

# 29 Statistical Analysis and Representation of Data

One of the most remarkable developments in recent years has been a tremendous increase in the availability of calculators, computers, terminals, and other devices as aids to analyses of numerical data. Still, many persons avoid learning methods for handling and interpreting numbers because the terminology looks strange or they presume that they do not have an adequate mathematical background. In most instances, these fears are unwarranted because one can learn many techniques for analyses quickly and easily.

The purpose of this chapter is to provide an overview of **statistics**, the scientific analysis of numerical data. Learning how to collect, organize, analyze, and interpret numerical data is vitally important if one is to be an effective scientist and researcher. As biologists, we are increasingly involved with the collection and analysis of large quantities of data on environmental, morphological, and physiological variables. Statistics provides techniques for objectively identifying sources of variation, for comparing data sets, and for establishing confidence limits for various estimates.

Students are strongly urged to enroll in a statistics course or courses sometime during their tenure at college. There are many statistics books on the market but the following deserve special comment. Steel and Torrie (1960), while technically good, is oriented toward applied and experimental agriculture. Three books that directly concern biologists include the introductory account by Simpson *et al.*, (1960) and the more extensive treatments by Sokal and Rohlf (1969) and Zar (1974).

## Data and Sampling Units

Like most scientific disciplines, statistics does have special terminology to aid standardization in publications and clarity in communication between workers in the field. In many instances, these words are familiar to you but may be defined more rigorously.

**Data** (singular **datum**) are observations or measurements taken on a sampling unit. These data or observations are taken from objects or entities termed **sampling units**. Sampling units could be individuals, leaves, soil samples, or other objects.

A **variable** or **character** is the property being measured or observed on the sampling unit. For example, a mammalogist interested in the variable "body weight in grams" weighed ten rats (the sampling units) producing the following set of data: 105, 110, 90, 85, 120, 115, 110, 100, 90, 95.

Many people use the term statistics to refer to the observations or data. Statisticians restrict the usage of this term to apply only to the discipline *or* a computed quantity such as the mean (arithmetic average).

29-A. In the following list, circle all words or statements representing variables and underline all those representing sampling units:

Rat 1421	Weight	Female 201
Row 114	Temperature	Glucose Level
Ear Length	Sex	Quadrat 1102

## Kinds of Variables

Sokal and Rohlf (1969:11-12) stated that variables may be classified into those resulting from measurements (continuous or discrete), counts based on attributes, and those based on relative rankings. **Discontinuous variables** (meristic or discrete) have fixed values with no intermediate values possible, e.g., 1, 2, or 3 toes present, but never 2.4 toes present. **Continuous variables** can assume, theoretically, an infinite number of values between any two fixed points, e.g., between points 1 and 2 there exist 1.1, 1.112, 1.0113, etc., depending on the accuracy of the measuring instrument and the patience of the data recorder. An **attribute** is

a qualitative measure or property, e.g., black, brown, or blue eyes; male or female; dead or alive. When attributes are combined with frequencies into tables of numbers they are referred to as **enumeration data**. **Ranked variables** can be placed on an ordinal scale but differences in ranks remain relative rather than absolute. For example, the widths of body stripes on a skunk might be ranked from 1 to 5, but a stripe in a class "1" may not be five times narrower than a class "5" stripe. The letter grades that you receive in a course are ranked values derived from measurement values (e.g., percentages).

29-B Using the letters in parentheses, classify the following data as continuous (C), discrete (D), or rank (R) variables:

- |                 |                    |
|-----------------|--------------------|
| ___ 90.1 °C     | ___ Code 3         |
| ___ density "3" | ___ 23.02 grams    |
| ___ 14 (stems)  | ___ 6 (bites)      |
| ___ I Class     | ___ male 6 Largest |
| ___ 114.22 mm   | ___ 242 Kilos      |
| ___ 0.01 grams  | ___ Incisors 3/3   |

### Accuracy and Precision

The researcher must strive for accuracy and precision in measurements. **Accuracy** is the nearness of a measured value to its true value. **Precision** is the nearness of values of successive measurements of the same character of the same specimen. Both accuracy and precision depend on the skill of the person making the measurement. Accuracy, however, can also be lost by a measuring device that is incapable of measuring to the researcher's goal of accuracy; e.g., vernier calipers accurate to 0.1 mm cannot produce results accurate to 0.01 mm. But an improperly calibrated balance could produce precise successive measurements even though these measurements may be inaccurate (e.g., the balance may show that animals weigh 0.5 grams less than their true weights).

29-C. To the right of each numerical observation listed, record the range of accuracy implied by the value. For example, for the value 0.1, the range of accuracy implied is 0.050-0.149.

5	1000
0.01	100.1
100	10
12.3	0.001
10.5	2.06

### Populations, Samples, Sampling

In statistics a **population** consists of all values of a particular variable within a specified space or time. For practical and logistical reasons it is usually impossible to record observations on all individuals in the population. Instead, the biologist works with randomly selected individuals from the population. A set of such individual observations is called a **sample**. The biologist then makes inferences about the entire population based on the available samples. Thus, **sample statistics** (usually symbolized by Latin letters) are estimators of **population parameters** (usually symbolized by Greek letters).

A population or **universe** may refer to all possible individuals or objects (units) or may be restricted by a space or time limitation. For example, a population could consist of all human inhabitants of the world or be all males 15 years of age in New York City. It would be extremely difficult to devise a sampling scheme to estimate the population parameters of the entire world population. Most scientists and statisticians make estimates of populations that have been more narrowly defined.

Many statistical procedures assume that samples will be obtained in a **random** fashion, such that each member of a population has an equal and independent chance of being selected. Randomness in sampling might be achieved by assigning all individuals in a population a number and then drawing numbers (representing the individuals) out of a hat. Better still, a **Table of Random Digits** (Table 29-1 in this text; Table 0 in Rohlf and Sokal, 1969; Table D.45 in Zar, 1974) should be consulted for selecting samples. Suppose that we wish to sample plants in 15 of 64 possible quadrants in an 8 × 8 grid. Since each quadrat can be represented by an X and Y coordinate, then pairs of numbers can be selected, in order, from the table of random numbers. Prior to consulting the table, determine the direction that you will follow in the table (e.g., across page, top to bottom, etc.). Then select a starting point in the table, proceed in the predetermined direction and write down pairs of unique numbers between 1 and 8. Fifteen of these pairs will then indicate the quadrants that should be sampled.

The number of sampling units or the **sample size** (N) that is used to estimate the population parameters is a matter of prime importance in any statistical analysis. In many instances, the biologist has no control over the sample size available for study. This frequently occurs in taxonomic studies when the scientist must work with existing collections of specimens (see Simpson et al., 1960:102-107). Ideally, the investigator should work with a sample size sufficient for

**Table 29-1**

Two thousand and five hundred random digits. (Brower and Zar (1974:4).

