frequencies** of those values.

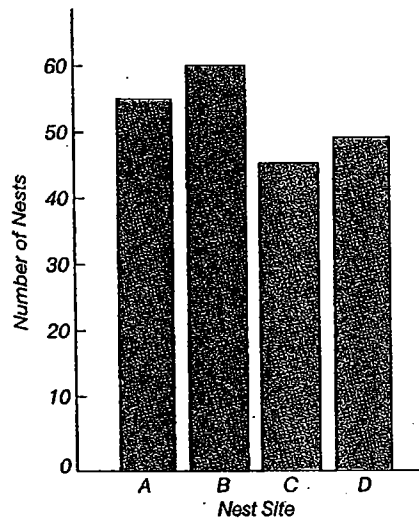
A **bar graph** can be drawn as a summary of **any qualitative** (**nominal** or **ordinal**) **variable**; there is no unique place called “yes” or “red” on the number line, but you can just **invent arbitrary horizontal locations** and make a useful plot anyway.

Strictly speaking, what makes a **histogram** a histogram is that the variable in question is **quantitative** (so that the values do have unique locations on the number line) — histograms can be made for either **discrete** or **continuous** variables.

# More Graphical Examples

**EXAMPLE 1.1** The location of sparrow nests. A frequency table of nominal data.

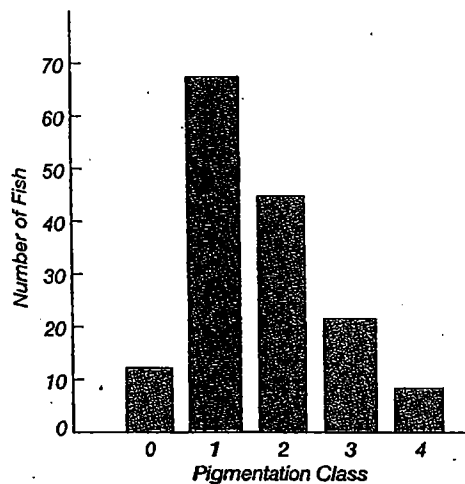
<i>Nest site</i>	<i>Number of nests observed</i>
A. Vines	56
B. Building eaves	60
C. Low tree branches	46
D. Tree and building cavities	49



**Figure 1.1** A bar graph of the sparrow nest data of Example 1.1. An example of a bar graph for nominal data.

**EXAMPLE 1.2** Numbers of sunfish, tabulated according to amount of black pigmentation. A frequency table of ordinal data.

<i>Pigmentation class</i>	<i>Amount of pigmentation</i>	<i>Number of fish</i>
0	No black pigmentation	13
1	Faintly speckled	68
2	Moderately speckled	44
3	Heavily speckled	21
4	Solid black pigmentation	8



**Figure 1.3** A bar graph of the sunfish pigmentation data of Example 1.2. An example of a bar graph for ordinal data.

# Graphical Examples (continued)

**EXAMPLE 1.3** Frequency of occurrence of various litter sizes in foxes. A frequency table of discrete, ratio-scale data.

<i>Litter size</i>	<i>Frequency</i>
3	10
4	27
5	22
6	4
7	1

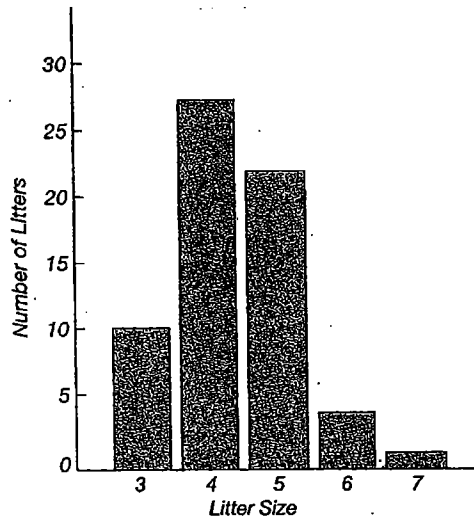


Figure 1.4 A bar graph of the fox litter data of Example 1.3. An example of a bar graph for discrete, ratio-scale data.

**EXAMPLE 1.4a** Number of aphids observed per clover plant. A frequency table of discrete, ratio-scale data.

<i>Number of aphids on a plant</i>	<i>Number of plants observed</i>	<i>Number of aphids on a plant</i>	<i>Number of plants observed</i>
0	3	20	17
1	1	21	18
2	1	22	23
3	1	23	17
4	2	24	19
5	3	25	18
6	5	26	19
7	7	27	21
8	8	28	18
9	11	29	13
10	10	30	10
11	11	31	14
12	13	32	9
13	12	33	10
14	16	34	8
15	13	35	5
16	14	36	4
17	16	37	1
18	15	38	2
19	14	39	1
		40	0
		41	1

Total number of observations = 424

# Graphical Examples (continued)

**EXAMPLE 1.4b** Number of aphids observed per clover plant. A frequency table grouping the discrete, ratio-scale data of Example 1.4a.

