300 - level Physiology Lab.

Terms and Definitions

null hypothesis (H_0)-a statement of no difference. Experiments and statistical analyses are designed to determine if there are differences between two or more groups. Statistical results provide information for determining if there is sufficient reason to reject the H_0 (the hypothesis that there are no differences between or among groups). In science we do not prove, but rather disprove. Scientific advancement is made by rejecting null hypotheses and accepting alternative hypotheses (H_A) that bring us closer to the absolute truth.

population-a defined group in space and time that includes all individuals about which inferences are made. For example, in discussions of the average or mean SAT score this year at your school, the student body at your school is the population. A population is symbolized by N.

sample-a collection of individual observations selected by a specific procedure and criteria to represent the population being sampled. A sample is a subset of the population that is used to make inferences about the population at large. For example, only 100 randomly selected SAT scores (a sample) may be used to estimate the mean SAT score of this year's student body at your school (the population). A sample is symbolized by n.

parameter-a quantity or absolute measure of central tendency to describe and/or define a characteristic of a population. Parameters are rarely reported because it is difficult to measure all individuals within a population.

statistic-an estimation of a population parameter, achieved by the use of a sample.

variable-a characteristic of an object or individual that can vary in magnitude or amount.

variate-a single score, reading, or observation of a variable. It is symbolized by $X_1, X_2, X_3 \ldots$

accuracy-the nearness of a measurement to the actual value of the variable being measured.

precision-the closeness of repeated measurements of the same quantity; repeatability of measurement.

frequency-how often a particular value of a variable occurs within a sample.

relative frequency-frequency divided by the total number in the sample; often expressed as a percentage.

mean-the sum of all variates divided by the total number of variates in the population or sample.

Population mean: $\mu = \dfrac{\Sigma x}{N}$

Sample mean: $\bar{x} = \dfrac{\Sigma x}{n}$

BOX 1.1 Example calculation of mean.

Sample: 9, 3, 5, 3

Sample size (n) = 4

$$\bar{x} = \frac{9 + 3 + 5 + 3}{4} = \frac{20}{4} = 5$$

median-the value or variate that divides the ordered sample (data set) into two equal halves. The median can be calculated by ranking all observations and using the following equations:

Odd number of observations: $X_{\frac{n+1}{2}}$

Even number of observations: $\dfrac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$

BOX 1.2 Example calculation of median (for an even number of observations).

Sample: 9, 3, 5, 3

Sample size (n) = 4

Ranked sample:

$X_1 = 3, X_2 = 3, X_3 = 5, X_4 = 9$

$$Median = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \frac{X_2 + X_3}{2} = \frac{3 + 5}{2} = 4$$

mode-the variate in the frequency distribution with the greatest number of observations.

BOX 1.3 Example calculation of mode.

Sample: 9, 3, 5, 3

Sample size (n) = 4

Variate with the greatest frequency = 3

range-an indication of magnitudinal difference between the smallest and largest observations.

BOX 1.4 Example calculation of range.

Sample: 9, 3, 5, 3

Sample size (n) = 4

Range = 9 − 3 = 6

variance-the degree of deviation or spread in a distribution about the mean. Variance is mathematically defined by:

$$\text{Population variance: } \sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

$$\text{Sample variance: } s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

BOX 1.5 Example calculation of variance.

Sample: 9, 3, 5, 3

Sample size (n) = 4; Mean ($\bar{x}$) = 5

$$s^2 = \frac{(9-5)^2 + (3-5)^2 + (5-5)^2 + (3-5)^2}{4-1}$$

$$s^2 = \frac{(4)^2 + (-2)^2 + (0)^2 + (-2)^2}{3}$$

$$s^2 = \frac{16 + 4 + 0 + 4}{3} = \frac{24}{3} = 8$$

standard deviation-like variance, this describes the degree of deviation or spread in a distribution about a mean. Standard deviation is the square root of the variance and is mathematically defined by:

$$\text{Population standard deviation: } \sigma = \sqrt{\sigma^2}$$

$$\text{Sample standard deviation: } s = \sqrt{s^2}$$

BOX 1.6 Example calculation of standard deviation.

Sample: 9, 3, 5, 3

Sample size (n) = 4; Variance (s^2) = 8

$$s = \sqrt{8} = 2.83$$

standard error-also called the standard deviation of means because it standardizes sample means based upon sample size. Standard error is mathematically defined by:

$$SE = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$

BOX 1.7 Example calculations of standard error.