72965	92280	85318	98478	05200	26558	04697	63195	41679	24133
25182	09959	91375	97794	50193	25930	47938	95633	22271	15628
78812	39100	81576	84683	47466	04204	86339	31919	83404	48293
87264	75327	92529	25409	52589	20914	58768	46171	32657	89750
21571	57796	67813	88705	52576	51712	12407	00644	81748	04204
98532	11191	63198	79306	04193	00859	83906	30625	67175	37774
38981	76006	33931	22225	00014	37716	67499	90402	08962	88602
11305	19964	22932	62300	64508	32996	05699	06536	22619	89725
96753	89989	67869	65743	65353	55722	91650	77833	05353	05950
28316	27206	32507	96140	83430	75357	57822	75247	93486	20481
24390	09214	19493	94975	71393	54675	51712	00581	11187	73464
23995	32726	41075	32118	63946	62464	60599	81670	73097	78553
41920	60706	55864	70343	61238	06810	53263	07815	56588	29384
78281	15410	26154	70445	27828	38282	29051	13433	84405	82969
92910	17017	92704	25210	63833	04909	02571	58402	62649	86771
29265	89779	95437	51929	75534	70858	54623	99661	87146	16775
60422	65242	57037	95091	25582	76743	95890	09033	08368	62677
42748	43783	94238	97764	64110	68935	21057	14994	94235	53722
39611	11320	52913	20490	84147	59510	45967	93742	71756	09298
74011	92403	54878	91689	20402	20287	05402	16617	86101	28192
49056	17282	52320	73306	91759	85329	88229	62615	25802	28655
06572	13935	69948	12322	84900	85760	67583	36717	75897	39169
32726	45220	41600	61236	55701	08181	26259	49841	88968	83197
13800	03061	28494	09432	95359	92550	11251	76533	51923	34450
09838	95794	39792	06406	81584	49541	20520	91941	43448	91692
86499	23583	61444	72616	78692	50822	10283	23499	17883	21908
19618	23145	32406	91793	50163	72615	61939	18183	20368	51482
04145	26409	44737	98157	14158	94981	66518	84956	65372	00578
44083	35657	49215	93131	41815	34454	46347	02783	27988	86461
13883	40605	76333	56473	27866	16074	00939	05149	14090	70080
08697	34971	19204	70701	56065	23839	45794	62036	07594	36604
86447	56887	61107	63246	88350	51579	95387	03708	16441	64848
37914	39110	60363	95348	96498	17447	18058	36020	57301	50492
08771	12569	06379	51277	88233	45879	89353	82759	16691	20680
65529	84747	61160	19575	98709	23055	37992	82397	62884	63738
53783	03060	00563	21869	41559	85468	37401	81331	62733	10999
40881	01466	66439	92600	95878	43878	76006	93166	20603	76173
81424	81842	17993	63784	39351	41580	89006	47888	92753	45323
47362	92940	89774	05283	49461	21521	72572	37403	90574	22562
79898	44180	49706	58783	47012	90892	89032	56904	56473	38246
98433	36491	48288	53653	77220	82969	70063	58551	20025	83414
79849	94549	69691	11789	43233	46831	08737	25992	11296	69195
26004	14598	80743	25043	45287	35345	46914	71487	10345	48236
46218	40835	82386	91946	14266	77484	02759	92164	77842	21600
49618	10730	47690	44746	09566	36769	39108	47001	62935	10227
66259	25266	88651	56018	68181	45119	91387	37257	83610	53138
65170	81485	14727	22898	63815	17317	68293	06449	91890	49994
82679	72969	04512	11079	95969	87389	46263	96780	78124	04120
37900	90316	47434	60701	89649	51773	26139	39231	72264	17654
27111	31679	71539	61375	58691	20215	91170	44290	91396	90173

the chosen statistical technique and aims of the investigator. Sample sizes numbering fewer than 6 or 8 individuals are rarely useful for statistical purposes since the investigator can place little confidence in the statistics derived from such small numbers. Most statistics textbooks give guidelines for determining adequate sample sizes and discussions of the value of larger sample sizes. A procedure for estimating adequate sample sizes is described in Sokal and Rohlf (1969:246-249).

29-D. Examine a table of random digits (e.g., Table 29-1). Select numbers to take 20 random samples from each of the following situations:

1. Location of vegetation plots alongside 100-meter transect divided into 100 one-meter segments
2. Locations for taking ten soil samples in 100 x 100 meter grid divided into 100 10 x 10 meter quadrats

What was similar about the methods used to select the locations? What was different?

### Frequency Distributions

Inspection of the distribution of a variable is an important step in the analysis of data since it frequently guides our decisions on what statistical analyses to utilize. A **frequency distribution** is prepared by listing all observed values and then noting how many times each value is observed. For example, we autopsy a sample of 29 female rats and wish to examine the frequencies of various numbers of embryos as an indication of potential litter size. A **frequency table** of the distribution of these data appears below (Table 29-2).

If the shape of a frequency distribution needs to be examined, a graphical technique is employed. For discrete, ranked, or attribute data, a **bar graph** (Fig.

**Table 29-2**

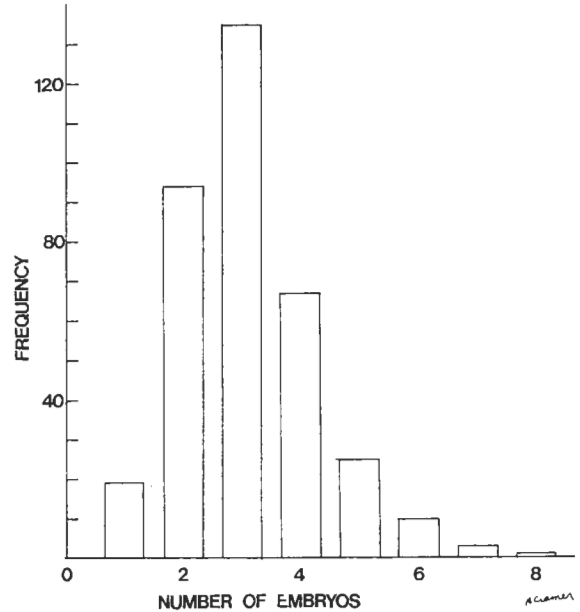
Frequency table of numbers of embryos in a sample of the cricetine rodent, *Holochilus sciureus berbicensis*. (Twigg 1965:271)

Number of Embryos	Frequency
1	19
2	94
3	135
4	67
5	25
6	10
7	3
8	1

29-1) should be prepared, keeping each vertical bar separate. For continuous data, a **histogram** (Fig. 29-2) is prepared with the vertical bars adjoining one another. Data of continuous variables may also be represented by a **frequency polygon** (Fig. 29-3) formed by a line that connects points. A relative **cumulative frequency polygon** (Fig. 29-4) may be plotted when you wish to examine the contribution of particular values to overall totals.

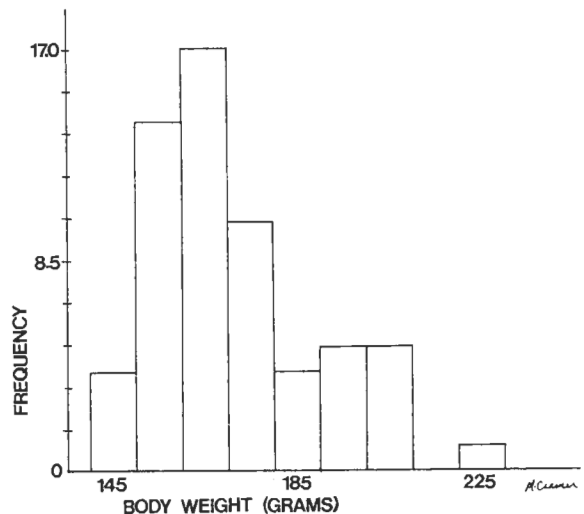
**Figure 29-1.**

Bar graph of the frequency data in Table 29-2. (Mary Ann Cramer)