<i>Number of aphids on a plant</i>	<i>Number of plants observed</i>
0-3	6
4-7	17
8-11	40
12-15	54
16-19	59
20-23	75
24-27	77
28-31	55
32-35	32
36-39	8
40-43	1

Total number of observations = 424

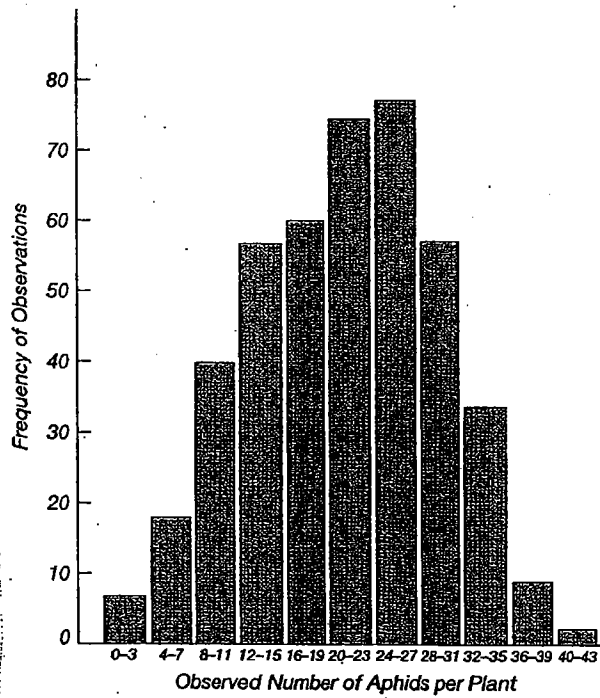


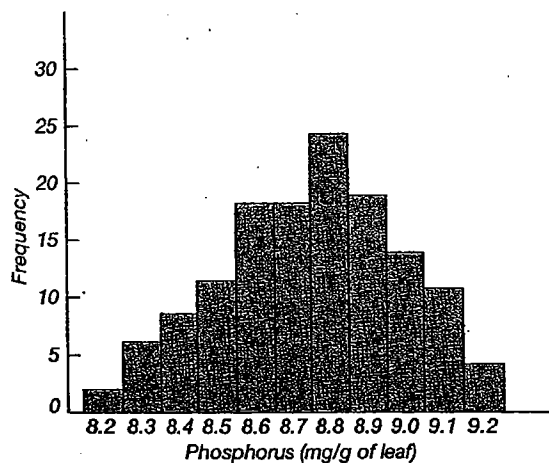
Figure 1.5 A bar graph of the aphid data of Example 1.4b. An example of a bar graph for grouped discrete, ratio-scale data.

# Graphical Examples (continued)

**EXAMPLE 1.5** Determinations of the amount of phosphorus in leaves. A frequency table of continuous data.

Phosphorus (i.e., number of mg/g of leaf)	Frequency (i.e., number of determinations)	Cumulative frequency	
		Starting with low values	Starting with high values
8.15–8.25	2	2	130
8.25–8.35	6	8	128
8.35–8.45	8	16	122
8.45–8.55	11	27	114
8.55–8.65	17	44	103
8.65–8.75	17	61	86
8.75–8.85	24	85	69
8.85–8.95	18	103	45
8.95–9.05	13	116	27
9.05–9.15	10	126	14
9.15–9.25	4	130	4

Total frequency = 130



**Figure 1.6a** A histogram of the leaf phosphorus data of Example 1.5. An example of a histogram for continuous data.

## 1.3.2 Numerical Descriptive Methods

In addition to **summarizing** a variable **graphically** to look for **patterns**, it's also useful to **summarize it numerically**, to capture **most of the information it contains** in fewer than  $n$  numbers.

People have found **two main types** of **numerical** summary useful: measures of **center** (or **central tendency**), and measures of **spread** (or **dispersion**, or **variability**).

**Measures of center.** The three most useful are the **mean**, the **median** (and other **quantiles**, or **percentiles**), and the **mode**.