Sample: 9, 3, 5, 3

Sample size (n) = 4; standard deviation (s) = 2.83; variance (s^2) = 8

$$SE = \frac{s}{\sqrt{n}} = \frac{2.83}{\sqrt{4}} = \frac{2.83}{2} = 1.41$$

$$SE = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{8}{4}} = \sqrt{2} = 1.41$$

You have now finished reviewing the calculations for some simple but important statistics used to summarize and describe data. These statistics are **descriptive statistics** because they simply describe data. The mean, median, and mode are **measures of central tendency** because they describe how a set of observations estimate or are centered about a true population parameter. Range, variance, standard deviation, and standard error are **measures of dispersion** because they describe the amount of variation about the mean.

Statistical Exercises

EXERCISE 1.1 Descriptive Statistics

Observe the following sample of testosterone concentrations (ng/ml) measured in the blood plasma of 10 adult men. Using this data set, calculate all descriptive statistics by hand with the use of a calculator. Be sure to write out all mathematical expressions and show all work.

Sample: 4, 6, 9, 8, 10, 8, 4, 3, 5, 8

Measures of Central Tendency

Mean =

Median =

Mode =

Measures of Dispersion

Variance =

Standard deviation =

Standard error =

Now enter these data into the statistical software package provided and explained to you by your laboratory instructor. Use the "descriptive statistics function" in this software program to check your calculations; print a copy of your results. Ask your laboratory instructor for help and additional instruction as needed.

EXERCISE 2 The Normal Probability Distribution and The 95% Confidence Interval

You have, no doubt, taken a course in which grades assigned were based on a "normal curve." This reference is to the familiar bell-shaped normal probability distribution that is produced when some parameter of a population (e.g., height of adult men) is plotted on the abscissa (x-axis) versus the frequency of a given measurement on the ordinate (y-axis). Theoretically, the tails of the normal curve extend infinitely in both directions. If the total area under the curve is taken as 100%, approximately 34% of the population will be distributed between the mean measurement and +1.0 standard deviations of the mean. Similarly, approximately 34% of the population will lie between the mean and −1.0 standard deviations of the mean. This indicates that 68.27% of the population is distributed within the range of the mean −1.0 and +1.0 standard deviations ($\mu \pm 1\sigma = 68.27\%$) (Figure 1-1).

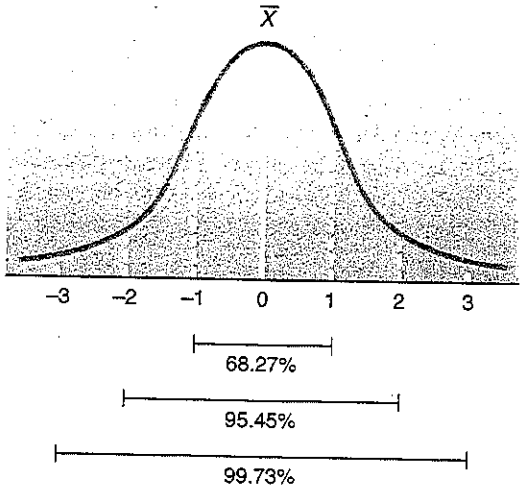

Figure 1-1 Normal distribution of a population.

For example, let us assume that the height distribution of adult men is such that the mean ($\overline{x}$) height is 70 inches and the standard deviation is 3 inches. This indicates that 68.27% of all men in the population are between 67 and 73 inches. The range $\mu \pm 2\sigma$ includes 95.45% of the population, and the range $\mu \pm 3\sigma$ includes 99.73% of the population.

If you sample a person's height, there is a 95% chance that it would fall between ±1.96 standard deviations about the mean. There is a 5% chance that it would not be included in this range. Thus, we accept or reject an observation belonging to the normal population based upon a probability of $P = 0.05$. If $P < 0.05$ we say it is statistically different and that it belongs to a different population. For example, suppose you sample an individual's height at 89 inches (7 feet 5 inches). How do you decide whether this person is one of the extremes in the normal population or if this person is part of a different population (one affected by the overproduction of growth hormone leading to gigantism)?