**Figure 29-2.**

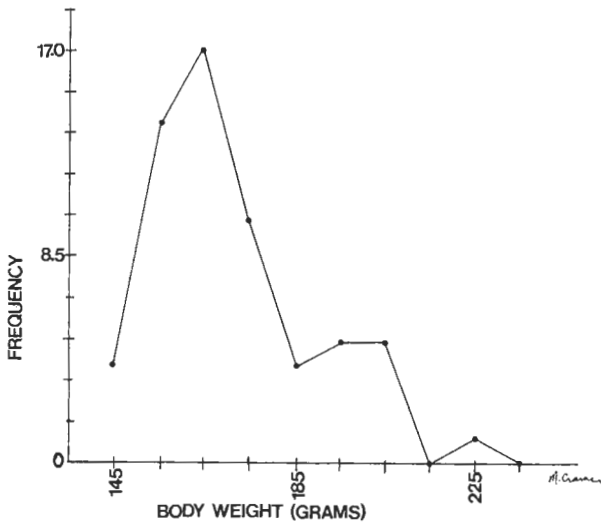
Histogram of body weights of the degu, *Octodon degus*. (Mary Ann Cramer)



**Figure 29-3.**

Frequency polygon of body weights of the degu, *Octodon degus*.

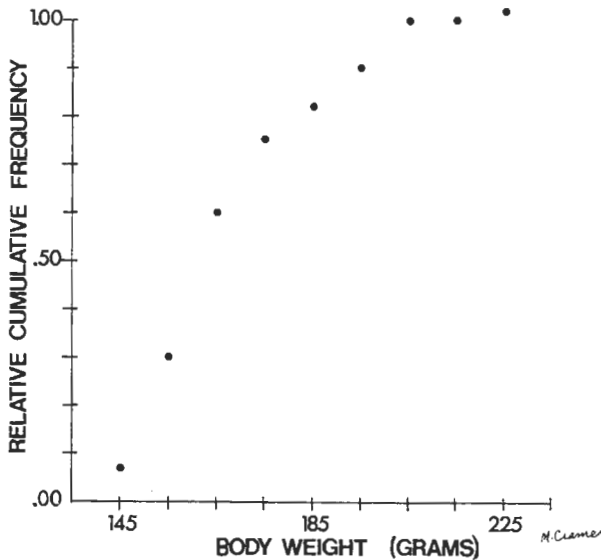
(Mary Ann Cramer)



**Figure 29-4.**

Cumulative frequency polygon of body weights of the degu, *Octodon degus*.

(Mary Ann Cramer)



### Central Tendency and Dispersion

In analyses of data, we generally wish to know something about **central tendency**, the localization of values near a central point, and **dispersion**, the scatter of these values from the central region. The **sample mean** ( $\bar{X}$ ), the average of a set of numbers ( $N$ ), is one of the best estimates of central tendency and the best and most consistent estimator of the **population mean** ( $\mu$ ).

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N}, \text{ where } \Sigma = \text{summation (formula 29-1)}$$

Other measures of central tendency will be discussed in the section on "Basic Statistics."

The **range**, the difference between the highest and lowest values, is a measure of dispersion or variability familiar to most persons. For statistical purposes, it is a crude measure since it frequently underestimates the range of the population.

The sum of all deviations from the mean is equal to zero.

$$\sum_{i=1}^N (X_i - \bar{X}) = 0 \quad (29-2)$$

The **population variance** ( $\sigma^2$ , sigma squared) is defined as the mean sum of squares of the deviations from the population mean ( $\mu$ ).

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N} \quad (29-3)$$

The population (parametric) standard deviation ( $\sigma$ ), the square root of the variance, is a useful measure since it is expressed in the same scale as the population values.

The **sample variance** ( $s^2$ ) is the best estimate of the population variance and is distinctly superior to the mean deviation for hypothesis testing (Zar 1974). Computational methods for calculating  $s^2$  and its derivatives will be described in the section on "Basic Statistics."

### Basic Statistics

**Descriptive statistics** provide a numerical summary on the properties of an observed frequency distribution. Measurements made for computing these statistics are generally recorded in tabular form (Table 29-3). Standard descriptive statistics include **sample size** ( $N$ ), **degrees of freedom** (generally  $N-1$ ), **range**, **arithmetic mean** ( $\bar{X}$  or  $\bar{Y}$ ), **variance** ( $s^2$ ) **standard deviation** ( $s$ ), **standard error of the mean** ( $s_{\bar{x}}$ ), and the **coefficient of variation** ( $CV$ ).

The symbol  $\Sigma$  indicates that a set of observations must be summed. By reference to Table 29-3 the mean for the character "condylobasal length" can be computed as follows:

$$\begin{aligned} \bar{X} &= \frac{\sum X_i}{N} = \frac{45.4 + 48.7 + \dots + 43.9}{8} \\ &= \frac{384}{8} = 48.0 \end{aligned}$$

The range of values for character 1 of taxon B is 43.9 to 51.8. Variance is a measure of dispersion of a set of data about the mean.

The variance is expressed in squared units. The standard deviation is the square root of the variance and is expressed in the same units as the original observations ( $X$ ). From data in Table 29-3 the values for variance and standard deviation are then calculated using computational formulas:

$$s^2 = \frac{\sum X^2 - (\sum X)^2/N}{N-1} \quad (29-4)$$

$$s^2 = \frac{(45.4)^2 + \dots + (43.9)^2 - (45.4 + \dots + 43.9)^2/8}{7}$$

$$s^2 = \frac{18478.54 - (384.0)^2/8}{7}$$

$$s^2 = 6.648$$

$$s = \sqrt{\frac{\sum X^2 - (\sum X)^2/N}{N-1}} \text{ or } s = \sqrt{s^2} \quad (29-5)$$

$$s = \sqrt{6.648} = 2.578$$

The standard error of the mean ( $s_{\bar{x}}$ ) is the standard deviation of the means for a sample. From the quantities above,  $s_{\bar{x}}$  is computed as follows:

$$s_{\bar{x}} = \sqrt{\frac{s^2}{N}} = \sqrt{\frac{6.648}{8}} = 0.912 \quad (29-6)$$

The coefficient of variation is the standard deviation expressed as a percentage of the mean. This permits comparison of variation in data when the mean or standard deviation values are very different; e.g., the measurements of a horse can be compared with those of a small shrew. The CV is computed on the basis of the above data as follows:

$$CV = \frac{(s)}{\bar{X}} (100) = \frac{257.8}{48.0} = 5.37 \quad (29-7)$$

One disadvantage of the coefficient of variation, as pointed out by Lewontin (1966) and Moriarty (1977), is the inability to perform exact statistical tests to compare CV values. Lewontin suggests transforming the measurements to logarithms (to any base) and then computing descriptive statistics on the characters, (e.g., A and B). Then, the ratio:  $s^2 \log A/s^2 \log B$  can be compared with an F-distribution to test the magnitude of the difference (if any). Lande (1977) points out other precautions that must be observed when using coefficients of variation (e.g., should not compare CV's based on discrete data).

Table 29-3

Cranial and bacular measurements and descriptive statistics based on them. Measurements are of a small sample of a single species at a single locality. Refer to text for explanation of symbols.