The **mean** is, I'm sure, an old familiar object: with a variable like **butterfly wing length** above ( $n = 24$ )

$$\begin{array}{cccccc} 4.4, & 3.6, & 4.1, & \dots, & 4.0, & 3.9 \\ y_1, & y_2, & y_3, & \dots, & y_{23}, & y_{24} \\ y_1, & y_2, & y_3, & \dots, & y_{n-1}, & y_n \end{array}$$

to get the **mean** (for a variable  $y$ , let's call the mean  $\bar{y}$ ) I just **add up** all the data values and **divide** by **how many there are**:

$$\bar{y} = \frac{4.4 + 3.6 + 4.1 + \dots + 4.0 + 3.9}{24} = \frac{95}{24} \doteq 3.96 \text{ cm.} \quad (1)$$

**Symbolically**, using the idea of **summation notation**, this can be written more **succinctly** as

$$\begin{aligned} \bar{y} &= \frac{y_1 + y_2 + y_3 + \dots + y_{n-1} + y_n}{n} \\ &= \frac{1}{n} (y_1 + y_2 + \dots + y_n) = \frac{1}{n} \sum_{i=1}^n y_i. \end{aligned} \quad (2)$$

## Measures of Center

It turns out that the **mean** has the **graphical interpretation** of the **center of gravity** of the data: if you visualize the **histogram** of the variable  $y$  as made of **bricks** that are sitting on a **number line** made of **plywood**, which in turn is put **on top of a saw-horse**, the mean is **the place where the histogram would exactly balance**.

The **median**  $\tilde{y}$  of a column of numbers  $(y_1, \dots, y_n)$  is defined to be the **middle** value in the list in which  $(y_1, \dots, y_n)$  has been **sorted** from smallest to largest — if  $n$  is an **odd** number this is **uniquely defined**; if  $n$  is **even** there is **no single middle number** and people define the median to be the **mean** of the two middle values.

By examining the **sorted list** of the **butterfly wing lengths** on page 12 above ( $n = 24$ ), you can see that the **middle two numbers** are both **4.0**, so the **median**  $\tilde{y}$  of that variable is **4.0cm** (not very far from the **mean** in this case).

Since **histograms** graphically display the **frequency distribution** of a variable, the **graphical interpretation** of the **median** must be that it's the place where **half of the data is to the left of that place** and **half to the right** in the histogram.

The **median** is a **special case** of the general idea of finding places in the distribution where a particular **percentage** of the data is **to the left** of that place — these are called **quantiles** or **percentiles**.

By definition the **median** is the **50th percentile**, but it's also useful sometimes to look at **other percentiles** — for example, the **25th percentile** (also called the **first quartile**) is the place where  $\frac{1}{4}$  of the data is to the **left** of that place, and similarly the **75th percentile** (the **third quartile**) is the place where  $\frac{1}{4}$  of the data is to the **right** of that place.

# Mean, Median, Mode

The **mode** is defined to be the point of highest frequency in the distribution of a variable, so by definition its **graphical interpretation** is just the **highest point in the histogram** (Note: there are **many different possible histograms** for the same set of data [as a function of **how wide you choose the bars to be and where you decide to start the first bar**], and the mode is **sensitive to these choices**).

Some distributions have **more than one mode** — these are called **multimodal** (a common special case is **two modes**, which defines a **bimodal** distribution), and distributions with only one mode are therefore **unimodal**; usually multimodality means that there are **two or more subgroups** in the sample that should perhaps be studied **separately**.

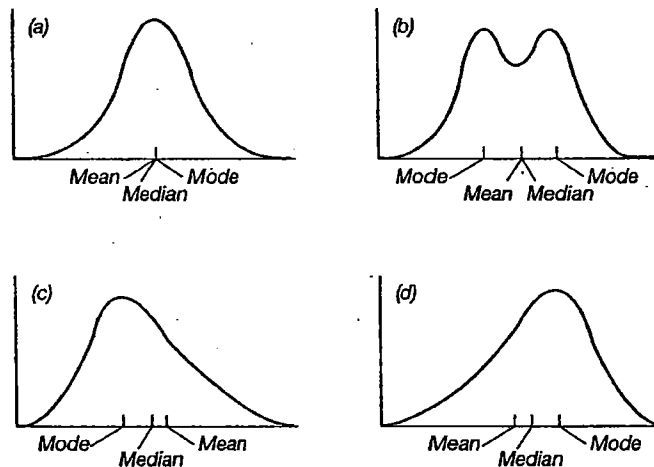


Figure 3.2 Frequency distributions showing measures of central tendency. Values of the variable are along the abscissa (horizontal axis), and the frequencies are along the ordinate (vertical axis). Distributions (a) and (b) are symmetrical, (c) is positively skewed, and (d) is negatively skewed. Distributions (a), (c), and (d) are unimodal, and distribution (b) is bimodal. In a unimodal asymmetric distribution, the median lies about one-third the distance between the mean and the mode.\*

Distributions that are composed of **mirror images to the left and right of a central folding point** are called **symmetric**; combining two of these terms, a **fairly common distributional shape is symmetric unimodal** (like Figure (a) above) — for such distributions the mean, median and mode all **coincide with the point of symmetry**, which makes choosing a measure of center for them **easy**.