An easy way to determine if groups belong to different populations because they differ statistically is to calculate and plot the **95% confidence intervals** (95% CI) for each sample of a population. In theory, if the 95% CI overlap, the groups are not different from one another (i.e., they belong to the same population) as defined by the normal probability distribution. Calculations of the upper and lower limits of the 95% CI are:

Lower limit = $\overline{x} - 1.96(SE)$
Upper Limit = $\overline{x} + 1.96(SE)$

Imagine that you have just collected resting heart rate data in beats per minute (BPM) for three

groups of patients. These groups are patients with low, normal, and high blood pressures. You wish to know if the resting heart rates are different among these groups. A 95% CI for each group of patients may give you an idea of similarities or differences among the resting heart rates of these patients.

Calculate the 95% CI for each of these three groups. Show all work in the space provided.

Low Blood Pressure: 73, 72, 76, 74, 71

$\bar{x}$ =
Lower limit =
Upper limit =

Normal Blood Pressure: 82, 80, 70, 75, 88

$\bar{x}$ =
Lower limit =
Upper limit =

High Blood Pressure: 92, 93, 100, 102, 98

$\bar{x}$ =
Lower limit =
Upper limit =

Now inspect Figure 1-2 and answer the questions regarding the heart rates of these three groups of patients.

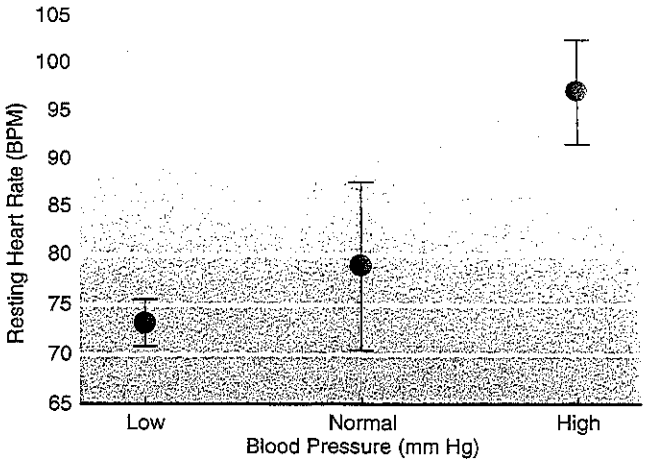

Figure 1-2 Example plot of group means and their 95% confidence intervals.

QUESTIONS

1. Compare your calculations with Figure 1-2. Do your calculations agree with the graph?

2. Is the heart rate of patients with low blood pressure statistically different from that of patients with normal blood pressure? Why or why not?

3. Of these three groups, which has the statistically different heart rate?

4. Which group has the most variable heart rate? How did you come to this conclusion?

EXERCISE 1B Student's *t*-test

A *t*-test is a statistical test used to compare the means of two populations or distributions. If we were interested in comparing the heart rates of patients with low and normal blood pressures, we could use a *t*-test because we are only comparing two groups. Enter the heart rate data for patients

with low and normal blood pressures into your statistics program and obtain the descriptive statistics, perform a *t*-test, and report your results. Your lab instructor will help you to interpret the results of the test.

The statistical report will show a **probability** or ***P*-value.** This value will tell you if the groups are significantly different statistically. If $P > 0.05$ the individuals are not different. If $P < 0.05$, the individuals are significantly different from each other with respect to heart rate. Were your conclusions from using the 95% CI in Exercise 1.2 accurate with respect to the results of the *t*-test?

The results that should be reported for a *t*-test are the *t*-statistic, degrees of freedom, and the *P*-value following a statement of your findings.

BOX 1.8 Example result statement for a *t*-test.

The heart rates of patients with low and normal blood pressures are not significantly different from one another ($t = -1.82$; $df = 8$; $P > 0.05$).

To understand how these results were obtained, use the descriptive statistics and calculate the *t*-statistic by hand. The equation for the *t*-statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}; \; df = (n_1 + n_2) - 2$$

Show your work here.

EXERCISE 1.4 The Analysis of Variance

The analysis of variance (ANOVA) is somewhat like a *t*-test but is used to statistically compare more than two groups. You cannot use a *t*-test to compare several groups as pairs; an ANOVA must be used. For example, let's imagine that you now wish to determine if heart rate is statistically different among the three groups of patients with low, normal, and high blood pressures. Because you are now comparing more than two groups in this example, the ANOVA is used. Enter the data into your statistics program, calculate the descriptive statistics, perform an ANOVA, and report your results. Once again, your lab instructor will help you with interpretation of the results.