Identification Number of Individual ( $X_i$ )	Measurements			
	Condyllo- basal Length	Zygo- matic Breadth	Bacular Length	Bacular Width
A01	45.4	25.3	6.3	1.6
A02	48.7	25.8	—	—
A03	51.8	27.3	8.3	1.7
A04	49.3	25.5	8.4	2.0
A05	47.5	25.4	6.3	1.5
A06	47.1	24.6	—	—
A07	50.3	26.4	6.7	1.7
A08	43.9	23.8	7.4	1.7

Sample Statistics				
N	8	8	6	6
$\sum$	384.0	204.1	43.4	10.2
$\bar{X}$	48.00	25.51	7.23	1.70
$s^2$	6.648	1.127	0.911	0.028
s	2.578	1.062	0.954	0.167
$s_{\bar{x}}$	0.912	0.375	0.390	0.068
CV	5.37	4.16	13.19	9.82
$\bar{X} \pm t_{.05, df}$	48.0 $\pm$ 2.2	25.5 $\pm$ 2.1	7.2 $\pm$ 1.0	1.7 $\pm$ 0.2

To compute the confidence limits on these values we utilize the  $t$ -distribution (see section on Two-Sample Comparisons). Thus, to compute the 95% confidence limit for the measurement "condylobasal length," we must utilize the following values: degrees of freedom ( $df$ ),  $t$ -statistic from table ( $t_{tab.}$ ), standard error of the mean ( $s_{\bar{x}}$ ), and mean ( $\bar{X}$ ).

$$df = N - 1 \text{ or } 7 \quad t_{tab. (.05), 7df} = 2.365$$

$$s_{\bar{x}} = 0.912 \quad \bar{X} = 48.0$$

The general formula for obtaining the confidence limits for the mean is the following:

$$\bar{X} - t_{\alpha, df} s_{\bar{x}} \leq \mu \leq \bar{X} + t_{\alpha, df} s_{\bar{x}} \quad (29-8)$$

Thus,

$$48.0 - (2.365) (0.912) \leq \mu \leq 48.0 + (2.365) (0.912)$$

or, in abbreviated form,

$$\bar{X} \pm t_{.05, df} s_{\bar{x}} \text{ or } 48.0 \pm 2.16.$$

29-E. Verify this calculation and then substitute an  $\alpha$ -level of .01 to see how this changes the value.

$$\bar{X} \pm t_{.01,7} S_{\bar{x}} \quad \text{or } 48.0 \pm 3.19.$$

29-F. Compute the mean, standard deviation, standard error of the mean, coefficient of variation, and confidence limits for the measurements of bacular width and bacular length in Table 29-4. Check your answers with the values in the table. Compare the values of coefficient of variation for all four measurements. What do these values indicate?

Descriptive statistics, such as mean, range, standard deviation and standard error, can be presented in the form of Dice-Leraas diagrams or Dice-grams (Fig. 29-5). These diagrams are helpful to see overall patterns of variation but should not be used for

extensive testing of differences between means (see section on Multiple Samples and Comparisons).

## Probability Distributions

### Probability and Binomial Distribution

Probability and probability distributions are important concepts for understanding many statistical procedures. **Probability** is the chance for the occurrence of a particular event given the total number of possible outcomes of all events. For example, when a die (pl. dice) is thrown, one of six possible numbers may appear. Thus, the probability ( $p$ ), that the number "5" will appear is

$$p_{(5)} = 1/6 \text{ or } 0.167.$$

The probability ( $k$ ) of any of the other numbers occurring (i.e., 1,2,3,4,6) is

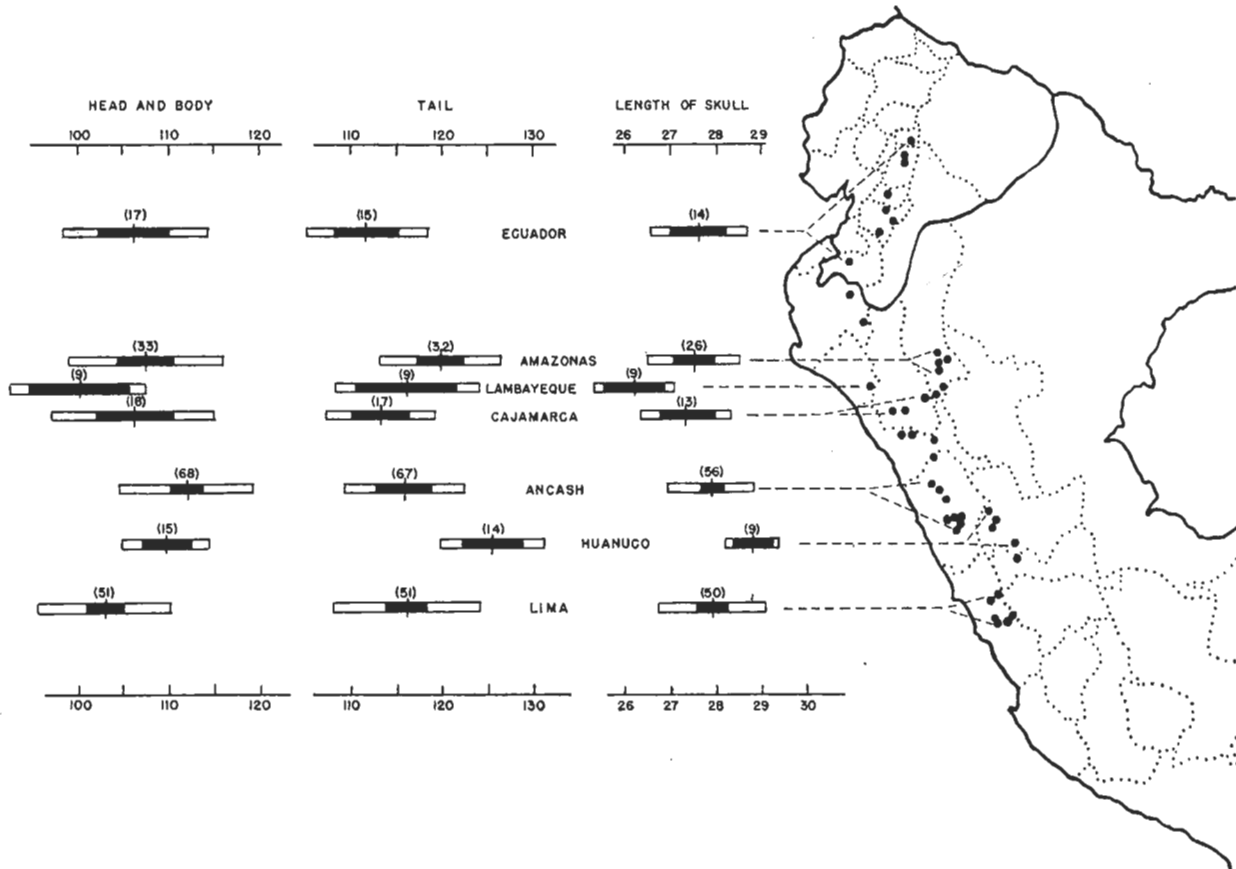
$$1 - 0.167 = 0.833.$$

Probability values always range between 0 and 1.

**Figure 29-5.**

Variation in measurements of two external characters and one cranial character between populations of the rodent, *Phyllotis andium*. Descriptive statistics are indicated by modified Dice-Leraas diagrams. Each bar shows the mean

(vertical line), twice the standard error of the mean (black rectangle), and standard deviation (black plus open rectangles). Sample sizes ( $N$ ) are indicated in the parentheses above the diagrams. (Pearson 1958: 438)



When a pair of dice is thrown, the probability of obtaining a pair of "5's" is the product of the independent probabilities:

$$p_{(5,5)} = (p_1) (p_2) = (1/6) (1/6) = 1/36 \text{ or } 0.028.$$

Further details on methods for computing probabilities can be found in textbooks such as Snedecor and Cochran (1967:199-202) and Sokal and Rohlf (1969:69-71).