The results that should be reported are the *F*-statistic, degrees of freedom, and the *P*-value following a statement of your findings.

BOX 1.9 Example result statement for an ANOVA.

We found heart rate to differ significantly among individuals with low, normal, and high blood pressures ($F = 33.14$; $df = 2,12$; $P < 0.05$).

You may be wondering why we state that these heart rates are different. We just showed with a *t*-test that heart rates of patients with low and normal blood pressures are not statistically different. So why does the ANOVA tell us that there are differences among these patients? Remember that an ANOVA determines if there are any differences among groups. We know that patients with high blood pressure have significantly different heart rates (Figure 1-2). Thus, the ANOVA gives us a *P*-value less than 0.05, indicating that at least one group's mean heart rate is statistically different from the others. Now you can use your 95% CIs to determine which mean is different. A multiple comparisons test could also be used to determine which means are different. Ask your lab instructor to give further instruction on multiple comparisons tests if interested.

Please note that the *df* in an ANOVA are calculated as follows: *df* = number of groups minus 1, number of observations minus the number of groups. Thus $df = 3 - 1, 15 - 3 = 2, 12$.

EXERCISE 1.5 Regression Analysis

Let us imagine that you were interested in investigating the effects of testosterone treatment duration on bone density in males over 65 years of age. You collect the following data set for analysis. Enter these data into the computer, perform a regression analysis, and report your results.

A regression analysis will generate an ANOVA table as you have seen in earlier analyses. A regression analysis will:

a. Determine if there is a cause-and-effect relationship between the dependent and independent variables.
b. Determine the nature of the relationship as defined by a mathematical equation.
c. Evaluate how accurately the mathematical equation defines the relationship.

Duration of Testosterone Treatment (Months)	Bone Mineral Density (% Increase)
0	0
6	0.8
6	1.0
6	0.6
12	1.8
12	2.4
12	1.5
18	2.3
18	2.7
18	1.9
24	2.9
24	3.4
24	2.6
30	3.6
30	3.7
30	3.1
36	4.2
36	5.1
36	3.3

Data taken from Snyder et al. 1999.

To properly report the results of a regression analysis, you should report the *F*-statistic, the two degrees of freedom, the *P*-value from the ANOVA table, and the r^2 value following a statement of your findings.

Note that the *df* in a regression are always 1 and the number of observations minus the number of variables (2 in a simple linear regression).

BOX 1.10 Example result statement for a regression.

We found a significant relationship between testosterone treatment duration and bone density in men over 65 years of age ($F = 137.4$; $df = 1,17$; $P < 0.05$). The duration of testosterone treatment explains nearly 88.3% of the variation in bone density ($r^2 = 0.883$).

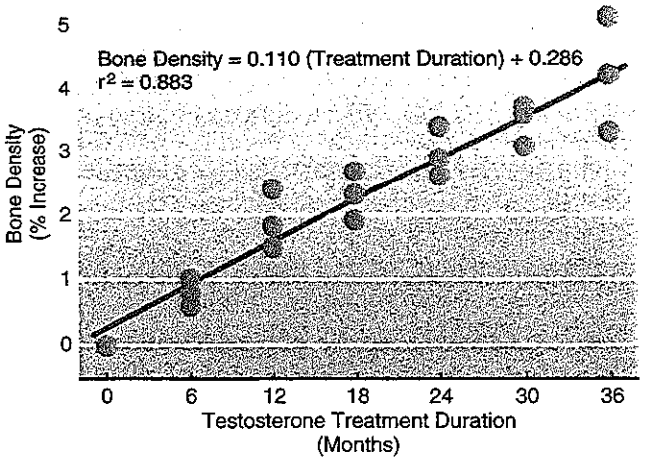

Figure 1-3 Example plot of a regression.

Now to complete the analysis, use your statistics program or a graphing program to plot these data showing the independent variable (treatment duration) on the abscissa (x-axis) versus the dependent variable (bone density) on the ordinate (y-axis). Recall that the dependent variable is the variable that is affected or is dependent upon the independent variable. Plot the best-fit line defined by the regression equation and report the regression equation on the graph. This equation is in the form $y = m(x) + b$ where m is the slope and b is the y-intercept. Your graph should look something like Figure 1-3.