The theoretical frequency distribution or **probability distribution** of events that can occur in two classes is known as the **binomial distribution** (Sokal and Rohlf 1969:71-81). Actual proportional data of a given sample size for two classes can then be compared with the theoretical distribution.

29-G. Compute probability values for the following situations:

1. Probability of obtaining two heads in two tosses of a coin: \_\_\_\_\_\*
2. Probability of selecting on one occasion a male from a cage containing five male and 10 female rats: \_\_\_\_\_\*

### Normal Distribution

Data that approximate a normal probability density function or bell-shaped **normal distribution** (Fig. 29-6) are necessary for conducting most **parametric** kinds of statistical tests. Many kinds of biological data such as lengths, weights, heights, and rates conform reasonably well to these distributions (Brower and Zar 1974). Data based on counts, frequencies, and percentages generally are not normally distributed and thus **nonparametric** methods of data analysis must be utilized (unless these data can be transformed into approximate normal distributions by the use of logarithms or square roots; Sokal and Rohlf 1969:380-387).

29-H. Examine Figure 29-6. What would happen to the shape of the normal distribution if the mean ( $\mu$ ) was as 10 and the standard deviation ( $\sigma$ ) 0.5? if  $\mu$  was 10 and  $\sigma$  1.5?

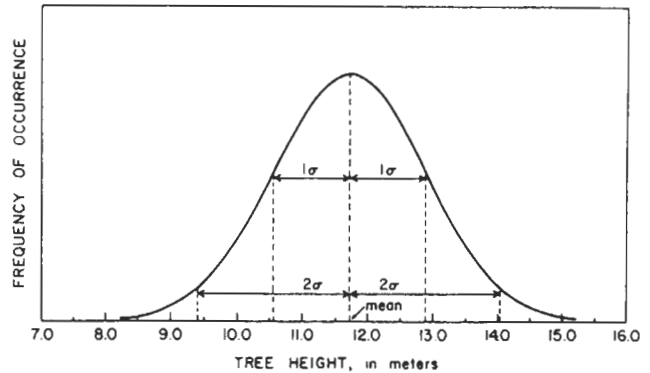
A **standardized normal distribution** is one in which the  $\mu = 0$  and  $\sigma = 1$ . Standard tables (e.g., Table D.9, Zar, 1974; Appendix I, Simpson, *et al.*, 1960) enable one to determine proportions of normal distributions.

\*Answers: 1.) = 0.25; 2.) = 0.33

**Figure 29-6.**

A normal distribution. These data are a hypothetical population of tree heights ( $X$ ), with a mean,  $\mu$ , of 11.72 m, and a standard deviation,  $\sigma$ , of 1.16 m. The mean  $\pm 1$  standard deviation includes 68.3% of any normal curve; the mean  $\pm 2$  standard deviations encompasses 95.5%; and  $\mu \pm 3\sigma$  includes 99.7%.

(Brower and Zar 1974: 12)



Thus, for a normal population with a mean ( $\mu$ ) and standard deviation ( $\sigma$ ), the expression

$$Z = \frac{X_i - \mu}{\sigma} \quad (29-9)$$

yields a **Z-value**, which indicates the number of standard deviations from the mean that an  $X$  value is located. These **Z-values** are termed **normal deviates** or **standard scores** and the calculation is referred to as normalizing or standardizing  $X_i$ -values. Since we rarely know the population mean ( $\mu$ ) and the standard deviation ( $\sigma$ ), we must use the sample approximations,  $\bar{X}$  and  $s$ , respectively. However, for small samples these statistics are poor approximations of the population parameters (Zar 1974:87).

### Testing Hypotheses

In statistics, an hypothesis is phrased very carefully and consists of two components: the null hypothesis and the alternative hypothesis. The **null hypothesis** (abbreviated  $H_0$ ) is a statement that there is no difference (e.g., between sample groups) and is formulated for the purpose of being rejected. The **alternative hypothesis** (abbreviated  $H_A$ ) is the operational statement or hypothesis that the researcher is testing. Thus, to test the assertion that differences exist between the mean body weights of two groups of rats, the hypothesis would be stated as follows:

$$H_0: \text{Group A} = \text{Group B or } H_0: A = B$$

$$H_A: A \neq B$$

Once the null hypothesis ( $H_0$ ) has been formulated, there must be an objective method for determining when to reject this hypothesis. First, let us examine the two types of errors that can be made when hypotheses are tested. A **Type-I error** (symbolized by  $\alpha$ ) is the *rejection of  $H_0$  when it is true* (or accepting a difference when there is none). A **Type-II error** (symbolized by  $\beta$ ) is the *acceptance of  $H_0$  when it is false* (or failing to find a difference when there is one).

Since our primary goal in hypothesis testing is to reject the null hypothesis when it is false, we generally wish to keep the probability of a Type-I error minimized to a stated  $\alpha$ -level. The larger the value of  $\alpha$ , the more probable that the null hypothesis will be rejected falsely (i.e., committing a Type-I error). Remember that all probability values, whether  $\alpha$  or  $\beta$  or some other parameter, range from 0 to 1.

The levels of  $\alpha$  and  $\beta$  are inversely related to each other and dependent on the sample size ( $N$ ). Thus, to decrease the possibility of both types of error,  $N$  must be increased. The *power* of a statistical test is the probability of rejecting  $H_0$  when it is false (or the probability of finding a difference when there is one). Stated in another form, the power of a test is

$$1 - \text{probability of a Type-II error or } 1 - \beta. \quad (29-10)$$

Generally, the power of a test increases with an increase in the sample size ( $N$ ) (Siegel 1956). A **test of significance** evaluates the probability of rejecting the null hypothesis when it is true. The probability of making a Type-I error ( $\alpha$ ), expressed as a percentage, is termed the **significance level**. For example, if  $\alpha$  is .01, then the significance level is 1% for a given test. If we choose a 5% level of significance, then we expect that only 5 of 100 samples examined will result in making a Type-I error (i.e., rejection of a true null hypothesis). In many scientific disciplines,  $\alpha$ -levels of .05, .01, and .001, are utilized for hypothesis testing. However, the choice of an appropriate  $\alpha$ -level is somewhat arbitrary and will depend on the nature of the investigation and the degree of predictability required. Refer to Sokal and Rohlf (1969:155-166) for further discussion on the selection of appropriate  $\alpha$ -levels.

- 29-1. If the significance level was set at 1% ( $\alpha = .01$ ) rather than 5%, would you be more likely to make a Type-I error at the 1% level if the sample size was smaller? How can you decrease the probability of making Type-I and Type-II errors?

Once the hypothesis has been formulated, an appropriate statistical procedure and test must be utilized. Data obtained from continuous variables are generally analyzed by **parametric statistics** since these variables most nearly follow a normal distribution (see section on "Probability Distributions"). Data from enumeration, discontinuous, and ranked types are generally analyzed by **nonparametric statistics** since no assumptions about the shape of the distributions are required to utilize these procedures. Large samples of enumeration and discontinuous data often have a nearly normal distribution and can be analyzed as if they were continuous. Ranked data, however, can never be analyzed with parametric statistics since ranks are relative and cannot be multiplied or divided.

## Two-Sample Comparisons

In biological problems, we frequently wish to know whether or not the means of two sample groups are significantly different from one another (e.g.,  $H_0 : \bar{X}_A = \bar{X}_B$  vs  $H_A : \bar{X}_A \neq \bar{X}_B$ ). If the sample values for these two groups (1) follow a normal distribution (or nearly so) and (2) the sample variances are not significantly different from one another, then parametric statistical procedures can generally be used for testing. If not, nonparametric statistics must be utilized.

One of the most common and useful tests for comparing sample means is Student's *t*-test or, simply, the *t*-test. The *t*-distribution is like a normal distribution when sample sizes approach infinity but the curve is more flattened for smaller sample sizes. Prior to utilizing the *t*-test or other parametric test for two samples, the **homogeneity of variances** (i.e., equality of variances) between the groups must be tested. For this purpose, we use the following statistic:

$$F_s = \frac{s^2_{\text{larger}}}{s^2_{\text{smaller}}} \quad (29-11)$$

for  $df_{\text{larger } s^2}, df_{\text{smaller } s^2}$

Then, substituting the appropriate variances ( $s^2$ ) from (Table 29-5) into the formula, we obtain the following:

$$F_s = \frac{5.67}{2.00}, \text{ df } 9, 8 \text{ (numerator, denominator).}$$

Since the  $F_{\text{tab } .05}$  value for 9 (numerator) and 8 (denominator) degrees of freedom is 3.39, we accept the null hypothesis that the variances are equivalent and thus can proceed with testing the equality of the group means.

To calculate a *t*-statistic ( $t_{\text{crl.}}$ ), we utilize the same procedures that were employed for obtaining basic

statistics, with these exceptions. A sum of squares (SS) is calculated for each group utilizing the basic formula for estimating variance ( $s^2$ ) with the exception of omitting the step where the quantity is divided by  $N-1$ . Thus, utilizing the values from Table 29-5, the sums of squares for the two groups are calculated as follows:

$$\begin{aligned} SS_A &= \Sigma X_A^2 - (\Sigma X_A)^2/N & (29-12) \\ &= 23622.07 - (485.5)^2/10 \\ &= 51.045 \\ SS_B &= \Sigma X_B^2 - (\Sigma X_B)^2/N \\ &= 20227.37 - (426.5)^2/9 \\ &= 16.009 \end{aligned}$$

To evaluate the means of the two groups, it is necessary to obtain estimate of the pooled variance ( $s_{pooled}^2$ ) and from this the standard error of the pooled mean ( $s_{x_{pooled}}$ ) as follows:

$$\begin{aligned} s_p^2 &= \frac{SS_A + SS_B}{df_A + df_B} & (29-13) \\ &= \frac{51.045 + 16.009}{9 + 8} \\ &= 3.94 \\ s_{\bar{X}_A} - \bar{X}_B &= \sqrt{\frac{s_p^2}{N_A} + \frac{s_p^2}{N_B}} & (29-14) \\ &= \sqrt{\frac{3.94}{10} + \frac{3.94}{9}} \\ &= \sqrt{.832} \\ &= 0.912 \end{aligned}$$

Then,  $s_{x_{pooled}}$  is substituted into the formula below to obtain  $t_{cal.}$ :

$$\begin{aligned} t_{cal.} &= \frac{X_A - X_B}{s_{\bar{X}_A} - \bar{X}_B} & (29-15) \\ &= \frac{48.6 - 47.4}{0.912} \\ &= 1.32 \end{aligned}$$

Inspection of Table 29-4 reveals that the tabulated value of the  $t$ -statistic for 17 df and  $\alpha = 0.05$  is 2.11. Thus, the null hypothesis of no difference in mean skull lengths of the male and female samples is accepted at the 5% level because  $t_{cal.} < t_{tab.}$

**Table 29-4**

Critical values of student's  $t$  (Brower and Zar 1974:10)

DF	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$
1	6.31	12.71	31.82	63.66
2	2.92	4.31	6.96	9.92
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.01	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71
7	1.89	2.36	3.00	3.50
8	1.86	2.31	2.90	3.36
9	1.83	2.26	2.82	3.25
10	1.81	2.23	2.76	3.17
11	1.80	2.20	2.72	3.11
12	1.78	2.18	2.68	3.06
13	1.77	2.16	2.65	3.01
14	1.76	2.14	2.62	3.00
15	1.75	2.13	2.60	2.95
16	1.75	2.12	2.58	2.92
17	1.74	2.11	2.57	2.90
18	1.73	2.10	2.55	2.88
19	1.73	2.09	2.54	2.86
20	1.72	2.09	2.53	2.85
22	1.72	2.07	2.51	2.82
24	1.71	2.06	2.49	2.80
26	1.71	2.06	2.48	2.78
28	1.70	2.05	2.47	2.76
30	1.70	2.04	2.46	2.75
35	1.69	2.03	2.44	2.72
40	1.68	2.02	2.42	2.70
45	1.68	2.01	2.41	2.69
50	1.68	2.01	2.40	2.68
60	1.67	2.00	2.39	2.66
70	1.67	1.99	2.38	2.65
80	1.66	1.99	2.37	2.64
90	1.66	1.99	2.37	2.63
100	1.66	1.98	2.36	2.63
120	1.66	1.98	2.36	2.62
150	1.66	1.98	2.35	2.61
200	1.65	1.97	2.35	2.61
300	1.65	1.97	2.34	2.59
500	1.65	1.96	2.33	2.59
$\infty$	1.65	1.96	2.33	2.58

The above values were computed as described by Zar (1974: 414). More extensive tables of Student's  $t$  are found in Rohlf and Sokal (1969:160-161) and Zar (1974:413-414).

**Table 29-5**

Skull measurements (condylobasal lengths) of two samples (males and females) of a single species from the same locality.

Males		Females	
Specimen No.	CBL in mm	Specimen No.	CBL in mm
068	51.4	059	48.2
064	51.6	051	48.9
067	51.4	073	49.0
056	49.8	009	47.1
048	46.7	062	46.2
071	46.4	061	45.2
053	49.1	064	47.7
065	47.2	057	48.5
072	45.6	052	45.7
054	46.3		

<i>N</i>	10	9
<i>df</i>	9	8
$\bar{X}$	48.6	47.4
SS	51.045	16.009
<i>s</i> <sup>2</sup>	5.67	2.00
<i>s</i>	2.38	1.41
<i>s</i> <sub><math>\bar{x}</math></sub>	0.753	0.472
$\bar{X} \pm t_{.05} s_{\bar{x}}$	48.6 $\pm$ 1.70	47.4 $\pm$ 1.09

$t_{\text{tab. } (.05), 17 \text{ df}} = 2.110$
$t_{\text{cal.}} = \frac{\bar{X}_m - \bar{X}_f}{s_{\bar{x}} - \bar{x}} = \frac{48.6 - 47.4}{0.9} = 1.32$

Since  $t_{\text{tab.}} > t_{\text{cal.}}$ ,  
the null hypothesis,  
 $H_0 : A=B$ , is accepted

A single-classification analysis of variance (ANOVA) can also be used to compare group means (Steel and Torrie 1960; Sokal and Rohlf 1969:218-219).

For nonparametric data, the Mann-Whitney U test (Siegel 1956:116-127; Sokal and Rohlf 1969:392-394) and the Kolmogorov-Smirnov two-sample test (Siegel 1956:127-136) are appropriate. The latter test should be applied only to nonparametric data of a continuous variable (e.g., continuous data not meeting the assumptions of a normal distribution).