EXERCISE 1.6 Correlation

There are times in biology when an investigator is interested in the possible *association* between two variables but is uncertain if there is a *cause-and-effect* relationship between the variables. In this case, the investigator performs a correlation analysis rather than a regression analysis. A correlation assumes no cause-and-effect relationship between the variables and, rather, attempts to find possible correlates. Think about this in the following way: Do wet roads after a rainfall cause automobile accidents? The answer is no. Wet roads do not cause accidents, or else no one would ever attempt to drive on wet roads. However, wet roads increase the likelihood that some people may get into accidents because they attempt to drive on wet roads as they would on dry roads. Thus, there may be a correlation between wet roads and automobile accidents, but wet roads do not cause automobile accidents.

For our purposes, a correlation analysis is calculated just like a regression. However, your question and hypothesis do not imply a cause-and-effect relationship. A correlation is simply the measure of association between two variables. In this test you also

report the same results as a regression except for the r^2 value. A correlation coefficient is reported, which is the square root of the r^2 from the regression. If $r = 0.0$ there is no association. If $r = -1.0$ or $+1.0$, there is a strong negative or positive association, respectively. Therefore, r ranges from -1.0 to $+1.0$. Most importantly, remember never to use the word *relationship* when performing a correlation analysis. We say either that there is an *association* or that there is *no association* between the variables being investigated.

Let us imagine that you are interested in investigating the association between age and testosterone levels in men. You collect data from 20 individuals ranging from 5 to 30 years of age. Is there a significant association between age (independent variable) and testosterone level (dependent variable)?

You may be tempted to analyze these data using a regression analysis. However, you cannot assume any causal effect of age on testosterone levels because you have not directly tested the effects of age on testosterone. There may be other reasons for the observed changes in your response variable. For example, other hormones produced by the brain may be driving the change in testosterone levels. Thus, although increasing age is associated with changing testosterone levels, age may or may not directly cause these changes. Note that in many situations, the direct effect of a variable (e.g., age, body size, brain size) cannot be manipulated. Thus, no cause-and-effect relationship can be implied. In such instances, you may only address the association between two variables, and the most appropriate way to do so is with a correlation analysis.

Now complete a correlation analysis using the following data set and create a graph showing the association between testosterone and age. Your results and graph should look like Figure 1-4. Notice that there is no best-fit line shown in the correlation graph, for this would indicate a cause-and-effect relationship and an ability to calculate y from x.

BOX 1.11 Example result statement for a correlation.

We found a significant association between age and testosterone level in men ($F = 62.29$; df = 1,18; $P < 0.05$).

Testosterone level in men is strongly correlated with age ($r = 0.873$).

(Note that your lab instructor may request that you use a Pearson-Moment correlation. Many statistical software packages now commonly run correlation analyses and may be more appropriately used for such analyses.)

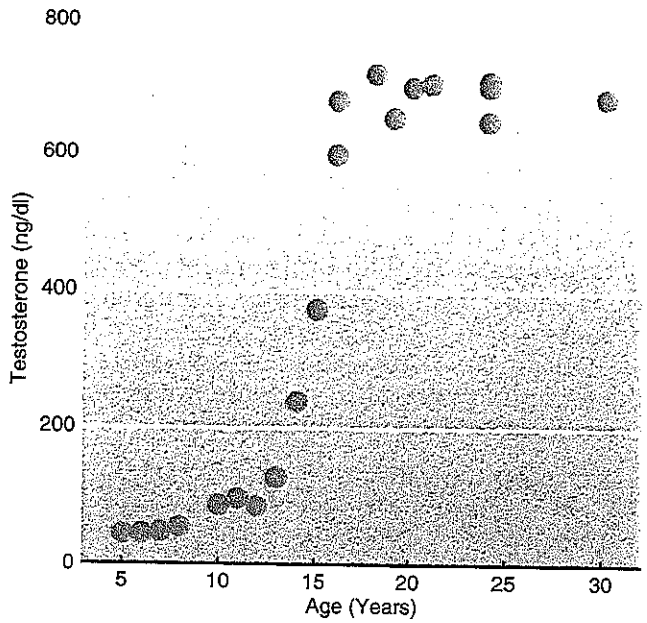

Figure 1-4 Example plot of a correlation.

Age (years)	Testosterone (ng/dl)
10	85
6	42
5	40
8	52
14	240
16	680
18	720
24	708
15	375
12	83
21	705
24	650
7	45
11	95
20	700
13	125
16	600
19	655
24	702
30	685