For nonparametric discrete data, the chi-square test for independent samples is frequently very useful. Suppose, for example, that we wish to compare the sex ratios in several litters (pooled) of a species of rodent. The null hypothesis to be tested states that half of the sample will be male and the other half female ( $H_0 : P = 0.5$ ). Upon examination of the litters, we discover that 20 are male and 17 are female.

The chi-square ( $X_s^2$ ) statistic\* is computed according to the following formula:

$$X_s^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (29-16)$$

where  $O_{ij}$  = observed number of cases in the *i*th row (horizontal) of *j*th column (vertical) and

$E_{ij}$  = number of cases expected under  $H_0$  in *i*th row and *j*th column

and  $\sum_{i=1}^r \sum_{j=1}^k$  indicates to sum over all rows (*r*)

and all columns (*k*), i.e., over all cells.

Since, there are only two cells, there is only one k-category and the expected value is determined by multiplying *N* times the predicted probability of occurrence (i.e., 0.5). Thus, the expected value for each cell is 18.5 ( $0.5 \times 37$ ). Substituting the observed and expected values into the equation, the  $X_s^2$  statistic is computed as follows:

$$\begin{aligned} X_s^2 &= \frac{(20 - 18.5)^2}{18.5} + \frac{(17 - 18.5)^2}{18.5} \\ &= (.1216) + (.1216) \\ &= 0.243 \end{aligned}$$

In order to compare the calculated value of  $X^2$  with the tabulated value, we must determine the number of degrees of freedom and then utilize Table 29-6. The general formula for determining the degrees of freedom is  $(r-1)(k-1)$ . Since  $k = 1$  in the present example, the appropriate degrees of freedom is  $r-1$  or  $2-1 = 1$ . Thus, the tabulated value of chi-square for an  $\alpha$ -level of .05 is 3.84. Since  $X_{\text{cal.}}^2 < X_{\text{tab.}}^2$ , we accept the null hypothesis that the sex ratio does not differ significantly from a 1:1 ratio.

There are special formulas for calculating  $X^2$  values when data are arranged in  $2 \times 2$  contingency tables and when the expected frequencies must be calculated from the marginal totals. Refer to Siegel (1956:42-47, 104-111, and 175-179) or Sokal and Rohlf (1969:549-620) for additional information.

## Multiple Samples and Comparisons

Several statistics can be used to test for the significance of differences between the means of samples. An analysis of variance (ANOVA) not only gives an indication of differences between means but provides

\*See Sokol and Rohlf (1969:553) for the rationale behind using the symbol  $X^2$  rather than  $\chi^2$  for this quantity.

**Table 29-6**

Critical values of Chi-square. (Brower and Zar 1974:15-16)

DF	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
1	2.706	3.841	5.024	6.635
2	4.605	5.991	7.378	9.210
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.070	12.833	15.086
6	10.645	12.592	14.449	16.812
7	12.017	14.067	16.013	18.475
8	13.362	15.507	17.535	20.090
9	14.684	16.919	19.023	21.666
10	15.987	18.307	20.483	23.209
11	17.275	19.675	21.920	24.725
12	18.549	21.026	23.337	26.217
13	19.812	22.362	24.736	27.688
14	21.064	23.685	26.119	29.141
15	22.307	24.996	27.488	30.578
16	23.542	26.296	28.845	32.000
17	24.769	27.587	30.191	33.409
18	25.989	28.869	31.526	34.805
19	27.204	30.144	32.852	36.191
20	28.412	31.410	34.170	37.566
21	29.615	32.671	35.479	38.932
22	30.813	33.924	36.781	40.289
23	32.007	35.172	38.076	41.638
24	33.196	36.415	39.364	42.980
25	34.382	37.652	40.646	44.314
26	35.563	38.885	41.923	45.642
27	36.741	40.113	43.195	46.963
28	37.916	41.337	44.461	48.278
29	33.711	39.087	42.557	45.722
30	40.256	43.773	46.979	50.892
31	41.422	44.985	48.232	52.191
32	42.585	46.194	49.480	53.486
33	43.745	47.400	50.725	54.776
34	44.903	48.602	51.966	56.061
35	46.059	49.802	53.203	57.302
36	47.212	50.998	54.437	58.619
37	48.363	52.192	55.668	59.893
38	49.513	53.384	56.896	61.162
39	50.660	54.572	58.120	62.428
40	51.805	55.758	59.342	63.691

The above values were computed as described by Zar (1974: 411). More extensive tables of chi-square are found in Rohlf and Sokal (1969:164-167) and Zar (1974:409-410).

a measure of variation within samples. Sokal and Rohlf (1969:173-366), Steel and Torrie (1960:99-160), and Zar (1974) give extensive accounts of the use of ANOVA's.

A *t*-test is inappropriate for making multiple paired comparisons of means (Sokal 1965). When more than two samples are involved an ANOVA can be used to test for overall difference between the means, although significant differences between pairs of means cannot be established. A *posteriori* tests, such as the sum of squares simultaneous test procedures (SS-STP) described in Sokal and Rohlf (1969), permit determination of homogeneous subsets of means within the total collection of means, *e.g.*, between means from different geographic localities. The Student-Newman-Keuls (SNK) test (Sokal and Rohlf 1969) and Duncan's multiple range test (Steel and Torrie 1960) have also been used to test multiple means. Some researchers believe that the last two tests are more useful, since the experimental error rate is not altered. In contrast, the SS-STP procedure generates more possible answers than can be realistically evaluated.

### Covariate Analysis

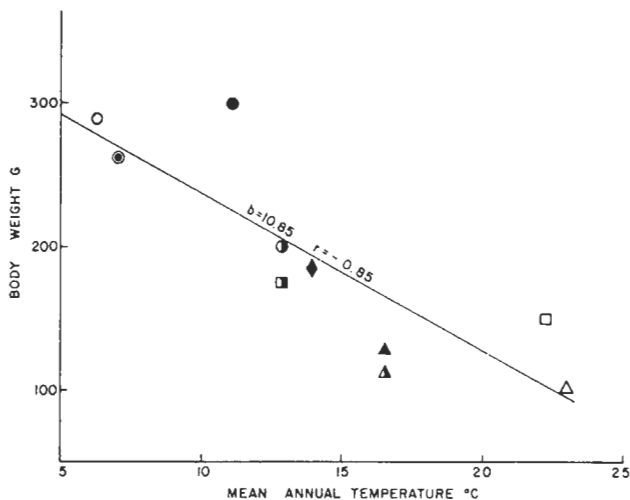
Correlation and regression are techniques of covariate analysis. **Correlation analysis** is an investigation of the degree of association between pairs of variables (*e.g.*, forelimb length versus hind limb length). Correlation analysis estimates the strength of the relationship between variables but implies no cause and effect relationship between the two. **Regression analysis** seeks to estimate the dependence of one variable (Y) on another, independent (X) variable (Fig. 29-7). Such a relationship, expressed mathematically, is generally written as a function (termed the regression equation), such as  $Y = fX$ , where the magnitude of a given Y is dependent upon the value of a given X. The slope of a regression line is termed the regression coefficient (b).

Regression analysis can be used in studies of differential growth (allometry) of body parts or regions. Differences between regression coefficients can be tested using a *t*-test or analysis of covariance (Steel and Torrie 1960; Sokal and Rohlf 1969).

A correlation coefficient (*r*) is a measure of interrelation between two variables, independent of the scale of measurement. The most commonly used correlation coefficient is Pearson's product-moment correlation coefficient. Procedures for the calculation of this statistic may be found in Sokal and Rohlf (1969: 508-515) and Zar (1974:236-240). In addition, many of the programmable calculators currently on the market have routines to calculate this statistic.

**Figure 29-7.**

Relationship between mean body weights of three species of *Neotoma* from ten populations and mean annual temperatures. The regression line, its slope ( $b$ ), and the correlation coefficient ( $r$ ) are given. *N. cinerea*, circles; *N. albigula*, squares; *N. lepida*, triangles. (Brown and Lee, 1969)



### Multivariate Analysis

Multivariate analysis simultaneously considers variation and covariation of two or more variables. Computations involving three or more variables are extremely complex and time consuming. Thus, the wide application of multivariate statistical techniques in biological studies awaited the development of electronic digital computers with their capability for rapidly processing numerous variables and data points.

General references on multivariate analysis include Cooley and Lohnes (1971), Morrison (1967), Seal (1964), and Anderson (1958). A working knowledge of matrix algebra is helpful, though not essential, for using multivariate statistical procedures (Searle 1966 is a useful reference on matrix manipulations). Sneath and Sokal (1973) provide information on multivariate analyses in phenetic classification studies. A useful key for determining what types of multivariate analyses to utilize may be found in Atchley and Bryant (1975:3-4) and Bryant and Atchley (1975:2-3).

In multivariate statistical analyses, the biologist is interested in one or all of the following: (1) a measure of similarity between groups (e.g., taxa); (2) reduction in the number of variables; and (3) discrimination between groups. Similarity can be measured by correlation, association, or distance coefficients. Phenograms are frequently constructed using the Unweighted Pair Group Method of Analysis (UPGMA)

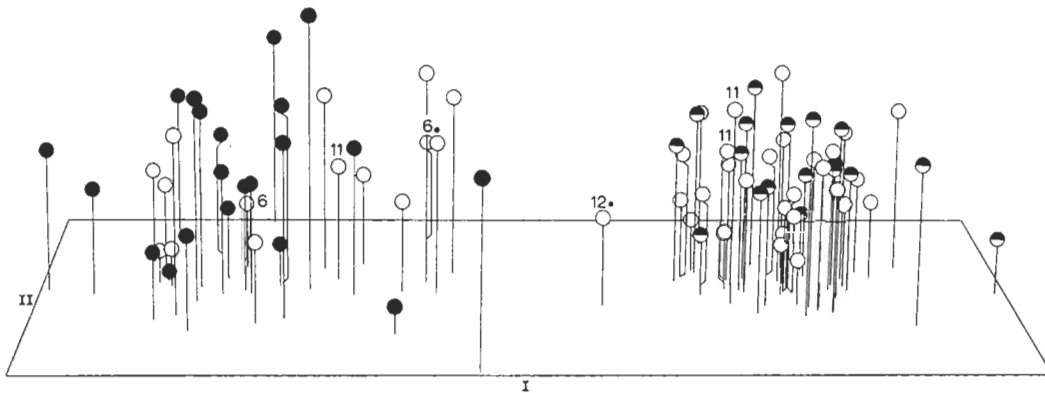
on correlation and average taxonomic distance matrices. Choate (1970), Genoways and Jones (1971), and Johnson and Selander (1970) and Patton, et al. (1975) are examples of phenetic cluster schemes using the UPGMA technique. Seal (1964), in contrast, recommended the use of a distance coefficient that considers relative correlation, such as generalized or Mahalanobis distance ( $D^2$ ).

Factor analysis is a general term for several multivariate techniques that convert a large number of original variables into a smaller set of new variables. Two techniques that are used in systematics research include principal components analysis and multiple-factor analysis (frequently with rotation to simple structure). The principal components analysis has a sound mathematical basis although interpretation is sometimes difficult. Multiple-factor analysis is less exact mathematically since there are no unique solutions for obtaining communalities (summarization of inter-correlations among variables) or for estimating the number of factors to extract from the many potential factors. Despite these difficulties, factor analysis is an important summarization technique. Genoways and Choate (1972) utilized principal components analysis in a study of geographic variation in Nebraska populations of *Blarina*. In their study the first three principal components accounted for approximately ninety-two per cent of the total variance in nine cranial and three external measurements (Fig. 29-8). Multiple-factor analysis, with rotation to simple structure, was used by Wallace and Bader (1967) in a study of twenty-seven morphometric variables in a single sample of the house mouse, *Mus musculus*. To improve interpretability and understanding of the forces affecting tooth size, the twenty-seven variables were reduced to five factors, of which the first three were identified as width, anterior length, and posterior length factors. Poole (1971, 1974) utilized factor analysis for modeling natural communities of plants and animals and for measuring the structural similarity of communities composed of the same species.

Discriminant functions were developed by Fisher as a means to distinguish members of closely related taxa. The computations produce differential weights for the various characters. Those characters with the highest weights, loadings, are the most useful "discriminators" for separating two groups or taxa. A step-wise discriminant analysis can be used if more than two reference groups or taxa must be separated. Summed values of the discriminant scores are frequently plotted on a frequency histogram (Y-axis, individuals; X-axis, discriminant scores) to illustrate the separation between taxa (Fig. 11-8). Genoways and Choate (1972) were interested in analyzing the specific relationships

**Figure 29-8.**

Three-dimensional projection of 83 specimens of *Blarina* onto the first three principal components (the third component is indicated by height). Solid circles, *B. b. brevicauda* reference sample; half solid circles, *B. b. carolinensis* reference sample; open circles test specimens of both taxa collected near zones of contact. (Genoway and Choate, 1972)



of two previously defined subspecies of short-tailed shrews (*Blarina*) occurring in a contact zone in Nebraska. After collecting specimens of these shrews from the contact zone, they wished to compare the morphology of these specimens with the morphology of reference specimens representing each of the two subspecies. The technique of discriminant function analysis permitted the calculation of discriminant scores for each of the two reference samples. Then, when the discriminant scores for the specimens in the contact zone were compared with the scores of the reference samples, the taxa were easily separated and potential hybrids (or intergrades) spotted (Fig. 11-8). Discriminant analyses were also utilized by Jolicoeur (1959) and Lawrence and Bossert (1967) in studies of canid populations and Robinson and Hoffmann (1975) in studies of geographical and interspecific cranial variation in big-eared ground squirrels (*Spermophilus*).

### Supplementary Readings

- Bliss, C. I. 1967. *Statistics in biology*, Vol. 1. McGraw-Hill Book Company, New York. 558 pp.
- Campbell, R. G. 1974. *Statistics for biologists*, 2nd ed. Cambridge Univ. Press, London. 385 pp.
- Rohlf, F. J., and R. R. Sokal. 1969. *Statistical tables*. W. H. Freeman and Co., San Francisco. 253 pp.
- Siegel, S. 1956. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill Book Co., New York. 312 pp.
- Simpson, G. G., A. Roe, and R. C. Lewontin. 1960. *Quantitative zoology*, rev. ed. Harcourt, Brace and Co., New York. 440 pp.
- Sokal, R. R. 1965. Statistical methods in systematics. *Biol. Rev.*, 40:337-391.
- Sokal, R. R., and F. J. Rohlf. 1969. *Biometry*. W. H. Freeman and Co., San Francisco. 776 pp.
- Steele, R. G. D., and J. H. Torrie. 1960. *Principles and procedures of statistics*. McGraw-Hill Book Co., New York. 481 pp.
- Zar, J. H. 1974. *Biostatistical analysis*. Prentice-Hall, Englewood Cliffs, New Jersey. 620 pp.